Numerical exploration of statistical physics
Bukva, A.

Citation

5

Criticality versus Uniformity in Deep Neural Networks

Attribution

Abstract

Deep feedforward networks initialized along the edge of chaos exhibit exponentially superior training ability as quantified by maximum trainable depth. In this work, we explore the effect of saturation of the tanh activation function along the edge of chaos. In particular, we determine the line of uniformity in phase space along which the post-activation distribution has maximum entropy. This line intersects the edge of chaos, and indicates the regime beyond which saturation of the activation function begins to impede training efficiency. Our results suggest that initialization along the edge of chaos is a necessary but not sufficient condition for optimal trainability.

5.1. Introduction

Over the past decade or so, deep learning has emerged as one of the most powerful tools for processing and analyzing data, and has proven successful on an increasingly wide range of computational challenges. These remarkable feats include highly accurate image classification [83], advanced generative modelling of images [84], natural language processing [85], accurate protein structure predictions [86], and besting humans in a wide range of games [87]. Key to these neural networks' success is the extremely large number of parameters—generally speaking, the *expressivity* of a neural network increases with depth [88]. Expressivity refers to the range of functions that a network can approximate, with the network being understood as simply a function from the space of inputs to the space of outputs. However, the price we must pay for larger and more powerful networks is that they are more difficult to train; for example, the risk of vanishing or exploding gradients is exacerbated with depth [89]. Hence, an improved understanding of how the network parameters impact trainability is highly valuable, as even small improvements in the initialization of deep neural networks can make intractable problems tractable.

In this work, we study trainability in deep random feedforward neural networks. Such networks are frequently used in the literature due to their analytical tractability: the phase space is two-dimensional and parameterized by the variances of the initial weight and bias distributions: σ_w^2 and σ_b^2.[1] This makes them useful models for investigating general features of deep networks. In particular, we will be concerned with the behavior of the pre- and post-activations, in terms of both their distributions as well as the accuracy of the network on a classic image classification task, namely MNIST (numerical digit recognition) and CIFAR-10 (colored images, which we convert to grayscale).

More specifically, we build on previous work [14, 15] which demonstrated the presence of an order-to-chaos phase transition in this class of deep networks. Intuitively, correlations in the input that we wish to learn are exponentially suppressed with depth in the ordered (analogously, low-temperature) phase, and washed-out by noise in the chaotic (high-temperature) phase; these two phases are characterized by vanishing or exploding gradients, respectively. The boundary between these two phases is a critical line called the *edge of chaos*,[2] which is a continuous phase transition characterized by a diverging correlation length ξ for the layer-to-layer two-point function of the neurons. Since the correlation length sets the depth scale at which information can propagate, this theoretically enables networks of arbitrary depth to be trained at criticality (more generally, networks are trainable provided their depth

---

[1] As is standard in the literature, we restrict to zero-mean networks, as initializing with a small non-zero mean does not qualitatively change our results.

[2] Technically, this should be called the edge of stability, but we will use edge of chaos synonymously with criticality for consistency with the literature.

does not exceed the scale set by ξ). In other words, the deeper the network, the closer one must lie to the edge of chaos; this was demonstrated in [15] along a slice of parameter space at bias variance 0.05 and weight variance ranging from 1 to 4, and subsequently generalized/corroborated in, e.g., [90–92]

Several questions naturally arise from the above work. First, given that the network parameters will evolve under training in order to minimize the specified cost function and, in particular, develop interdependencies, why does the choice of initialization have such a decisive effect on network performance?[3] Indeed, it was observed in [92] that the hidden-layer pre-activation distributions (as quantified by their variance) rapidly approach some asymptotic value within 10 or fewer layers, and then remain relatively unchanged for arbitrarily many additional layers. We corroborate this fact at the level of the post-activation in fig. 5.6 of appendix 5.6.1.

Second, what role does the particular distribution of post-activations in a given layer play in determining network performance? For example, the activation function considered in [15] is hyperbolic tangent, which we adopt henceforth. When $\sigma_b^2 \ll 1$ and $\sigma_w^2 \lesssim 1$, the pre-activations z of the hidden layers are approximately Gaussian-distributed with small variance (cf. (5.8)). In this case, $\tanh(z) \approx z$, so the network behaves like a linear network. These are quite restrictive, being incapable of representing functions whose output data are non-linearly separable and cannot be generated by a combination of linearly separable data. In the opposite extreme, for large values of σ_w^2 and σ_b^2, the pre-activation variance becomes so large that the post-activation distribution becomes peaked at ± 1. In other words, large pre-activation variance saturates the tanh, causing it to behave like a discrete step-function. One expects this regime to also impair trainability, since the gradients on which the back-propagation algorithm depends become vanishingly small everywhere except near the origin.[4] Thus, it seems that one should seek to remain somewhere between these two extremes. Quantifying this is one of the main motivations for the present work.

In particular, note that in both the linear and the saturation regimes, one expects the expressibility of the network to be poor. In contrast, between these extremes lies a region in which the post-activation distribution is approximately uniform, and hence we might expect the expressibility of the network to be maximized at this point. To see this, recall that the uniform distribution has maximum entropy, which measures the number of possible states any particular system can have; a step function, in contrast, can only store a single bit of information, and hence has a low entropy of $\ln 2$. This leads to the conjecture that networks whose internal distributions are approximately uniform, i.e., maximally entropic, have higher expressibility, and hence

---

[3]In other words, why does the network remain near the initialization regime (e.g., the edge of chaos) as it evolves?

[4]Recall that the updates to the weights and biases under gradient descent contain products of the derivatives of the activation functions in all higher layers.

might enjoy a performance advantage. Of course, given approximately Gaussian pre-activations, the post-activation distribution of tanh cannot be exactly uniform, but we can quantify the degree of uniformity via the relative entropy (defined below). In fact, we will show that there is a *line of uniformity* on the (σ_w^2, σ_b^2) phase space along which the post-activation distribution is as uniform as possible. This line intersects the aforementioned edge of chaos (see fig. 5.1), and the relative importance of lying near this line is the primary question we shall explore below.

We shall begin by deriving an expression for the line of uniformity, defined by the condition that the distribution of the final hidden layer minimizes the relative entropy with respect to the uniform distribution. The computation uses many of the same ingredients as [15], and the interested reader is encouraged to turn there for more background. We then examine proximity to this line in relation to the edge of chaos considered in previous works.

We find that for deep networks away from the edge of chaos, the exponential suppression dominates, and no benefit from uniformity is observed. However, along the edge of chaos – where the suppression is only polynomial – we find a relatively sharp fall-off in the post-training accuracy to the right of the line of uniformity. The location of this fall-off depends on the learning rate, since decreasing the learning rate can increase the final accuracy, but at the cost of additional computing time (see fig. 5.2). This suggests that criticality is a necessary but not sufficient condition for optimal trainability.

This dependence on other hyperparameters illustrates that optimal trainability is not just a matter of final accuracy but also of efficiency, i.e., how quickly the final accuracy is reached. Since computational limits exist, we shall rely on an intuitive notion of efficiency per epochs in addition to accuracy; that is, we consider the accuracy achieved after a fixed number of training epochs. It is conceivable that in the limit of infinite training epochs accuracy differences disappear, so that formally, the configurations are equally good. In a practical sense however, they clearly are not.

Note that there can obviously be very many notions of efficiency depending on which resource(s) one considers most valuable. Here, we are implicitly prioritizing training time, i.e., number of epochs. If one were to put the premium on floating point operations used in training, then one would instead measure efficiency as in [93]. Yet another concept called learning efficiency has to do with how much time it takes to run a learning algorithm and, in particular, how this scales with the size of the input space [94].

Returning to our main question, to isolate the effects of uniformity *away* from the edge of chaos, we also examine networks which are both shallow (i.e., not yet exponentially suppressed) and narrow (i.e., low expressibility per layer), and confirm

that training efficiency, in the sense described above, degrades to the right of the line of uniformity (i.e., away from the origin), though final accuracy need not. In contrast to the edge of chaos, the line of uniformity is not a sharp phase boundary, but it does indicate coarsely the parameter boundary where activation saturation starts to affect training efficiency. This not only establishes the more obvious point that, even in deep random feedforward toy models on the edge of chaos, backpropagation training depends sensitively on activation function choice, as earlier emphasized in [95, 96], but also that for a given activation function choice there are optimal points or regions on the edge of chaos itself.

