DEEP INFORMATION PROPAGATION

Samuel S. Schoenholz* Google Brain
Justin Gilmer* Google Brain
Surya Ganguli Stanford University
Jascha Sohl-Dickstein Google Brain

ABSTRACT

We study the behavior of untrained neural networks whose weights and biases are randomly distributed using mean field theory. We show the existence of depth scales that naturally limit the maximum depth of signal propagation through these random networks. Our main practical result is to show that random networks may be trained precisely when information can travel through them. Thus, the depth scales that we identify provide bounds on how deep a network may be trained for a specific choice of hyperparameters. As a corollary to this, we argue that in networks at the edge of chaos, one of these depth scales diverges. Thus arbitrarily deep networks may be trained only sufficiently close to criticality. We show that the presence of dropout destroys the order-to-chaos critical point and therefore strongly limits the maximum trainable depth for random networks. Finally, we develop a mean field theory for backpropagation and we show that the ordered and chaotic phases correspond to regions of vanishing and exploding gradient respectively.

1 INTRODUCTION

Deep neural network architectures have become ubiquitous in machine learning. The success of deep networks is due to the fact that they are highly expressive (Montufar et al., 2014) while simultaneously being relatively easy to optimize (Choromanska et al., 2015; Goodfellow et al., 2014) with strong generalization properties (Recht et al., 2015). Consequently, developments in machine learning often accompany improvements in our ability to train increasingly deep networks. Despite this, designing novel network architectures is frequently equal parts art and science. This is, in part, because a general theory for neural networks that might inform design decisions has lagged behind the feverish pace of design.

A pair of recent papers (Poole et al., 2016; Raghu et al., 2016) demonstrated that random neural networks are exponentially expressive in their depth. Central to their approach was the consideration of networks after random initialization, whose weights and biases were i.i.d. Gaussian distributed. In particular the paper by Poole et al. (2016) developed a "mean field" formalism for treating wide, untrained, neural networks. They showed that these mean field networks exhibit an order-to-chaos transition as a function of the weight and bias variances. Notably the mean field formalism is not closely tied to a specific choice of activation function or loss.

In this paper, we demonstrate the existence of several characteristic "depth" scales that emerge naturally and control signal propagation in these random networks. We then show that one of these depth scales, ξ_c, diverges at the boundary between order and chaos. This result is insensitive to many architectural decisions (such as choice of activation function) and will generically be true at any order-to-chaos transition. We then extend these results to include dropout and we show that even small amounts of dropout destroys the order-to-chaos critical point and consequently removes the divergence in ξ_c. Together these results bound the depth to which signal may propagate through random neural networks.

We then develop a corresponding mean field model for gradients and we show that a duality exists between the forward propagation of signals and the backpropagation of gradients. The ordered and chaotic phases that Poole et al. (2016) identified correspond to regions of vanishing and exploding gradients, respectively. We demonstrate the validity of this mean field theory by computing gradients of random networks on MNIST. This provides a formal explanation of the 'vanishing gradients'

*Work done as a member of the Google Brain Residency program (g.co/brainresidency)

phenomenon that has long been observed in neural networks (Bengio et al., 1993). We continue to show that the covariance between two gradients is controlled by the same depth scale that limits correlated signal propagation in the forward direction.

Finally, we hypothesize that a necessary condition for a random neural network to be trainable is that information should be able to pass through it. Thus, the depth-scales identified here bound the set of hyperparameters that will lead to successful training. To test this ansatz we train ensembles of deep, fully connected, feed-forward neural networks of varying depth on MNIST and CIFAR10, with and without dropout. Our results confirm that neural networks are trainable precisely when their depth is not much larger than ξ_c. This result is dataset independent and is, therefore, a universal function of network architecture.

A corollary of these result is that asymptotically deep neural networks should be trainable provided they are initialized sufficiently close to the order-to-chaos transition. The notion of "edge of chaos" initialization has been explored previously. Such investigations have been both direct as in Bertschinger et al. (2005); Glorot & Bengio (2010) or indirect, through initialization schemes that favor deep signal propagation such as batch normalization (Ioffe & Szegedy, 2015), orthogonal matrix initialization (Saxe et al., 2014), random walk initialization (Sussillo & Abbott, 2014), composition kernels (Daniely et al., 2016), or residual network architectures (He et al., 2015). The novelty of the work presented here is two-fold. First, our framework predicts the depth at which networks may be trained even far from the order-to-chaos transition. While a skeptic might ask when it would be profitable to initialize a network far from criticality, we respond by noting that there are architectures (such as neural networks with dropout) where no critical point exists and so this more general framework is needed. Second, our work provides a formal, as opposed to intuitive, explanation for why very deep networks can only be trained near the edge of chaos.

2 Background

We begin by recapitulating the mean-field formalism developed in Poole et al. (2016). Consider a fully-connected, untrained, feed-forward, neural network of depth L with layer width N_l and some nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$. Since this is an untrained neural network we suppose that its weights and biases are respectively i.i.d. as $W^l_{ij} \sim N(0, \sigma_w^2/N_l)$ and $b^l_i \sim N(0, \sigma_b^2)$. Notationally we set z^l_i to be the pre-activations of the lth layer and y^{l+1}_i to be the activations of that layer. Finally, we take the input to the network to be $y^0_i = x_i$. The propagation of a signal through the network is described by the pair of equations,

$$z^l_i = \sum_j W^l_{ij} y^l_j + b^l_i \qquad\qquad y^{l+1}_i = \phi(z^l_i). \tag{1}$$

Since the weights and biases are randomly distributed, these equations define a probability distribution on the activations and pre-activations over an ensemble of untrained neural networks. The "mean-field" approximation is then to replace z^l_i by a Gaussian whose first two moments match those of z^l_i. For the remainder of the paper we will take the mean field approximation as given.