5.2. The line of uniformity

We can estimate the location of the line of uniformity by capitalizing on the fact that wide networks, with a large number N of neurons in each hidden layer, are approximate Gaussian processes. At finite N, the neurons in a given layer are not independent due to their shared dependence on the neurons in the previous layer. Physically however, the non-Gaussianities that can be seen by marginalizing over the previous layer(s) can be thought of as interactions that are $1/N$ suppressed [97, 98]. Hence, in the limit $N \to \infty$, the distribution of pre-activations becomes Gaussian, essentially by the central limit theorem. This greatly simplifies the analysis, and is the reason for the widespread use of such models in previous studies, including [15].[5]

Thus, at large-N, the distribution of pre-activations z for any hidden layer takes the form

$$p(z;\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{z^2}{2\sigma^2}}\,, \tag{5.1}$$

where σ^2 is the variance, and we assume the mean $\mu = 0$ since adding a small finite mean does not qualitatively change our results. If the activation function $\phi(z)$ is one-to-one and once-differentiable, then the distribution of post-activations x will be given by

$$p_\phi(x;\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma\,\phi'\big(\phi^{-1}(x)\big)}\, e^{-\frac{\phi^{-1}(x)^2}{2\sigma^2}}\,. \tag{5.2}$$

Concretely, for $\phi(z) = \tanh(z)$, this yields

$$p_\phi(x;\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma(1-x^2)}\, e^{-\frac{\mathrm{arctanh}(x)^2}{2\sigma^2}}\,, \tag{5.3}$$

with $x \in [-1,1]$. The corresponding variance is given by

$$\sigma_\phi^2 = \int_{-1}^{1} \mathrm{d}x\, x^2\, p_\phi\big(x;\sigma^2\big)\,. \tag{5.4}$$

---

[5]One will often see the phrase "mean-field theory" used in place of the central limit theorem in this context; however, as pointed out in [98], this is not technically correct, and mean-field theory does not necessarily correspond to the $N \to \infty$ limit.

As mentioned above, we quantify the uniformity of the post-activation distribution p_ϕ by the relative entropy or Kullback-Leibler divergence with respect to the uniform distribution p_{uni},

$$S(p_{\text{uni}}||p_\phi) = \int_{-1}^{1} dx \, p_{\text{uni}}(x) \ln \frac{p_{\text{uni}}(x)}{p_\phi(x)} \, . \tag{5.5}$$

Substituting in (5.3) and $p_{\text{uni}} = \frac{1}{2}$, this yields

$$S(p_{\text{uni}}||p_\phi) = \frac{1}{2}\ln\left(8\pi\sigma^2\right) + \frac{\pi^2}{24\sigma^2} - 2 \, . \tag{5.6}$$

This has a minimum at

$$\sigma_{\text{min}}^2 = \frac{\pi^2}{12} \approx 0.822 \, . \tag{5.7}$$

Therefore, we wish to find the set of points (σ_w^2, σ_b^2) at which the variance of the final hidden layer is σ_{min}^2; this will define the line of uniformity. To proceed, we use the recursion relation

$$\sigma_\ell^2 = \sigma_w^2 \sigma_{\phi,\ell-1}^2 + \sigma_b^2 \, , \tag{5.8}$$

which follows from the large-N condition discussed above (i.e., the neurons on any given layer can be treated as i.i.d. random variables). Note that this is exactly the same as eq. (3) of [15], where our σ_ℓ^2 is their q_{aa}^ℓ and our $\sigma_{\phi,\ell-1}^2$ is the corresponding integral expression.[6] This recursion relation ostensibly requires the variance of the first hidden layer, σ_1^2, as an input. However, it turns out that (5.8) quickly converges to a fixed value σ_*^2, which (by definition) is a function of σ_w^2 and σ_b^2, but not of σ_1^2:

$$\sigma_*^2 = \sigma_w^2 \sigma_{\phi,*}^2 + \sigma_b^2, \tag{5.9}$$

where $\sigma_{\phi,*}^2$ is σ_ϕ^2 evaluated at σ_*^2; see [14] for further discussion of this convergence. In appendix 5.6.1, we have demonstrated numerically that the corresponding post-activation distribution indeed converges rapidly to one which depends only on the initialization point (σ_w^2, σ_b^2).

Now, consider a fixed value of σ_*^2 (and hence also of $\sigma_{\phi,*}^2$). Then we can consider (5.9) as an expression for σ_b^2 as a function of σ_w^2, which defines a line in phase space of the form

$$\sigma_b^2 = \sigma_*^2 - \sigma_{\phi,*}^2 \sigma_w^2. \tag{5.10}$$

where σ_*^2 is the y-intercept, and $-\sigma_{\phi,*}^2$ is the slope. Since the relative entropy (5.6) of the final hidden layer is only a function of its variance, the lines of constant σ_* given by (5.10) are also lines of constant relative entropy. In particular, the line of

---

[6]Explicitly, the variance can be written as $\sigma_\phi^2 = \int \mathscr{D}z \, [\phi(\sigma z)]^2$, where $\mathscr{D}z = \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ is the standard Gaussian measure.
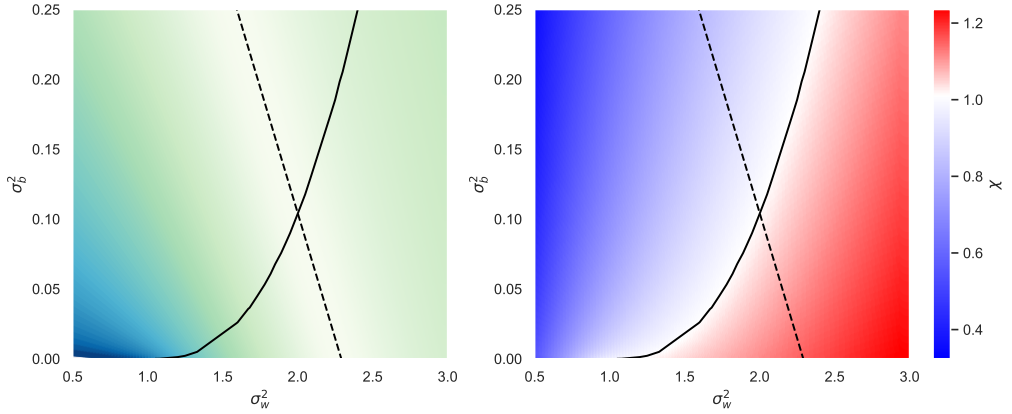
Figure 5.1: (Left) Contour plot of the logarithm of the relative entropy in the (σ_w^2, σ_b^2) plane. The dashed line is the line of uniformity—saturation increases to the right of it and linearity increases to the left of it. (Right) Contour plot of $\chi = e^{-1/\xi}$. The ordered/low-temperature phase is shaded blue, while the chaotic/high-temperature phase is shaded red. In both, the solid black line is the edge of chaos, while the dashed black line is the line of uniformity.

uniformity (minimum relative entropy) is given by (5.10) with $\sigma_*^2 = \sigma_{\min}^2 = \frac{\pi^2}{12}$, cf. eq. (5.7). There is no closed-form expression for $\sigma_{\phi,\min}^2$, but we can evaluate (5.4) numerically to obtain $\sigma_{\phi,\min}^2 \approx 0.359$. In summary, the line of uniformity (LOU) is given by

$$\text{LOU}: \quad \sigma_b^2 = \sigma_{\min}^2 - \sigma_{\phi,\min}^2 \sigma_w^2, \tag{5.11}$$

with $\sigma_{\min}^2 = \frac{\pi^2}{12} \approx 0.822$ and $\sigma_{\phi,\min}^2 \approx 0.359$. In the left panel of fig. 5.1, we present a contour plot of the logarithm of the relative entropy. The line of uniformity is the dashed black line—to the left of it, as one approaches the origin, is the linear regime; and to the right, the activation becomes more and more saturated. For comparison, the edge of chaos is the solid black line.

5.3. The edge of chaos

The method for computing the edge of chaos as a function of σ_w^2 and σ_b^2 is described in [14, 15]. Once we have σ_*^2, as described previously, then we can define the quantities

$$\chi = \sigma_w^2 \int \mathscr{D}z \left[\phi'(\sigma_* z)\right]^2, \qquad\qquad \xi = -\frac{1}{\ln \chi}, \tag{5.12}$$

where $\mathscr{D}z$ is the standard Gaussian measure, cf. footnote 6, and ξ is the correlation length mentioned in the introduction (note that this is denoted ξ_c in [15]).

The meaning of χ will be discussed in the next paragraph, while the meaning of ξ is as follows: we consider two identical copies of the network and feed them slightly different inputs. Then, we can study the correlation (i.e., covariance) between a neuron in one copy and the same neuron in the second copy as a function of the layer. This correlation will decay exponentially for deeper layers with a characteristic length scale, ξ. (Strictly speaking, this is only true in the ordered phase: in the chaotic phase, the quantity ξ is complex-valued and cannot be interpreted as a correlation length). The edge of chaos is defined as the critical point, where the correlation length ξ diverges.