Consider first the evolution of a single input, $x_{i;a}$, as it evolves through the network (as quantified by $y^l_{i;a}$ and $z^l_{i;a}$). Since the weights and biases are independent with zero mean, the first two moments of the pre-activations in the same layer will be,

$$\mathbb{E}[z^l_{i;a}] = 0 \qquad\qquad \mathbb{E}[z^l_{i;a} z^l_{j;a}] = q^l_{aa} \delta_{ij} \tag{2}$$

where δ_{ij} is the Kronecker delta. Here q^l_{aa} is the variance of the pre-activations in the lth layer due to an input $x_{i;a}$ and it is described by the recursion relation,

$$q^l_{aa} = \sigma_w^2 \int \mathcal{D}z \phi^2 \left(\sqrt{q^{l-1}_{aa}} z \right) + \sigma_b^2 \tag{3}$$

where $\int \mathcal{D}z = \frac{1}{\sqrt{2\pi}} \int dz e^{-\frac{1}{2}z^2}$ is the measure for a standard Gaussian distribution. Together these equations completely describe the evolution of a single input through a mean field neural network. For any choice of σ_w^2 and σ_b^2 with bounded ϕ, eq. 3 has a fixed point at $q^* = \lim_{l \to \infty} q^l_{aa}$.

The propagation of a pair of signals, $x^0_{i;a}$ and $x^0_{i;b}$, through this network can be understood similarly. Here the mean pre-activations are trivially the same as in the single-input case. The independence

of the weights and biases implies that the covariance between different pre-activations in the same layer will be given by, $\mathbb{E}[z^l_{i;a} z^l_{j;b}] = q^l_{ab}\delta_{ij}$. The covariance, q^l_{ab}, will be given by the recurrence relation,

$$q^l_{ab} = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1)\phi(u_2) + \sigma_b^2 \tag{4}$$

where $u_1 = \sqrt{q^{l-1}_{aa}} z_1$ and $u_2 = \sqrt{q^{l-1}_{bb}} \left(c^{l-1}_{ab} z_1 + \sqrt{1-(c^{l-1}_{ab})^2} z_2 \right)$, with $c^l_{ab} = q^l_{ab}/\sqrt{q^l_{aa} q^l_{bb}}$, are Gaussian approximations to the pre-activations in the preceding layer with the correct covariance matrix. Moreover c^l_{ab} is the correlation between the two inputs after l layers.

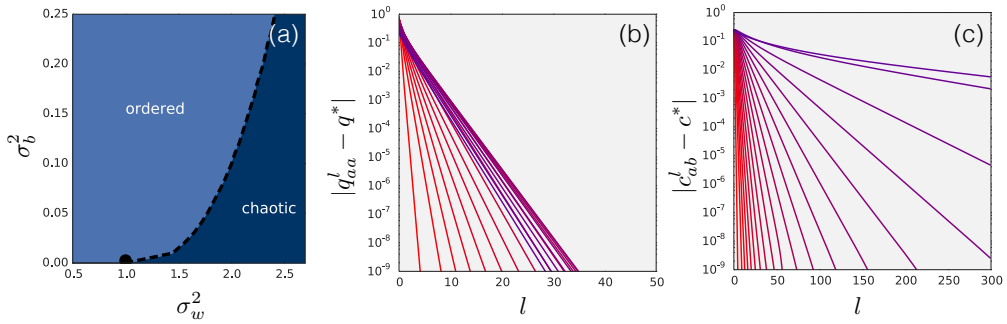

Figure 1: Mean field criticality. (a) The mean field phase diagram showing the boundary between ordered and chaotic phases as a function of σ_w^2 and σ_b^2. (b) The residual $|q^* - q^l_{aa}|$ as a function of depth on a log-scale with $\sigma_b^2 = 0.05$ and σ_w^2 from 0.01 (red) to 1.7 (purple). Clear exponential behavior is observed. (c) The residual $|c^* - c^l_{ab}|$ as a function of depth on a log-scale. Again, the exponential behavior is clear. The same color scheme is used here as in (b).

Examining eq. 4 it is clear that $c^* = 1$ is a fixed point of the recurrence relation. To determine whether or not the $c^* = 1$ is an attractive fixed point the quantity,

$$\chi_1 = \frac{\partial c^l_{ab}}{\partial c^{l-1}_{ab}} = \sigma_w^2 \int \mathcal{D}z \left[\phi'\left(\sqrt{q^*}z\right)\right]^2 \tag{5}$$

is introduced. Poole et al. (2016) note that the $c^* = 1$ fixed point is stable if $\chi_1 < 1$ and is unstable otherwise. Thus, $\chi_1 = 1$ represents a critical line separating an ordered phase (in which $c^* = 1$ and all inputs end up asymptotically correlated) and a chaotic phase (in which $c^* < 1$ and all inputs end up asymptotically decorrelated). For the case of $\phi = \tanh$, the phase diagram in fig. 1 (a) is observed.

3 Asymptotic Expansions and Depth Scales

Our first contribution is to demonstrate the existence of two depth-scales that arise naturally within the framework of mean field neural networks. Motivating the existence of these depth-scales, we iterate eq. 3 and 4 until convergence for many values of σ_w^2 between 0.1 and 3.0 and with $\sigma_b^2 = 0.05$ starting with $q^0_{aa} = q^0_{bb} = 0.8$ and $c^0_{ab} = 0.6$. We see, in fig. 1 (b) and (c), that the manner in which both q^l_{aa} approaches q^* and c^l_{ab} approaches c^* is exponential over many orders of magnitude. We therefore anticipate that asymptotically $|q^l_{aa} - q^*| \sim e^{-l/\xi_q}$ and $|c^l_{ab} - c^*| \sim e^{-l/\xi_c}$ for sufficiently large l. Here, ξ_q and ξ_c define depth-scales over which information may propagate about the magnitude of a single input and the correlation between two inputs respectively.

We will presently prove that q^l_{aa} and c^l_{ab} are asymptotically exponential. In both cases we will use the same fundamental strategy wherein we expand one of the recurrence relations (either eq. 3 or eq. 4) about its fixed point to get an approximate "asymptotic" recurrence relation. We find that this asymptotic recurrence relation in turn implies exponential decay towards the fixed point over a depth-scale, ξ_x.