As discussed in more detail in [14, 15], χ is obtained as the derivative of the aforementioned covariance with respect to that in the previous layer, and probes the stability of the fixed point when the covariance is unity: $\chi > 1$ implies that we approach this point from below (unstable), while $\chi < 1$ implies that we approach this point from above (stable).[7] The edge of chaos corresponds to $\chi = 1$, where ξ diverges.

To find the edge of chaos, we can scan over the space of tuples (σ_w, σ_*) to find those which satisfy the condition $\chi = 1$. We then feed these into (5.8) to find the corresponding value of σ_b. In this manner, we can find arbitrarily many points on the edge of chaos (EOC). Within some finite range of σ_w^2 values, we can find a good fit to the EOC. In the range $1 \le \sigma_w^2 \le 10$, a good polynomial fit is

$$\text{EOC}: \quad \sigma_b^2 = \sum_{n=2}^{9} \frac{c_n}{n!} (\sigma_w^2 - 1)^n, \tag{5.13}$$

with fit coefficients

n	c_n	n	c_n
2	0.0190	6	-1.15
3	0.778	7	0.769
4	-1.07	8	-0.328
5	1.25	9	0.0672

(5.14)

Of course, we can reduce the number of fit coefficients needed by reducing the range of σ_w^2 values over which we require the fit to be good.

The form of this fit is designed such that it contains the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$, and that the edge of chaos has zero slope at this point. We justify these conditions analytically in appendix 5.6.2. In the right plot in fig. 5.1, we present a contour plot of χ. Again, the edge of chaos is drawn as a solid black line and the line of uniformity as a dashed line. The point of intersection of the edge of chaos and line of uniformity is found to be

$$(\sigma_w^2, \sigma_b^2)_{\text{intersect}} = (2.00, 0.104) \,. \tag{5.15}$$
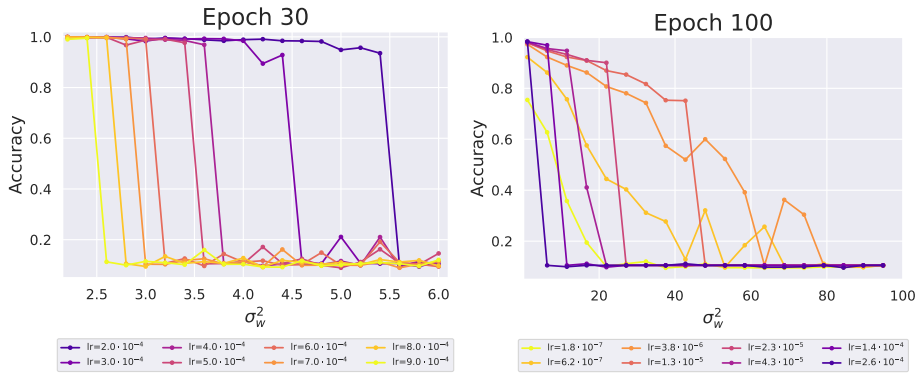
---

[7]See [99] for a pedagogical explanation.

Figure 5.2: Accuracies on MNIST for distributions of initial weights along the edge of chaos in a deep ($L = 100$), wide ($N = 784$) neural network with tanh activation function, for a range of learning rates, after 30 epochs (left) and 100 epochs (right). We observe a drop-off in accuracy beyond a value σ_w^2 which is up to an order of magnitude larger than the point at which the line of uniformity is crossed. For learning rates of the order typically used in the literature, this point is near the intersection of the LOU and the EOC, but moves to higher values of σ_w^2 for smaller learning rates. When learning rates become extremely small ($r < 10^{-5}$), learning becomes highly inefficient, and the drop-off less sharp for the training duration considered. Networks were trained via stochastic gradient descent with batch size 64 and momentum 0.8.

5.4. The impact of uniformity along the edge of chaos

To the right of the line of uniformity, neurons begin to saturate the tanh activation function, i.e., approach ± 1. This implies that backpropagation based on gradient descent should be less efficient, and hence networks should reach a lower accuracy in a fixed amount of training time. The Google Brain collaboration has already established that at the edge of chaos, learning accuracy is enhanced due to polynomial rather than exponential decay of correlations as a function of network depth [15]. Combining the two insights, optimal learning should therefore take place on the edge of chaos near the line of uniformity.

To test this hypothesis, we have performed the MNIST image classification task in networks ranging up to a depth of $L = 100$ hidden layers at various points along the edge of chaos. The resulting learning accuracy is shown in fig. 5.2. We see that this expectation is partially validated. On the left side of the line of uniformity – but to the right of the linear regime – all points on the edge of chaos are equally good at learning. But beyond a certain point, which lies to the right of the intersection point (5.15) of the edge of chaos and line of uniformity, the final accuracy decreases. However, this drop-off point is substantially (up to an order of magnitude) displaced to the right of the intersection point, indicating that the line of uniformity is perhaps better thought of as a region rather than a narrow band, and depends on hyperparameters (such as the learning rate) as mentioned above. Nevertheless, for typical learning rates used in the literature of order 10^{-3}, such as used in [15], the drop-off

point at approximately $\sigma_w^2 \sim 2.5$ is indeed fairly close to the intersection between the line of uniformity and the edge of chaos at $\sigma_w^2 = 2$.

We repeated this exercise for the CIFAR-10 image classification task, and present the corresponding results in fig. 5.3. We converted the colored images to grayscale to reduce the input size by a factor of 3. The drop-off in accuracy along the edge of chaos towards larger values of σ_w^2 is still present, though the effect is not as dramatic as it is for MNIST. This is not surprising as CIFAR is a much more difficult task than MNIST and so we expect that the saturation of slightly more or fewer neurons will have a much less decisive effect. We note however that in the regime of extremely small learning rates, where training MNIST becomes highly inefficient, the MNIST and CIFAR results appear similar insofar as neither exhibits the obvious sharp drop-off observed for MNIST at the higher learning rates generally used in practice.
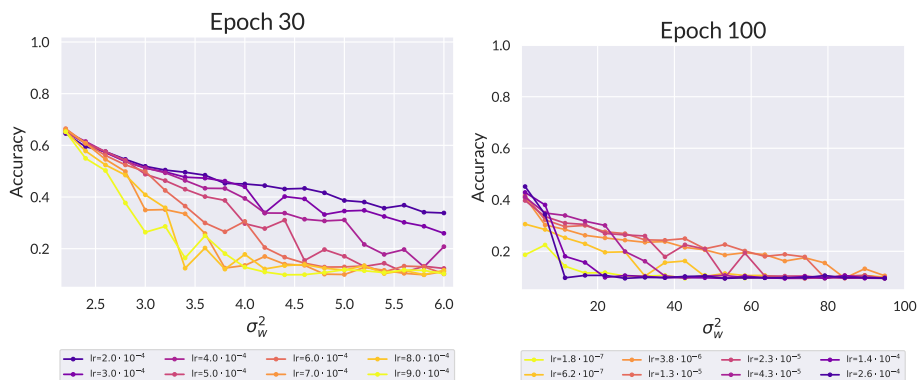

Figure 5.3: Accuracies on CIFAR-10 for distributions of initial weights along the edge of chaos in a deep ($L = 100$), wide ($N = 1024$) neural network with tanh activation function, for a range of learning rates, after 30 epochs (left) and 100 epochs (right). The drop-off in accuracy towards higher values of σ_w^2 here is much more gradual than the sharp drop-offs observed for MNIST in fig. 5.2 (at all but the lowest learning rates, $r < 10^{-5}$). Networks were trained via stochastic gradient descent with batch size 64 and momentum 0.8.

Thus, the line of uniformity is not a sharp boundary, unlike the edge of chaos. This is somewhat inherent in its definition, which selects proximity to the uniform distribution of final hidden layer weights as a condition for efficient learning based on the entropic argument given above, but does not specify any particular fall-off behavior. The line of uniformity does, however, give an estimate of where the saturation of the activation function should start to affect learning, and by extension, the point at which saturation of the activation function begins to hinder learning efficiency. To summarize: on the left side of the line of uniformity, the distributions are sufficiently narrow that saturation of the tanh activation function does not occur, and all initial weight distributions along the EOC learn equally well. Conversely, on the right side of uniformity, neurons saturate the activation function and hence hamper learning, even along the EOC. This is our main observation. Importantly, we note that the

studies by [15] were performed to the left of the point where the line of uniformity crosses the edge of chaos and hence at optimal efficiency.

Before moving on to our final set of experiments, we note that the above conclusion is of course specific to saturating activation functions, specifically tanh. This is one motivation for the use of non-saturating activation functions such as ReLU or SWISH, though the unbounded nature of such functions presents its own set of training difficulties. While a similar analysis of uniformity, as quantified by the maximally entropic distribution, for non-saturating activation functions is beyond the scope of this work, a brief inspection of learning efficiency along the EOC for SWISH shows no loss of accuracy in agreement with the absence of saturation effects; see appendix 5.6.4.[8]

5.5. Uniformity away from the EOC

Thus far, we have examined the impact of uniformity on training efficiency along the edge of chaos. Now, we would like to explore whether the line of uniformity still affords training advantages even for networks initialized far from criticality. In attempting to exhibit this however, one quickly finds that the edge of chaos represents a far more dominant effect than the line of uniformity. A close inspection of the learning accuracy of deep (L=300) and wide (N=784) MNIST learning networks shows that there is no discernible difference in learning accuracy away from the edge of chaos: it is simply poor everywhere (see fig. 5.7 in appendix 5.6.3, also [15].) This can be understood from the form of the correlation functions: away from the edge of chaos, correlations damp exponentially $\sim e^{-L/\xi}$. For a deep network, this exponential damping will erase any finer difference in accuracy results. Along the edge of chaos, the damping is only polynomial and, therefore, the finer difference remains, as seen in fig. 5.2. In shallow networks however, the exponential damping does not have sufficient time to compound, and if the network is also narrow and hence has low expressibility per layer, we can explore the effect of uniformity even away from criticality in such models.

Furthermore, it is common lore that efficient backpropagation needs sufficient gradients, and that such gradients are absent if most of the post-activation functions saturate to a fixed asymptotic value. However, if a sufficient number of weight and bias values are such that there remain trainable paths through a saturated landscape, the model will still learn, even though, distribution-wise, most of the neurons have saturated. Therefore, the inefficiency due to saturation discussed above can be displayed more clearly by choosing narrower networks with smaller N, where we might expect that uniformity – that is, maximally entropic distributions – may afford the

---

[8]For both SWISH and tanh, the edge of chaos is a line of critical initalizations through phase space, while for ReLU it is only a single point [97].
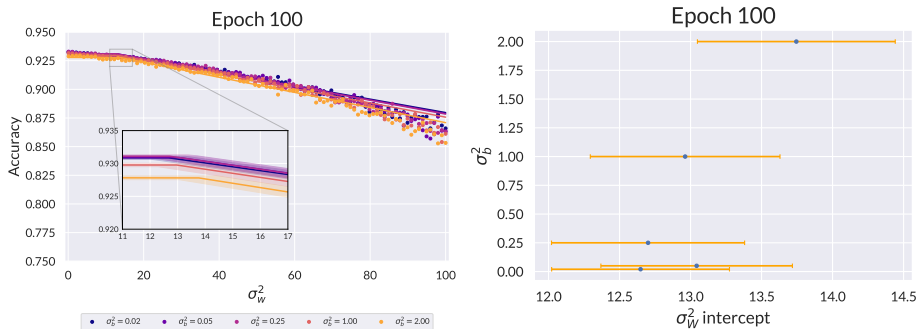
most advantage.

Figure 5.4: For small networks, the learning efficiency exhibits threshold behavior as a function of σ_w^2. Shown are results for MNIST trained on a $N = 8$, $L = 1$ network sampled over 50 network initializations. The inset shows the fits in the threshold region. The bottom figure shows that the location of this threshold in σ_w^2 decreases with increasing σ_b^2 consistent with the trend implied by the line of uniformity threshold. As explained in the text, there is a multiplicative factor involved and the large-N analysis cannot be straightforwardly transplanted to this small-N case. The uncertainty bars are propagated from the uncertainties in the accuracy versus σ_w^2 data points.

The effect of lying near uniformity is therefore strongest in shallow, narrow networks rather than deep, wide networks where the edge of chaos effect dominates. For these small networks, some of the asymptotic analysis above locating the LOU and EOC does not immediately apply, since the network is unable to reach the asymptotic value σ_*^2 of the pre-activation variance.[9] At the same time, the input variance and mean, σ_0^2 and μ_0, actually *do* matter in this case and, with this information, we can roughly estimate the location of the line of uniformity. For example, for $L = 1$, we have $\sigma_1^2 = \sigma_w^2(\sigma_0^2 + \mu_0^2) + \sigma_b^2$ and the line of uniformity would be where $\sigma_1^2 = \sigma_{min}^2 = \frac{\pi^2}{12}$. For example, for MNIST, $\sigma_0^2 \approx 0.095$ and $\mu_0^2 \approx 0.017$, so the line of uniformity can be estimated as $\sigma_b^2 \approx \frac{\pi^2}{12} - 0.112\sigma_w^2$. Equivalently, for fixed σ_b^2, this gives a σ_w^2-threshold of $\sigma_w^2 \sim 7.35 + 8.93\sigma_b^2$ beyond which we expect saturation effects to decrease training efficiency. For $L = 2$, we would iterate the above process once more, passing through the activation function; this gives an estimated threshold of $\sigma_w^2 \approx 3.5 + 8.93\sigma_b^2$.

Results for $L = 1$ are shown in fig. 5.4, and results for $L = 2$ are shown in fig. 5.5. As predicted, we observe that the accuracy retains a high, approximately constant value up to a σ_b^2-dependent threshold for σ_w^2, and then decays approximately linearly thereafter. To determine the threshold empirically, we fit the data to a function

---

[9]In this sense, we may take "shallow" to mean $L \le 5$, since as shown in fig. 5.6, by $L \approx 6$, the network has reached σ_*^2. Strictly speaking however, the predictions for the EOC as well as the LOU are ill-defined in narrow networks, since these are no longer approximately Gaussian, and also appear to be beyond the reach of current perturbative approaches [98].

of the form

$$A_{\text{fit}}(\sigma_w^2) = A_{\max} - r(\sigma_w^2 - \sigma_{w,\text{thr}}^2)\,\Theta(\sigma_w^2 - \sigma_{w,\text{thr}}^2), \tag{5.16}$$

where $A_{\max}$ is the maximum accuracy, $\sigma_{w,\text{thr}}^2$ is the threshold value, r is the rate of linear decay, and Θ is the Heaviside step function. Each accuracy vs. σ_w^2 data point is an average over 20 instantiations of the network and thus comes with its own variance. These propagate into uncertainty bars for the three fit parameters. We plot the threshold for different values of σ_b^2 in fig. 5.4 for $L = 1$. This qualitatively confirms our expectations, though the empirical value of the threshold is about a factor of 2 greater than the analytical prediction, and the slope about a factor of 8 smaller. However, given that we are applying a large-N analysis to a relatively narrow network ($N = 8$), an $\mathcal{O}(1)$ quantitative discrepancy is reasonable. For $L = 2$ the corresponding results are presented in fig. 5.5, again showing qualitative agreement. The empirical threshold in this case is about a factor of 4 greater than the theoretical value, and the slope is a factor of 8 smaller.
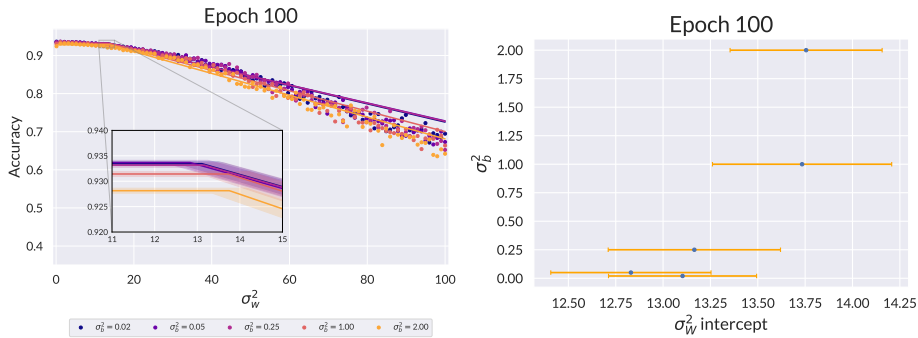

Figure 5.5: Small-network threshold behavior as in fig. 5.4, for MNIST trained on a network with $N = 8$ and $L = 2$, sampled over 50 initial conditions drawn from (σ_w^2, σ_b^2). The bottom figure shows that the location of this threshold decreases with σ_b^2 consistent with the trend implied by the line of uniformity.

Conclusion In this work, we establish that for deep random feedforward networks along the edge of chaos, the efficiency of training via stochastic gradient descent still depends on non-saturation of the activation function. Similar points have been made previously in [95, 96], which compared the performance of difference activation functions initialized at one point on their respective edges of chaos. However, what we demonstrate for the tanh activation function is that not all points on the edge of chaos are equally efficient at learning. Within a fixed number of training epochs (~ 100), activation function saturation eventually impedes learning if we push the weight and bias variances too far to the right of the line of uniformity, defined to be where the final layer post-activation is most uniformly distributed, i.e., maximally entropic. Unlike the edge of chaos, which separates chaotic and ordered outputs, the line of uniformity does not mark an abrupt change in the overall be-

havior of the network. Rather, it simply indicates roughly the point where the saturation of the activation function begins to impede learning. We demonstrate this for shallow and narrow networks as well, where the exponential damping of neuron correlations away from the edge of chaos becomes much less of a decisive factor in determining training efficiency.