We first analyze eq. 3 and identify a depth-scale at which information about a single input may propagate. Let $q^l_{aa} = q^* + \epsilon^l$. By construction so long as $\lim_{l\to\infty} q^l_{aa} = q^*$ exists it follows that

$\epsilon^l \to 0$ as $l \to \infty$. Eq. 3 may be expanded to lowest order in ϵ^l to arrive at an asymptotic recurrence relation (see Appendix 7.1),

$$\epsilon^{l+1} = \epsilon^l \left[\chi_1 + \sigma_w^2 \int \mathcal{D}z \phi'' \left(\sqrt{q^*} z \right) \phi \left(\sqrt{q^*} z \right) \right] + \mathcal{O}\left((\epsilon^l)^2 \right). \tag{6}$$

Notably, the term multiplying ϵ^l is a constant. It follows that for large l the asymptotic recurrence relation has an exponential solution, $\epsilon^l \sim e^{-l/\xi_q}$, with ξ_q given by

$$\xi_q^{-1} = -\log \left[\chi_1 + \sigma_w^2 \int \mathcal{D}z \phi'' \left(\sqrt{q^*} z \right) \phi \left(\sqrt{q^*} z \right) \right]. \tag{7}$$

This establishes ξ_q as a depth scale that controls how deep information from a single input may penetrate into a random neural network.

Next, we consider eq. 4. Using a similar argument (detailed in Appendix 7.2) we can expand about $c_{ab}^l = c^* + \epsilon^l$ to find an asymptotic recurrence relation,

$$\epsilon^{l+1} = \epsilon^l \left[\sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1^*) \phi'(u_2^*) \right] + \mathcal{O}((\epsilon^l)^2). \tag{8}$$

Here $u_1^* = \sqrt{q^*} z_1$ and $u_2^* = \sqrt{q^*}(c^* z_1 + \sqrt{1-(c^*)^2} z_2)$. Thus, once again, we expect that for large l this recurrence will have an exponential solution, $\epsilon^l \sim e^{-l/\xi_c}$, with ξ_c given by

$$\xi_c^{-1} = -\log \left[\sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1^*) \phi'(u_2^*) \right]. \tag{9}$$

In the ordered phase $c^* = 1$ and so $\xi_c^{-1} = -\log \chi_1$. Since the transition between order and chaos occurs when $\chi_1 = 1$ it follows that ξ_c diverges at any order-to-chaos transition so long as q^* and c^* exist.

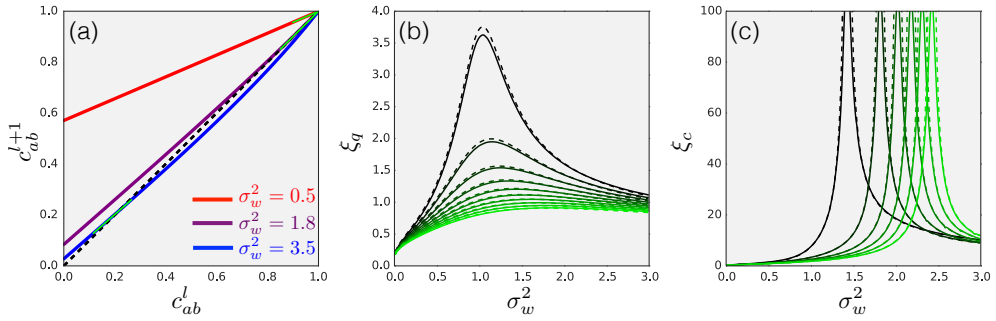

Figure 2: Depth scales. (a) The iterative correlation map showing c_{ab}^{l+1} as a function of c_{ab}^l for three different values of σ_w^2. Green inset lines show the linearization of the iterative map about the critical point, e^{-1/ξ_c}. The three curves show networks far in the ordered regime (red), at the edge of chaos (purple), and deep in the chaotic regime (blue). (b) The depth scale for information propagated in a single input, ξ_q as a function of σ_w^2 for $\sigma_b^2 = 0.01$ (black) to $\sigma_b^2 = 0.3$ (green). Dashed lines show theoretical predictions while solid lines show measurements. (c) The depth scale for correlations between inputs, ξ_c for the same values of σ_b^2. Again dashed lines are the theoretical predictions while solid lines show measurements. Here a clear divergence is observed at the order-to-chaos transition.

These results can be investigated intuitively by plotting c_{ab}^{l+1} vs c_{ab}^l in fig. 2 (a). In the ordered phase there is only a single fixed point, $c_{ab}^l = 1$. In the chaotic regime we see that a second fixed point develops and the $c_{ab}^l = 1$ point becomes unstable. We see that the linearization about the fixed points becomes significantly closer to the trivial map near the order-to-chaos transition.

To test these claims we measure ξ_q and ξ_c directly by iterating the recurrence relations for q_{aa}^l and c_{ab}^l as before with $q_{aa}^0 = q_{bb}^0 = 0.8$ and $c_{ab}^0 = 0.6$. In this case we consider values of σ_w^2 between

0.1 and 3.0 and σ_b^2 between 0.01 and 0.3. For each hyperparameter settings we fit the resulting residuals, $|q_{aa}^l - q^*|$ and $|c_{ab}^l - c^*|$, to exponential functions and infer the depth-scale. We then compare this measured depth-scale to that predicted by the asymptotic expansion. The result of this measurement is shown in fig. 2. In general we see that the agreement is quite good. As expected we see that ξ_c diverges at the critical point.

As observed in Poole et al. (2016) we see that the depth scale for the propagation of information in a single input, ξ_q, is consistently finite and significantly shorter than ξ_c. To understand why this is the case consider eq. 6 and note that for $\tanh$ nonlinearities the second term is always negative. Thus, even as χ_1 approaches 1 we expect $\chi_1 + \sigma_w^2 \int \mathcal{D}z \phi''(\sqrt{q^*}z)\phi(\sqrt{q^*}z)$ to be substantially smaller than 1.

3.1 DROPOUT

The mean field formalism can be extended to include dropout. The main contribution here will be to argue that even infinitesimal amounts of dropout destroys the mean field critical point, and therefore limits the trainable network depth. In the presence of dropout the propagation equation, eq. 1, becomes,

$$z_i^l = \frac{1}{\rho} \sum_j W_{ij}^l p_j^l y_j^l + b_i^l \tag{10}$$

where $p_j \sim \text{Bernoulli}(\rho)$ and ρ is the dropout rate. As is typically the case we have re-scaled the sum by ρ^{-1} so that the mean of the pre-activation is invariant with respect to our choice of dropout rate.