Acknowledgments

5.6. Appendix

5.6.1. Independence of σ_*^2 on σ_1^2

The exact pre- or post-activation distribution at a given layer obviously does depend on σ_1^2, the pre-activation variance at the first hidden layer. This dependence is generated via the recursion relation (5.8). However, at the fixed point, the asymptotic distributions do not depend on σ_1^2. Indeed, the relation that the asymptotic pre-activation variance satisfies is eq. (5.9), which does not depend on σ_1^2 at all. We can demonstrate this fact by plotting the evolution of the post-activation distribution for fixed σ_w^2 and σ_b^2, but for many values of σ_1^2. In fig. 5.6, we show this for $(\sigma_w^2, \sigma_b^2) = (1.76, 0.05)$ for several values of σ_1^2, both less than and greater than σ_*^2 which turns out to be $\sigma_*^2 \approx 0.57$ in this case. When $\sigma_1^2 < \sigma_*^2$, the post-activation distribution starts out narrower and spreads out, whereas when $\sigma_1^2 > \sigma_*^2$ it starts out more peaked at ± 1 and then flattens out.
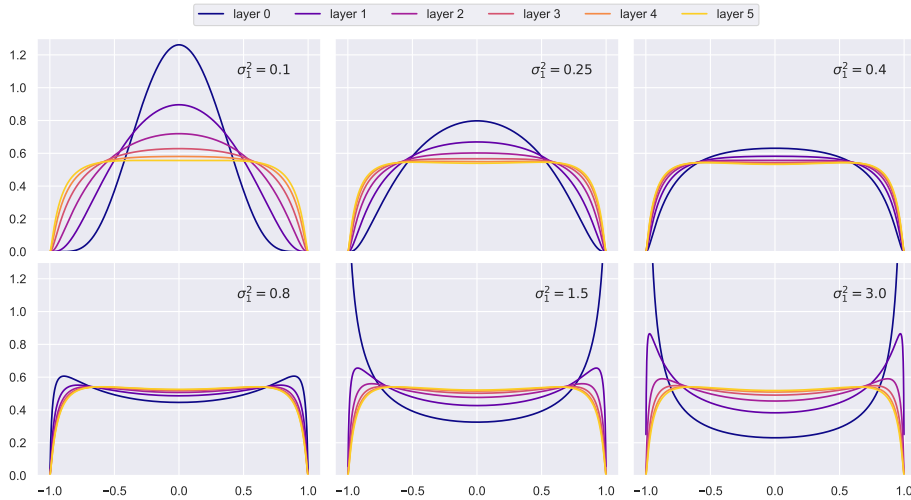

Figure 5.6: Layer-to-layer evolution of the post-activation distribution at $(\sigma_w^2, \sigma_b^2) = (1.76, 0.05)$ for six different values of the first hidden layer pre-activation variance σ_1^2. The post-activations converge to the asymptotic distribution within about five layers.

5.6.2. Analytic Details of the Fixed Point Computation

In this appendix, we will show that the edge of chaos contains the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$ and has zero slope there. At this point, the fixed-point equation (5.9) reads

$$\sigma_*^2 = \sigma_{\phi,*}^2 \,. \tag{5.17}$$

The left-hand side is the fixed-point pre-activation variance, whereas the right-hand side is the corresponding post-activation variance. As long as $|\phi(z)| < |z|$, which is the case for $\phi(z) = \tanh(z)$ except at $z = 0$, the variance of the post-activation will always be smaller than that of the pre-activation. Therefore, the only solution at this point is $\sigma_*^2 = \sigma_{\phi,*}^2 = 0$ and thus at this point $\phi'(\sigma_* z) = \text{sech}^2(0) = 1$ and $\chi = 1$ or $\xi = \infty$. Hence, this point is on the edge of chaos.

Now, consider eq. (5.10), but now along the edge of chaos rather than the lines of constant σ_*^2. Let σ_w^2 be our independent parameter along the edge of chaos and take a derivative with respect to it:

$$\frac{\partial \sigma_b^2}{\partial \sigma_w^2} = \left(1 - \sigma_w^2 \frac{\partial \sigma_{\phi,*}^2}{\partial \sigma_*^2} \right) \frac{\partial \sigma_*^2}{\partial \sigma_w^2} - \sigma_{\phi,*}^2 , \tag{5.18}$$

where we have used the fact that $\sigma_{\phi,*}^2$ depends on σ_w^2 only through its dependence on σ_*^2.

To compute the derivative $\frac{\partial \sigma_{\phi,*}^2}{\partial \sigma_*^2}$, it is convenient to first rewrite the integral expression for σ_ϕ^2 in (5.4) by changing back to the original pre-activation variable z:

$$\sigma_\phi^2 = \int_{-1}^{1} dx \, p_\phi(x; \sigma^2) \, x^2 = \int_{-\infty}^{\infty} dz \, p(z; \sigma^2) \, \phi(z)^2 . \tag{5.19}$$

We can easily compute the various derivatives of the pre-activation distribution:

$$\frac{\partial p(z; \sigma^2)}{\partial \sigma^2} = \left(\frac{z^2}{\sigma^2} - 1 \right) \frac{p(z; \sigma^2)}{2\sigma^2}, \quad \frac{\partial^2 p(z; \sigma^2)}{\partial z^2} = \left(\frac{z^2}{\sigma^2} - 1 \right) \frac{p(z; \sigma^2)}{\sigma^2} = 2 \frac{\partial p(z; \sigma^2)}{\partial \sigma^2} . \tag{5.20}$$

Therefore, using integration by parts, and the fact that we can ignore boundary terms due to the fast fall-off of the Gaussian, we find

$$\frac{\partial \sigma_\phi^2}{\partial \sigma^2} = \int dz \, \frac{\partial p(z; \sigma^2)}{\partial \sigma^2} \, \phi(z)^2 = \frac{1}{2} \int dz \, \frac{\partial^2 p(z; \sigma^2)}{\partial z^2} \, \phi(z)^2 = \int dz \, p(z; \sigma^2) \big(\phi'(z)^2 + \phi(z) \, \phi''(z) \big) . \tag{5.21}$$

By rescaling the variable to σz, the first integral term above can be written as

$$\int dz \, p(z; \sigma^2) \phi'(z)^2 = \int \mathscr{D}z \, \big[\phi'(\sigma z) \big]^2 . \tag{5.22}$$

Note that when this is evaluated at σ_*^2 and multiplied by σ_w^2, we get precisely χ, as defined in (5.12). Let us give a name to the remaining integral in (5.21) evaluated at σ_*^2. For future convenience, we will put a relative minus sign in the definition below, the reason being that, for $\phi = \tanh$, the object $\phi\phi''$ is *negative* semi-definite:

$$\tilde{\chi} = -\sigma_w^2 \int dz \, p(z; \sigma_*^2) \, \phi(z) \, \phi''(z) = -\sigma_w^2 \int \mathscr{D}z \, \phi(\sigma_* z) \, \phi''(\sigma_* z) . \tag{5.23}$$

Then, (5.21) evaluated at σ_*^2 and multiplied by σ_w^2 reads

$$\sigma_2^2 \frac{\partial \sigma_{\phi,*}^2}{\partial \sigma_*^2} = \chi - \tilde{\chi} \,. \tag{5.24}$$

Now, let us define

$$\tilde{\xi} = -\frac{1}{\ln(\chi - \tilde{\chi})} \,. \tag{5.25}$$

This is precisely the object called ξ_q in [15], which is the length scale that controls the exponential decay of information propagation through the neural network from a single input.

Plugging eq. (5.24) back into eq. (5.18) gives

$$\frac{\partial \sigma_b^2}{\partial \sigma_w^2} = (1 - \chi + \tilde{\chi}) \frac{\partial \sigma_*^2}{\partial \sigma_w^2} - \sigma_{\phi,*}^2 \,. \tag{5.26}$$

Along the edge of chaos, $\chi = 1$, and so

$$\frac{\partial \sigma_b^2}{\partial \sigma_w^2} = \tilde{\chi} \frac{\partial \sigma_*^2}{\partial \sigma_w^2} - \sigma_{\phi,*}^2 \,, \tag{5.27}$$

Now, we can establish a simple bound on $\tilde{\chi}$ by virtue of the fact that $|\phi(z)| \le |z|$, for $\phi = \tanh$. To do this, let us first rewrite $\tilde{\chi}$ using the identity

$$\phi''(z) = -2 \tanh(z) \operatorname{sech}^2(z) = -2\,\phi(z)\,\phi'(z) \,. \tag{5.28}$$

Therefore,

$$\phi(z)\,\phi''(z) = -2\,\phi(z)^2\,\phi'(z) = -\frac{2}{3}\left[\phi(z)^3\right]' \,, \tag{5.29}$$

and

$$\tilde{\chi} = \frac{2\sigma_2^2}{3} \int \mathrm{d}z\, p(z;\sigma_*^2)\left[\phi(z)^3\right]' = -\frac{2\sigma_w^2}{3} \int \mathrm{d}z\, \frac{\partial p(z;\sigma_*^2)}{\partial z}\,\phi(z)^3 = \frac{2\sigma_w^2}{3\sigma_*^2} \int \mathrm{d}z\, p(z;\sigma_*^2)\, z\,\phi(z)^3 \,. \tag{5.30}$$

Therefore, since $|\phi(z)| \le |z|$ for $\phi = \tanh$,

$$0 \le \tilde{\chi} \le \frac{2\sigma_w^2}{3\sigma_*^2} \int \mathrm{d}z\, p(z;\sigma_*^2)\, z^4 = 2\,\sigma_w^2\,\sigma_*^2 \,. \tag{5.31}$$

Therefore, since we have already shown that $\sigma_*^2 = \sigma_{\phi,*}^2 = 0$ at the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$, it follows that $\tilde{\chi} = 0$ at this point as well and, from eq. (5.27),

$$\left. \frac{\partial \sigma_b^2}{\partial \sigma_w^2} \right|_{(\sigma_w^2, \sigma_b^2) = (1,0)} = 0 \,. \tag{5.32}$$

In other words, the edge of chaos has zero slope at the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$.

5.6.3. Implementation Details

Throughout this work, we have used a vanilla feedforward neural network of L hidden layers, each having the same depth N. As described, initial weights and biases are drawn from zero-mean Gaussian distributions with $\frac{\sigma_w^2}{N}$ and σ_b^2 respectively. Both MNIST and CIFAR-10 were trained using the standard cross-entropy loss function and no optimizer. This reproduces the results of [15] (see fig. 5.7), confirming critical behavior.
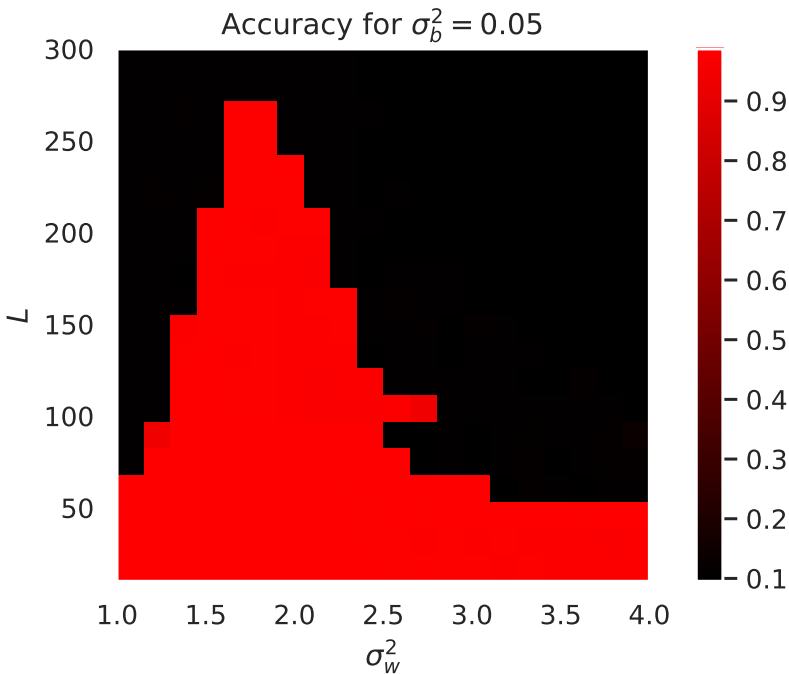

Figure 5.7: Optimal learning for deep neural networks at the edge of chaos as first shown by [15]. Shown is learning efficiency for MNIST training as a function of network depth L with $N = 784$ and choice of initial weight distribution σ_w^2 holding the initial bias distribution $\sigma_b^2 = 0.05$ fixed. At the edge of chaos $(\sigma_w^2, \sigma_b^2) = (1.76, 0.05)$, learning remains efficient even for very deep networks, but eventually ($L \sim 270$) goes down. This same behavior has been observed for deep feedforward networks in [15, 92]. The learning rate used is $\ell = 10^{-3}$ for $L < 100$ and $\ell = 10^{-4}$ for $L \geq 100$.

5.6.4. SWISH activation function

Throughout the text, we examined the impact of saturation via the line of uniformity for the tanh activation function. For non-saturating activation functions, it is an open question whether a similar notion of uniformity exists. While a full analysis of this is beyond the scope of this work, in this appendix we offer some preliminary

results for the SWISH activation function,

$$\text{swish}(z) = \frac{z}{1 + e^{-z}} \, , \tag{5.33}$$

which also features a line of critical points separating an ordered and chaotic phase. Note that unlike the EOC for tanh, which increases with increasing σ_w^2, the EOC for SWISH decreases with increasing σ_w^2, which prevents us from examining the impact of large weight variances. Conversely, for small values of σ_w^2, the corresponding value of σ_b^2 becomes so large that we are unable to satisfy the critical detection criteria $\chi = 1$ discussed in the main text.[10] The EOC for SWISH is plotted in fig. 5.8, which shows a computable range of approximately $\sigma_w^2 \in [1.97, 3.4]$. The same figure also shows the accuracy for an $L = 40$ network with SWISH activation function trained along the EOC, demonstrating no deterioration of performance within this range, which confirms the absence of saturation effects. See also [95, 96].
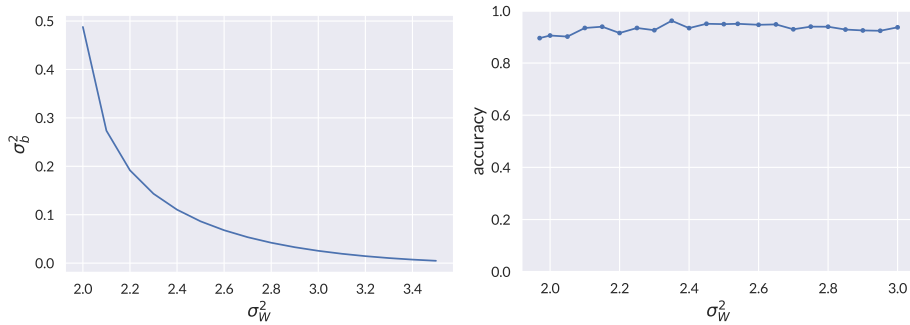

Figure 5.8: (Left) Edge of chaos for SWISH activation function. (Right) Accuracy for a feedforward network with $L = 40$ layers trained on MNIST for 21 equally-spaced points along the SWISH EOC. Over the limited range for which the criticality condition $\chi = 1$ is satisfied, we observe no significant differences in accuracy, though a slightly lower learning rate was used for the left-most two points; we believe this to be due to the large values of σ_b^2 in this regime.

----

[10]We do not claim that the EOC stops beyond this point, rather that it cannot be computed from the central limit method used in [14, 15]. It is conceivable that this could be computed via the NN/QFT correspondence developed in [98], but this has not been attempted for SWISH.

Bibliography

[1] S. Sachdev, *Quantum Phase Transitions*. Cambridge University Press, 2 ed., 2011.

[2] M. Newman and G. Barkema, *Monte Carlo Methods in Statistical Physics*. Clarendon Press, 1999.

[3] K. Binder, *Critical properties from monte carlo coarse graining and renormalization*, *Phys. Rev. Lett.* **47** (Aug, 1981) 693–696.

[4] G. Carleo and M. Troyer, *Solving the quantum many-body problem with artificial neural networks*, *Science* **355** (2017), no. 6325 602–606, [https://www.science.org/doi/pdf/10.1126/science.aag2302].

[5] A. Dawid, J. Arnold, B. Requena, A. Gresch, M. Płodzień, K. Donatella, K. A. Nicoli, P. Stornati, R. Koch, M. Büttner, R. Okuła, G. Muñoz-Gil, R. A. Vargas-Hernández, A. Cervera-Lierta, J. Carrasquilla, V. Dunjko, M. Gabrié, P. Huembeli, E. van Nieuwenburg, F. Vicentini, L. Wang, S. J. Wetzel, G. Carleo, E. Greplová, R. Krems, F. Marquardt, M. Tomza, M. Lewenstein, and A. Dauphin, *Modern applications of machine learning in quantum sciences*, 2022.

[6] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, *Neural Networks* **2** (1989), no. 5 359–366.