Following a similar procedure to the original mean field calculation consider the fate of two inputs, $x_{i;a}^0$ and $x_{i;b}^0$, as they are propagated through such a random network. We take the dropout masks to be chosen independently for the two inputs mimicking the manner in which dropout is employed in practice. With dropout the diagonal term in the covariance matrix will be (see Appendix 7.3),

$$\bar{q}_{aa}^l = \frac{\sigma_w^2}{\rho} \int \mathcal{D}z \phi^2 \left(\sqrt{\bar{q}_{aa}^{l-1}} z \right) + \sigma_b^2. \tag{11}$$

The variance of a single input with dropout will therefore propagate in an identical fashion to the vanilla case with a re-scaling $\sigma_w^2 \to \sigma_w^2/\rho$. Intuitively, this result implies that, for the case of a single input, the presence of dropout simply increases the effective variance of the weights.

Computing the off-diagonal term of the covariance matrix similarly (see Appendix 7.4),

$$\bar{q}_{ab}^l = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(\bar{u}_1)\phi(\bar{u}_2) + \sigma_b^2 \tag{12}$$

with $\bar{u}_1$, $\bar{u}_2$, and $\bar{c}_{ab}^l$ defined by analogy to the mean field equations without dropout. Here, unlike in the case of a single input, the recurrence relation is identical to the recurrence relation without dropout. To see that $\bar{c}^* = 1$ is no longer a fixed point of these dynamics consider what happens to eq. 12 when we input $\bar{c}^l = 1$. For simplicity, we leverage the short range of ξ_q to replace $\bar{q}_{aa}^l = \bar{q}_{bb}^l = \bar{q}^*$. We find (see Appendix 7.5),

$$\bar{c}_{ab}^{l+1} = 1 - \frac{1-\rho}{\rho \bar{q}^*} \sigma_w^2 \int \mathcal{D}z \phi^2 \left(\sqrt{\bar{q}^*} z \right). \tag{13}$$

The second term is positive for any $\rho < 1$. This implies that if $\bar{c}_{ab}^l = 1$ for any l then $\bar{c}_{ab}^{l+1} < 1$. Thus, $c^* = 1$ is not a fixed point of eq. 12 for any $\rho < 1$. Since eq. 12 is identical in form to eq. 4 it follows that the depth scale for signal propagation with dropout will likewise be given by eq. 9 with the substitutions $q^* \to \bar{q}^*$ and $c^* \to \bar{c}^*$ computed using eq. 11 and eq. 12 respectively. Importantly, since there is no longer a sharp critical point with dropout we do not expect a diverging depth scale.

As in networks without dropout we plot, in fig. 3 (a), the iterative map $\bar{c}_{ab}^{l+1}$ as a function of $\bar{c}_{ab}^l$. Most significantly, we see that the $\bar{c}_{ab}^l = 1$ is no longer a fixed point of the dynamics. Instead, as the dropout rate increases $\bar{c}_{ab}^l$ gets mapped to decreasing values and the fixed point monotonically decreases.

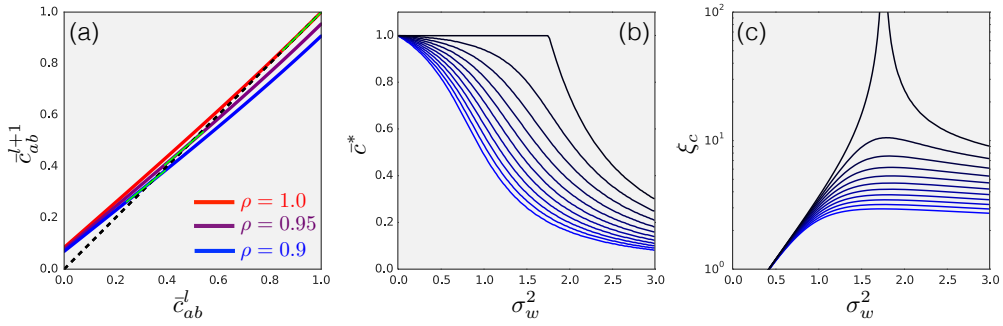

Figure 3: Dropout destroys the critical point, and limits the depth to which information can propagate in a deep network. (a) The iterative correlation map showing $\bar{c}^{l+1}_{ab}$ as a function of $\bar{c}^{l}_{ab}$ for three different values of the dropout rate ρ for networks tuned close to their critical point. Green inset lines show the linearization of the iterative map about the critical point, e^{-1/ξ_c}. (b) The asymptotic value of the correlation map, c^*, as a function of σ_w^2 for different values of dropout from $\rho = 1$ (black) to $\rho = 0.8$ (blue). We see that for all values of dropout except for $\rho = 1$, c^* does not show a sharp transition between an ordered phase and a chaotic phase. (c) The correlation depth scale ξ_c as a function of σ_w^2 for the same values of dropout as in (b). We see here that for all values of ρ except for $\rho = 1$ there is no divergence in ξ_c.

To test these results we plot in fig. 3 (b) the asymptotic correlation, c^*, as a function of σ_w^2 for different values of dropout from $\rho = 0.8$ to $\rho = 1.0$. As expected, we see that for all $\rho < 1$ there is no sharp transition between $c^* = 1$ and $c^* < 1$. Moreover as the dropout rate increases the correlation c^* monotonically decreases. Intuitively this makes sense. Identical inputs passed through two different dropout masks will become increasingly dissimilar as the dropout rate increases. In fig. 3 (c) we show the depth scale, ξ_c, as a function of σ_w^2 for the same range of dropout probabilities. We find that, as predicted, the depth of signal propagation with dropout is drastically reduced and, importantly, there is no longer a divergence in ξ_c. Increasing the dropout rate continues to decrease the correlation depth for constant σ_w^2.

4 Gradient Backpropagation

There is a duality between the forward propagation of signals and the backpropagation of gradients. To elucidate this connection consider the backpropagation equations given a loss E,

$$\frac{\partial E}{\partial W^l_{ij}} = \delta^l_i \phi(z^{l-1}_j) \qquad\qquad \delta^l_i = \phi'(z^l_i) \sum_j \delta^{l+1}_j W^{l+1}_{ji} \tag{14}$$

with the identification $\delta^l_i = \partial E / \partial z^l_i$. Within mean field theory, it is clear that the scale of fluctuations of the gradient of weights in a layer will be proportional to $\mathbb{E}[(\delta^l_i)^2]$ (see appendix 7.6). In contrast to the pre-activations in forward propagation (eq. 1), the δ^l_i will typically not be Gaussian distributed even in the large layer width limit.

Nonetheless, we can work out a recurrence relation for the variance of the error, $\tilde{q}^{\,l}_{aa} = \mathbb{E}[(\delta^l_i)^2]$, leveraging the Gaussian ansatz on the pre-activations. In order to do this, however, we must first make an additional approximation that the weights used during forward propagation are drawn independently from the weights used in backpropagation. This approximation is similar in spirit to the vanilla mean field approximation and is reminiscent of work on feedback alignment (Lillicrap et al., 2014). With this in mind we arrive at the recurrence (see appendix 7.7),

$$\tilde{q}^{\,l}_{aa} = \tilde{q}^{\,l+1}_{aa} \frac{N_{l+1}}{N_l} \chi_1. \tag{15}$$

The presence of χ_1 in the above equation should perhaps not be surprising. In Poole et al. (2016) they show that χ_1 is intimately related to the tangent space of a given layer in mean field neural

networks. We note that the backpropagation recurrence features an explicit dependence on the ratio of widths of adjacent layers of the network, N_{l+1}/N_l. Here we will consider exclusively constant width networks where this factor is unity. For a discussion of the case of unequal layer widths see Glorot & Bengio (2010).

Since χ_1 depends only on the asymptotic q^* it follows that for constant width networks we expect eq. 15 to again have an exponential solution with,

$$\tilde{q}^{\,l}_{aa} = \tilde{q}^{\,L}_{aa} e^{-(L-l)/\xi_\nabla} \qquad\qquad \xi_\nabla^{-1} = -\log \chi_1. \tag{16}$$

Note that here $\xi_\nabla^{-1} = -\log\chi_1$ both above and below the transition. It follows that ξ_∇ can be both positive and negative. We conclude that there should be three distinct regimes for the gradients.

1. In the ordered phase, $\chi_1 < 1$ and so $\xi_\nabla > 0$. We therefore expect gradients to vanish over a depth $|\xi_\nabla|$.
2. At criticality, $\chi_1 \to 1$ and so $\xi_\nabla \to \infty$. Here gradients should be stable regardless of depth.
3. In the chaotic phase, $\chi_1 > 1$ and so $\xi_\nabla < 0$. It follows that in this regime gradients should explode over a depth $|\xi_\nabla|$.

Intuitively these three regimes make sense. To see this, recall that perturbations to a weight in layer l can alternatively be viewed as perturbations to the pre-activations in the same layer. In the ordered phase both the perturbed signal and the unperturbed signal will be asymptotically mapped to the same point and the derivative will be small. In the chaotic phase the perturbed and unperturbed signals will become asymptotically decorrelated and the gradient will be large.

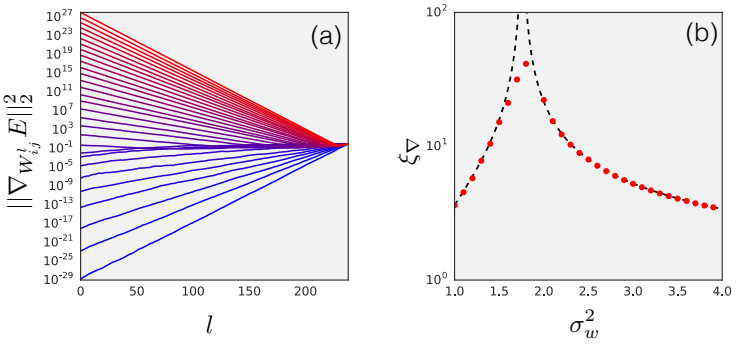

Figure 4: Gradient backpropagation behaves similarly to signal forward propagation. (a) The 2-norm, $||\nabla_{W^l_{ab}} E||_2^2$ as a function of layer, l, for a 240 layer random network with a cross-entropy loss on MNIST. Different values of σ_w^2 from 1.0 (blue) to 4.0 (red) are shown. Clear exponential vanishing / explosion is observed over many orders of magnitude. (b) The depth scale for gradients predicted by theory (dashed line) compared with measurements from experiment (red dots). Similarity between theory and experiment is clear. Deviations near the critical point are primarily due to finite size effects.

To investigate these predictions we construct deep random networks of depth $L = 240$ and layer-width $N_l = 300$. We then consider the cross-entropy loss of these networks on MNIST. In fig. 4 (a) we plot the layer-by-layer 2-norm of the gradient, $||\nabla_{W^l_{ab}} E||_2^2$, as a function of layer, l, for different values of σ_w^2. We see that $||\nabla_{W^l_{ab}} E||_2^2$ behaves exponentially over many orders of magnitude. Moreover, we see that the gradient vanishes in the ordered phase and explodes in the chaotic phase. We test the quantitative predictions of eq. 16 in fig. 4 (b) where we compare $|\xi_\nabla|$ as predicted from theory with the measured depth-scale constructed from exponential fits to the gradient data. Here we see good quantitative agreement between the theoretical predictions from mean field random networks and experimentally realized networks. Together these results suggest that the approximations on the backpropagation equations were representative of deep, wide, random networks.

Finally, we show that the depth scale for correlated signal propagation likewise controls the depth at which information stored in the covariance between gradients can survive. The existence of

consistent gradients across similar samples from a training set ought to be especially important for determining whether or not a given neural network architecture can be trained. To establish this depth-scale first note (see Appendix 7.8) that the covariance between gradients of two different inputs, $x_{i;1}$ and $x_{i;2}$, will be proportional to $(\nabla_{W^l_{ij}} E_a) \cdot (\nabla_{W^l_{ij}} E_b) \sim \mathbb{E}[\delta^l_{i;a}\delta^l_{i;b}] = \tilde{q}^{\,l}_{ab}$ where E_a is the loss evaluated on $x_{i;a}$ and $\delta_{i;a} = \partial E_a / \partial z^l_{i;a}$ are appropriately defined errors.