[7] M. Srednicki, *Chaos and quantum thermalization*, *Phys. Rev. E* **50** (Aug, 1994) 888–901.

[8] P. Sabella-Garnier, K. Schalm, T. Vakhtel, and J. Zaanen, *Thermalization/relaxation in integrable and free field theories: an operator thermalization hypothesis*, 2019.

[9] T. LeBlond and M. Rigol, *Eigenstate thermalization for observables that break hamiltonian symmetries and its counterpart in interacting integrable systems*, *Phys. Rev. E* **102** (Dec, 2020) 062113.

[10] F. J. Wegner, *Duality in generalized ising models and phase transitions without local order parameters*, *Journal of Mathematical Physics* **12** (1971), no. 10 2259–2272, [https://doi.org/10.1063/1.1665530].

[11] J. B. Kogut, *An introduction to lattice gauge theory and spin systems*, *Rev. Mod. Phys.* **51** (Oct, 1979) 659–713.

[12] E. Fradkin and S. H. Shenker, *Phase diagrams of lattice gauge theories with higgs fields*, *Phys. Rev. D* **19** (Jun, 1979) 3682–3697.

[13] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, *A high-bias, low-variance introduction to machine learning for physicists*, *Physics Reports* **810** (may, 2019) 1–124.

[14] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, *Exponential expressivity in deep neural networks through transient chaos*, .

[15] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, *Deep Information Propagation*, *arXiv e-prints* (Nov., 2016) arXiv:1611.01232, [arXiv:1611.01232].

[16] A. Bukva, P. Sabella-Garnier, and K. Schalm, *Operator thermalization vs eigenstate thermalization*, 2019.

[17] T. Langen, R. Geiger, and J. Schmiedmayer, *Ultracold atoms out of equilibrium*, *Annual Review of Condensed Matter Physics* **6** (2015), no. 1 201–217, [https://doi.org/10.1146/annurev-conmatphys-031214-014548].

[18] A. M. Kaufman, M. E. Tai, A. Lukin, M. Rispoli, R. Schittko, P. M. Preiss, and M. Greiner, *Quantum thermalization through entanglement in an isolated many-body system*, *Science* **353** (2016), no. 6301 794–800, [https://science.sciencemag.org/content/353/6301/794.full.pdf].

[19] P. Bordia, H. Lüschen, S. Scherg, S. Gopalakrishnan, M. Knap, U. Schneider, and I. Bloch, *Probing slow relaxation and many-body localization in two-dimensional quasiperiodic systems*, *Phys. Rev. X* **7** (Nov, 2017) 041047.

[20] G. Bentsen, T. Hashizume, A. S. Buyskikh, E. J. Davis, A. J. Daley, S. S. Gubser, and M. Schleier-Smith, *Treelike interactions and fast scrambling with cold atoms*, *Phys. Rev. Lett.* **123** (Sep, 2019) 130601.

[21] H. Labuhn, D. Barredo, S. Ravets, S. De Léséleuc, T. Macrì, T. Lahaye, and A. Browaeys, *Tunable two-dimensional arrays of single rydberg atoms for realizing quantum ising models*, *Nature* **534** (2016), no. 7609 667.

[22] C. J. Turner, A. A. Michailidis, D. A. Abanin, M. Serbyn, and Z. Papić, *Weak ergodicity breaking from quantum many-body scars*, *Nature Physics* **14** (2018), no. 7 745–749.

[23] A. Dymarsky, N. Lashkari, and H. Liu, *Subsystem ETH*, *Phys. Rev.* **E97** (2018) 012140, [arXiv:1611.08764].

[24] E. M. Brehm, D. Das, and S. Datta, *Probing thermality beyond the diagonal*, *Phys. Rev.* **D98** (2018), no. 12 126015, [arXiv:1804.07924].

[25] A. Romero-Bermúdez, P. Sabella-Garnier, and K. Schalm, *A Cardy formula for off-diagonal three-point coefficients; or, how the geometry behind the horizon gets disentangled*, *JHEP* **09** (2018) 005, [arXiv:1804.08899].

[26] Y. Hikida, Y. Kusuki, and T. Takayanagi, *Eigenstate thermalization hypothesis and modular invariance of two-dimensional conformal field theories*, Phys. Rev. **D98** (2018), no. 2 026003, [arXiv:1804.09658].

[27] T. Anous and J. Sonner, *Phases of scrambling in eigenstates*, [arXiv:1903.03143].

[28] P. Nayak, J. Sonner, and M. Vielma, *Eigenstate Thermalisation in the conformal Sachdev-Ye-Kitaev model: an analytic approach*, [arXiv:1903.00478].

[29] M. Besken, S. Datta, and P. Kraus, *Quantum thermalization and virasoro symmetry*, 2019.

[30] J. M. Deutsch, *Quantum statistical mechanics in a closed system*, Phys. Rev. A **43** (Feb, 1991) 2046–2049.

[31] M. Srednicki, *Chaos and quantum thermalization*, Phys. Rev. E **50** (Aug, 1994) 888–901.

[32] V. I. Yukalov, *Equilibration and thermalization in finite quantum systems*, Laser Phys. Lett. **8** (2011) 485–507, [arXiv:1201.2781].

[33] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, *From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics*, Adv. Phys. **65** (2016), no. 3 239–362, [arXiv:1509.06411].

[34] S. Grozdanov and J. Polonyi, *Dynamics of the electric current in an ideal electron gas: A sound mode inside the quasiparticles*, Phys. Rev. **D92** (2015), no. 6 065009, [arXiv:1501.06620].

[35] I. Amado, B. Sundborg, L. Thorlacius, and N. Wintergerst, *Black holes from large N singlet models*, JHEP **03** (2018) 075, [arXiv:1712.06963].

[36] D. E. Parker, X. Cao, A. Avdoshkin, T. Scaffidi, and E. Altman, *A Universal Operator Growth Hypothesis*, [arXiv:1812.08657].

[37] M. Medenjak, B. Buca, and D. Jaksch, *The isolated heisenberg magnet as a quantum time crystal*, [1905.08266].

[38] S. Banerjee, J. Engelsöy, J. Larana-Aragon, B. Sundborg, L. Thorlacius, and N. Wintergerst, *Quenched coupling, entangled equilibria, and correlated composite operators: a tale of two O(N) models*, [arXiv:1903.12242].

[39] B. Craps, M. De Clerck, O. Evnin, and S. Khetrapal, *Energy level splitting for weakly interacting bosons in a harmonic trap*, [arXiv:1903.04974].

[40] P. Sabella-Garnier, K. Schalm, T. Vakhtel, and J. Zaanen, *Thermalization/Relaxation in integrable and free field theories: an Operator Thermalization Hypothesis*, [arXiv:1906.02597].

[41] D. Fioretto and G. Mussardo, *Quantum quenches in integrable field theories*, *New Journal of Physics* **12** (may, 2010) 055015.

[42] A. C. Cassidy, C. W. Clark, and M. Rigol, *Generalized thermalization in an integrable lattice system*, *Phys. Rev. Lett.* **106** (Apr, 2011) 140405.

[43] V. Alba, *Eigenstate thermalization hypothesis and integrability in quantum spin chains*, *Phys. Rev. B* **91** (Apr, 2015) 155123.

[44] J. Riddell and M. P. Müller, *Generalized eigenstate typicality in translation-invariant quasifree fermionic models*, *Phys. Rev. B* **97** (Jan, 2018) 035129.

[45] S. Byju, K. Lochan, and S. Shankaranarayanan, *Generalized thermalization in quenched free Fermionic models*, [`arXiv:1808.07742`].

[46] P. Banerjee, A. Gaikwad, A. Kaushal, and G. Mandal, *Quantum quench and thermalization to GGE in arbitrary dimensions and the odd-even effect*, [`arXiv:1910.02404`].

[47] M. C. Bañuls, J. I. Cirac, and M. B. Hastings, *Strong and weak thermalization of infinite nonintegrable quantum systems*, *Phys. Rev. Lett.* **106** (Feb, 2011) 050405.

[48] A. Bukva, K. Schalm, and J. Zaanen, *Replicating Higgs fields in Ising gauge theory: The registry order*, May, 2023.

[49] X. G. Wen, *Mean-field theory of spin-liquid states with finite energy gap and topological orders*, *Phys. Rev. B* **44** (Aug, 1991) 2664–2672.

[50] F. Alexander Bais, P. van Driel, and M. de Wild Propitius, *Quantum symmetries in discrete gauge theories*, *Physics Letters B* **280** (1992), no. 1 63–70.

[51] T. Hansson, V. Oganesyan, and S. Sondhi, *Superconductors are topologically ordered*, *Annals of Physics* **313** (2004), no. 2 497–538.