It can be shown (see Appendix 7.9) that $\tilde{q}^{\,l}_{ab}$ features the recurrence relation,

$$\tilde{q}^{\,l}_{ab} = \tilde{q}^{\,l+1}_{ab} \frac{N_{l+1}}{N_{l+2}} \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1)\phi'(u_2) \tag{17}$$

where u_1 and u_2 are defined similarly as for the forward pass. Expanding asymptotically it is clear that to zeroth order in ϵ^l, $\tilde{q}^l_{ab}$ will have an exponential solution with $\tilde{q}^{\,l}_{ab} = \tilde{q}^{\,L}_{ab} e^{-(L-l)/\xi_c}$ with ξ_c as defined in the forward pass.

5 Experimental Results

Taken together, the results of this paper lead us to the following hypothesis: a necessary condition for a random network to be trained is that information about the inputs should be able to propagate forward through the network, and information about the gradients should be able to propagate backwards through the network. The preceding analysis shows that networks will have this property precisely when the network depth, L, is not much larger than the depth-scale ξ_c. This criterion is data independent and therefore offers a "universal" constraint on the hyperparameters that depends on network architecture alone. We now explore this relationship between depth of signal propagation and network trainability empirically.

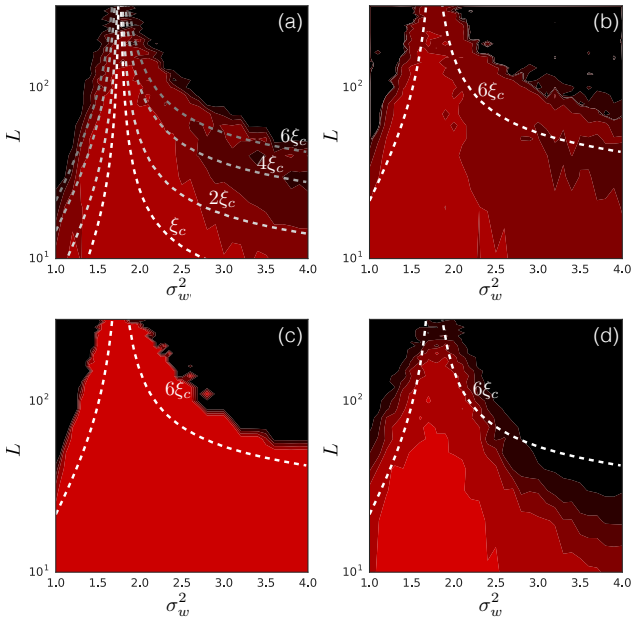

Figure 5: Mean field depth scales control trainable hyperparameters. The training accuracy for neural networks as a function of their depth and initial weight variance, σ_w^2 from a high accuracy (red) to low accuracy (black). In (a) we plot the training accuracy after 200 training steps on MNIST using SGD. Here overlayed in grey dashed lines are different multiples of the depth scale for correlated signal propagation, $n\xi_c$. We plot the accuracy in (b) after 2000 training steps on CIFAR10 using SGD, in (c) after 14000 training steps on MNIST using SGD, and in (d) after 300 training steps on MNIST using RMSPROP. Here we overlay in white dashed lines $6\xi_c$.

To investigate this prediction, we consider random networks of depth $10 \leq L \leq 300$ and $1 \leq \sigma_w^2 \leq 4$ with $\sigma_b^2 = 0.05$. We train these networks using Stochastic Gradient Descent (SGD) and RMSProp

on MNIST and CIFAR10. We use a learning rate of 10^{-3} for SGD when $L \lesssim 200$, 10^{-4} for larger L, and 10^{-5} for RMSProp. These learning rates were selected by grid search between 10^{-6} and 10^{-2} in exponentially spaced steps of size 10. We note that the depth dependence of learning rate was explored in detail in Saxe et al. (2014). In fig. 5 (a)-(d) we color in red the training accuracy that neural networks achieved as a function of σ_w^2 and L for different datasets, training time, and choice of minimizer (see Appendix 7.10 for more comparisons). In all cases the neural networks over-fit the data to give a training accuracy of 100% and test accuracies of 98% on MNIST and 55% on CIFAR10. We emphasize that the purpose of this study is to demonstrate trainability as opposed to optimizing test accuracy.

We now make the connection between the depth scale, ξ_c, and the maximum trainable depth more precise. Given the arguments in the preceding sections we note that if $L = n\xi_c$ then signal through the network will be attenuated by a factor of e^n. To understand how much signal can be lost while still allowing for training, we overlay in fig. 5 (a) curves corresponding to $n\xi_c$ from $n = 1$ to 6. We find that networks appear to be trainable when $L \lesssim 6\xi_c$. It would be interesting to understand why this is the case.

Motivated by this argument in fig. 5 (b)-(d) in white, dashed, overlay we plot twice the predicted depth scale, $6\xi_c$. There is clearly a relationship between the depth of correlated signal propagation and whether or not these networks are trainable. Networks closer to their critical point appear to train more quickly than those further away. Moreover, this relationship has no obvious dependence on dataset, duration of training, or minimizer. We therefore conclude that these bounds on trainable hyperparameters are universal. This in turn implies that to train increasingly deep networks, one must generically be ever closer to criticality.