[52] T. Senthil and M. P. A. Fisher, *Z_2 gauge theory of electron fractionalization in strongly correlated systems*, *Phys. Rev. B* **62** (Sep, 2000) 7850–7881.

[53] H. V. Kruis, I. P. McCulloch, Z. Nussinov, and J. Zaanen, *Geometry and topological order in the luttinger liquid state*, *Europhysics Letters* **65** (feb, 2004) 512.

[54] Y. Zhang, E. Demler, and S. Sachdev, *Competing orders in a magnetic field: Spin and charge order in the cuprate superconductors*, *Phys. Rev. B* **66** (Sep, 2002) 094501.

[55] A. Kitaev, *Fault-tolerant quantum computation by anyons*, *Annals of Physics* **303** (2003), no. 1 2–30.

[56] P. E. Lammert, D. S. Rokhsar, and J. Toner, *Topology and nematic ordering, Phys. Rev. Lett.* **70** (Mar, 1993) 1650–1653.

[57] K. Liu, J. Nissinen, R.-J. Slager, K. Wu, and J. Zaanen, *Generalized liquid crystals: Giant fluctuations and the vestigial chiral order of i, o, and t matter, Phys. Rev. X* **6** (Oct, 2016) 041025.

[58] D. A. Huse and S. Leibler, *Are sponge phases of membranes experimental gauge-higgs systems?, Phys. Rev. Lett.* **66** (Jan, 1991) 437–440.

[59] E. Babaev, A. Sudbø, and N. W. Ashcroft, *A superconductor to superfluid phase transition in liquid metallic hydrogen, Nature* **431** (Oct., 2004) 666–668.

[60] E. Babaev and N. W. Ashcroft, *Violation of the london law and onsager–feynman quantization in multicomponent superconductors, Nature Physics* **3** (June, 2007) 530–533.

[61] E. Babaev, J. Carlström, and M. Speight, *Type-1.5 superconducting state from an intrinsic proximity effect in two-band superconductors, Phys. Rev. Lett.* **105** (Aug, 2010) 067003.

[62] T. Senthil and O. Motrunich, *Microscopic models for fractionalized phases in strongly correlated systems, Phys. Rev. B* **66** (Nov, 2002) 205104.

[63] A. Polyakov, *Quark confinement and topology of gauge theories, Nuclear Physics B* **120** (Mar., 1977) 429–458.

[64] J. Nissinen, K. Liu, R.-J. Slager, K. Wu, and J. Zaanen, *Classification of point-group-symmetric orientational ordering tensors, Phys. Rev. E* **94** (Aug, 2016) 022701.

[65] K. Liu, J. Nissinen, J. de Boer, R.-J. Slager, and J. Zaanen, *Hierarchy of orientational phases and axial anisotropies in the gauge theoretical description of generalized nematic liquid crystals, Phys. Rev. E* **95** (Feb, 2017) 022704.

[66] K. Liu, J. Nissinen, Z. Nussinov, R.-J. Slager, K. Wu, and J. Zaanen, *Classification of nematic order in 2 + 1 dimensions: Dislocation melting and o(2)/Z_N lattice gauge theory, Phys. Rev. B* **91** (Feb, 2015) 075103.

[67] J. Zaanen, Y. Liu, Y.-W. Sun, and K. Schalm, *Holographic Duality in Condensed Matter Physics.* Cambridge University Press, Cambridge, 2015.

[68] S. A. Hartnoll, A. Lucas, and S. Sachdev, *Holographic quantum matter*, Mar., 2018.

[69] C. Holzhey, F. Larsen, and F. Wilczek, *Geometric and renormalized entropy in conformal field theory, Nucl. Phys. B* **424** (1994) 443–467, [hep-th/9403108].

[70] G. Evenbly and G. Vidal, *Quantum Criticality with the Multi-scale Entanglement Renormalization Ansatz*, pp. 99–130. No. arXiv:1109.5334. arXiv, Berlin, Heidelberg, 2013. `arXiv:1109.5334`.

[71] G. Carleo and M. Troyer, *Solving the quantum many-body problem with artificial neural networks*, Science **355** (2017), no. 6325 602–606, [`https://www.science.org/doi/pdf/10.1126/science.aag2302`].

[72] F. Vicentini et al., *NetKet 3: Machine Learning Toolbox for Many-Body Quantum Systems*, [`arXiv:2112.10526`].

[73] H.-Q. Shi, X.-Y. Sun, and D.-F. Zeng, *Neural-network quantum state of transverse-field ising model*, Communications in Theoretical Physics **71** (nov, 2019) 1379, [`arXiv:1905.11066`].

[74] E. Fradkin and L. Susskind, *Order and disorder in gauge systems and magnets*, Phys. Rev. D **17** (May, 1978) 2637–2658.

[75] E. Fradkin and S. H. Shenker, *Phase diagrams of lattice gauge theories with Higgs fields*, Physical Review D **19** (June, 1979) 3682–3697.

[76] J. B. Kogut, *An introduction to lattice gauge theory and spin systems*, Rev. Mod. Phys. **51** (Oct, 1979) 659–713.

[77] D. Luo, G. Carleo, B. K. Clark, and J. Stokes, *Gauge equivariant neural networks for quantum lattice gauge theories*, Physical Review Letters **127** (Dec., 2021) 276402, [`arXiv:2012.05232`].

[78] S. Ghosh, R. M. Soni, and S. P. Trivedi, *On The Entanglement Entropy For Gauge Theories*, JHEP **09** (2015) 069, [`arXiv:1501.02593`].

[79] H. Matsueda, *Entanglement Entropy and Entanglement Spectrum for Two-Dimensional Classical Spin Configuration*, arXiv e-prints (Sept., 2011) arXiv:1109.0104, [`arXiv:1109.0104`].

[80] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racanière, D. J. Rezende, and P. E. Shanahan, *Equivariant flow-based sampling for lattice gauge theory*, Physical Review Letters **125** (sep, 2020).

[81] M. Medvidovic, J. Carrasquilla, L. E. Hayward, and B. Kulchytskyy, *Generative models for sampling of lattice field theories*, 2021.

[82] A. Bukva, J. de Gier, K. T. Grosvenor, R. Jefferson, K. Schalm, and E. Schwander, *Criticality versus uniformity in deep neural networks*, 2023.

[83] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[84] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, *Zero-Shot Text-to-Image Generation*, *arXiv e-prints* (Feb., 2021) arXiv:2102.12092, [`arXiv:2102.12092`].

[85] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language Models are Few-Shot Learners*, *arXiv e-prints* (May, 2020) arXiv:2005.14165, [`arXiv:2005.14165`].

[86] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, *Highly accurate protein structure prediction with alphafold*, *Nature* **596** (2021), no. 7873 583–589.

[87] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, *Mastering atari, go, chess and shogi by planning with a learned model*, *Nature* **588** (2020), no. 7839 604–609.

[88] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, *On the Expressive Power of Deep Neural Networks*, *arXiv e-prints* (June, 2016) arXiv:1606.05336, [`arXiv:1606.05336`].

[89] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised learning techniques*. O'Reilly Media, Incorporated, 2019.

[90] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington, *Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks*, .

[91] 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*, .

[92] J. Erdmenger, K. T. Grosvenor, and R. Jefferson, *Towards quantifying information flows: relative entropy in deep neural networks and the renormalization group*, *SciPost Phys.* **12** (2022), no. 1 041, [`arXiv:2107.06898`].

[93] D. Hernandez and T. B. Brown, *Measuring the Algorithmic Efficiency of Neural Networks*, *arXiv e-prints* (May, 2020) arXiv:2005.04305, [`arXiv:2005.04305`].

[94] R. Livni, S. Shalev-Shwartz, and O. Shamir, *On the Computational Efficiency of Training Neural Networks*, *arXiv e-prints* (Oct., 2014) arXiv:1410.1141, [`arXiv:1410.1141`].

[95] S. Hayou, A. Doucet, and J. Rousseau, *On the Selection of Initialization and Activation Function for Deep Neural Networks*, *arXiv e-prints* (May, 2018) arXiv:1805.08266, [`arXiv:1805.08266`].

[96] S. Hayou, A. Doucet, and J. Rousseau, *On the impact of the activation function on deep neural networks training*, in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 2672–2680, PMLR, 09–15 Jun, 2019. `arXiv:1902.06853`.

[97] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory*, [`arXiv:2106.10165`].

[98] K. T. Grosvenor and R. Jefferson, *The edge of chaos: quantum field theory and deep neural networks*, *SciPost Phys.* **12** (2022), no. 3 081, [`arXiv:2109.13247`].

[99] R. Jefferson, "Criticality in deep neural nets." `https://rojefferson.blog/2020/06/19/criticality-in-deep-neural-nets/`, 2020. Accessed: 2022-11-17.