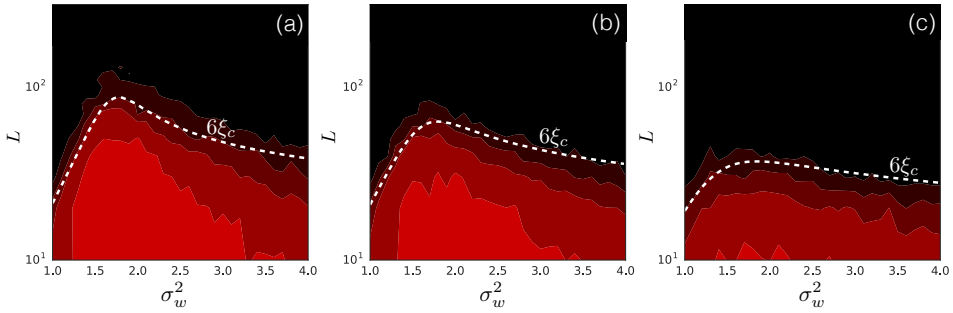

Figure 6: The effect of dropout on trainability. The same scheme as in fig. 5 but with dropout rates of (a) $\rho = 0.99$, (b) $\rho = 0.98$, and (c) $\rho = 0.94$. Even for modest amounts of dropout we see an upper bound on the maximum trainable depth for neural networks. We continue to see good agreement between the prediction of our theory and our experimental training accuracy.

Next we consider the effect of dropout. As we showed earlier, even infinitesimal amounts of dropout disrupt the order-to-chaos phase transition and cause the depth scale to become finite. However, since the effect of a single dropout mask is to simply re-scale the weight variance by $\sigma_w^2 \to \sigma_w^2/\rho$, the gradient magnitude will be stable near criticality, while the input and gradient correlations will not be. This therefore offers a unique opportunity to test whether the relevant depth-scale is $|1/\log \chi_1|$ or ξ_c.

In fig. 6 we repeat the same experimental setup as above on MNIST with dropout rates $\rho = 0.99, 0.98$, and 0.94. We observe, first and foremost, that even extremely modest amounts of dropout limit the maximum trainable depth to about $L = 100$. We additionally notice that the depth-scale, ξ_c, predicts the trainable region accurately for varying amounts of dropout.

6 Discussion

In this paper we have elucidated the existence of several depth-scales that control signal propagation in random neural networks. Furthermore, we have shown that the degree to which a neural network can be trained depends crucially on its ability to propagate information about inputs and gradients

through its full depth. At the transition between order and chaos, information stored in the correlation between inputs can propagate infinitely far through these random networks. This in turn implies that extremely deep neural networks may be trained sufficiently close to criticality. However, our contribution goes beyond advocating for hyperparameter selection that brings random networks to be nearly critical. Instead, we offer a general purpose framework that predicts, at the level of mean field theory, which hyperparameters should allow a network to be trained. This is especially relevant when analyzing schemes like dropout where there is no critical point and which therefore imply an upper bound on trainable network depth.

An alternative perspective as to why information stored in the covariance between inputs is crucial for training can be understood by appealing to the correspondence between infinitely wide Bayesian neural networks and Gaussian Processes (Neal, 2012). In particular the covariance, q^l_{ab}, is intimately related to the kernel of the induced Gaussian Process. It follows that cases in which signal stored in the covariance between inputs may propagate through the network correspond precisely to situations in which the associated Gaussian Process is well defined.

Our work suggests that it may be fruitful to investigate pre-training schemes that attempt to perturb the weights of a neural network to favor information flow through the network. In principle this could be accomplished through a layer-by-layer local criterion for information flow or by selecting the mean and variance in schemes like batch normalization to maximize the covariance depth-scale.

These results suggest that theoretical work on random neural networks can be used to inform practical architectural decisions. However, there is still much work to be done. For instance, the framework developed here does not apply to unbounded activations, such as rectified linear units, where it can be shown that there are phases in which eq. 3 does not have a fixed point. Additionally, the analysis here applies directly only to fully connected feed-forward networks, and will need to be extended to architectures with structured weight matrices such as convolutional networks.

We close by noting that in physics it has long been known that, through renormalization, the behavior of systems near critical points can control their behavior even far from the idealized critical case. We therefore make the somewhat bold hypothesis that a broad class of neural network topologies will be controlled by the fully-connected mean field critical point.

ACKNOWLEDGMENTS

We thank Ben Poole, Jeffrey Pennington, Maithra Raghu, and George Dahl for useful discussions. We are additionally grateful to RocketAI for introducing us to Temporally Recurrent Online Learning and two-dimensional time.

REFERENCES

Y Bengio, Paolo Frasconi, and P Simard. The problem of learning long-term dependencies in recurrent networks. In *Neural Networks, 1993., IEEE International Conference on*, pp. 1183–1188. IEEE, 1993.

Nils Bertschinger, Thomas Natschläger, and Robert A. Legenstein. At the edge of chaos: Real-time computations and self-organized criticality in recurrent neural networks. In L. K. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems 17*, pp. 145–152. MIT Press, 2005.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

A. Daniely, R. Frostig, and Y. Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. *arXiv:1602.05897*, 2016.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pp. 249–256, 2010.

Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. *arXiv:1412.6544*, 2014.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *ArXiv e-prints*, December 2015.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random feedback weights support learning in deep neural networks. *arXiv:1411.0247*, 2014.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2924–2932. Curran Associates, Inc., 2014.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. *arXiv:1606.05340*, June 2016.

M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the expressive power of deep neural networks. *arXiv:1606.05336*, June 2016.

Benjamin Recht, Moritz Hardt, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv:1509.01240*, 2015.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations*, 2014.

David Sussillo and LF Abbott. Random walks: Training very deep nonlinear feed-forward networks with smart initialization. *CoRR, vol. abs/1412.6558*, 2014.

7 APPENDIX

Here we present derivations of results from throughout the paper.

7.1 SINGLE INPUT DEPTH-SCALE

Result:

Consider the recurrence relation for the variance of a single input,

$$q^l_{aa} = \sigma_w^2 \int \mathcal{D}z \phi^2 \left(\sqrt{q^{l-1}_{aa}} z \right) + \sigma_b^2 \tag{18}$$

and a fixed point of the dynamics, q^*. q^l_{aa} can be expanded about the fixed point to yield the asymptotic recurrence relation,

$$\epsilon^{l+1} = \epsilon^l \left[\chi_1 + \sigma_w^2 \int \mathcal{D}z \phi' \left(\sqrt{q^*} z \right) \phi \left(\sqrt{q^*} z \right) \right] + \mathcal{O}\left((\epsilon^l)^2 \right). \tag{19}$$

Derivation:

We begin by first expanding to order ϵ^l,

$$q^* + \epsilon^{l+1} = \sigma_w^2 \int \mathcal{D}z \left[\phi\left(\sqrt{q^* + \epsilon^l} z\right)\right]^2 + \sigma_b^2 \tag{20}$$

$$\approx \sigma_w^2 \int \mathcal{D}z \left[\phi\left(\sqrt{q^*} z + \frac{1}{2}\frac{\epsilon^l z}{\sqrt{q^*}}\right)\right]^2 + \sigma_b^2 \tag{21}$$

$$\approx \sigma_w^2 \int \mathcal{D}z \left[\phi\left(\sqrt{q^*} z\right) + \frac{1}{2}\frac{\epsilon^l z}{\sqrt{q^*}}\phi'\left(\sqrt{q^*} z\right)\right]^2 + \sigma_b^2 + \mathcal{O}((\epsilon^l)^2) \tag{22}$$

$$\approx \sigma_w^2 \int \mathcal{D}z \phi^2\left(\sqrt{q^*} z\right) + \sigma_b^2 + \epsilon^l \frac{\sigma_w^2}{\sqrt{q^*}} \int \mathcal{D}z z \phi\left(\sqrt{q^*} z\right) \phi'\left(\sqrt{q^*} z\right) + \mathcal{O}((\epsilon^l)^2) \tag{23}$$

$$\approx q^* + \epsilon^l \frac{\sigma_w^2}{\sqrt{q^*}} \int \mathcal{D}z z \phi(\sqrt{q^*} z) \phi'\left(\sqrt{q^*} z\right) + \mathcal{O}((\epsilon^l)^2). \tag{24}$$

We therefore arrive at the approximate reccurence relation,

$$\epsilon^{l+1} = \epsilon^l \frac{\sigma_w^2}{\sqrt{q^*}} \int \mathcal{D}z z \phi(\sqrt{q^*} z) \phi'\left(\sqrt{q^*} z\right) + \mathcal{O}((\epsilon^l)^2). \tag{25}$$

Using the identity, $\int \mathcal{D}z z f(z) = \int \mathcal{D}z f'(z)$ we can rewrite this asymptotic recurrence relation as,

$$\epsilon^{l+1} = \epsilon^l \left[\sigma_w^2 \int \mathcal{D}z \left[\phi'\left(\sqrt{q^*} z\right)\right]^2 + \sigma_w^2 \int \mathcal{D}z \phi''\left(\sqrt{q^*} z\right) \phi\left(\sqrt{q^*} z\right)\right] + \mathcal{O}((\epsilon^l)^2) \tag{26}$$

$$= \epsilon^l \left[\chi_1 + \sigma_w^2 \int \mathcal{D}z \phi''\left(\sqrt{q^*} z\right) \phi\left(\sqrt{q^*} z\right)\right] + \mathcal{O}((\epsilon^l)^2) \tag{27}$$

as required.

7.2 Two input depth-scale

Result:

Consider the recurrence relation for the co-variance of two input,

$$q_{ab}^l = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1) \phi(u_2) + \sigma_b^2, \tag{28}$$

a correlation between the inputs, $c_{ab}^l = q_{ab}^l / \sqrt{q_{aa}^l q_{bb}^l}$, and a fixed point of the dynamics, c^*. c_{ab}^l can be expanded about the fixed point to yield the asymptotic recurrence relation,

$$\epsilon^{l+1} = \epsilon^l \left[\sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1) \phi'(u_2)\right] + \mathcal{O}\left((\epsilon^l)^2\right). \tag{29}$$

Derivation:

Since the relaxation of q_{aa}^l and q_{bb}^l to q^* occurs much more quickly than the convergence of q_{ab}^l we approximate $q_{aa}^l = q_{bb}^l = q^*$ as in Poole et al. (2016). We therefore consider the perturbation $q_{ab}^l / q^* = c_{ab}^l = c^* + \epsilon^l$. It follows that we may make the approximation,

$$u_2^l = \sqrt{q^*}\left(c_{ab}^l z_1 + \sqrt{1 - (c_{ab}^l)^2} z_2\right) \tag{30}$$

$$\approx \sqrt{q^*}\left(c^* z_1 + \sqrt{1 - (c^*)^2 - 2c^*\epsilon^l} z_2\right) + \sqrt{q^*}\epsilon^l z_1 + \mathcal{O}(\epsilon^2) \tag{31}$$

$$\tag{32}$$

We now consider the case where $c^* < 1$ and $c^* = 1$ separately; we will later show that these two results agree with one another. First we consider the case where $c^* < 1$ in which case we may safely expand the above equation to get,

$$u_2^l = \sqrt{q^*}\left(c^* z_1 + \sqrt{1 - (c^*)^2} z_2\right) + \sqrt{q^*}\epsilon^l \left(z_1 - \frac{c^*}{\sqrt{1 - (c^*)^2}} z_2\right) + \mathcal{O}(\epsilon^2). \tag{33}$$

This allows us to in turn approximate the recurrence relation,

$$c_{ab}^{l+1} = \frac{\sigma_w^2}{q^*}\int \mathcal{D}z_1\mathcal{D}z_2\phi(u_1^*)\phi(u_2^l) + \sigma_b^2 \tag{34}$$

$$\approx \frac{\sigma_w^2}{q^*}\int \mathcal{D}z_1\mathcal{D}z_2\phi(u_1^*)\left[\phi(u_2^*) + \sqrt{q^*}\epsilon^l\left(z_1 - \frac{c^*}{\sqrt{1-(c^*)^2}}z_2\right)\phi'(u_2^*)\right] + \sigma_b^2 + \mathcal{O}(\epsilon^2) \tag{35}$$

$$= c^* + \frac{\sigma_w^2}{\sqrt{q^*}}\epsilon^l\int \mathcal{D}z_1\mathcal{D}z_2\left(z_1 - \frac{c^*}{\sqrt{1-(c^*)^2}}z_2\right)\phi(u_1^*)\phi'(u_2^*) \tag{36}$$

$$= c^* + \frac{\sigma_w^2}{\sqrt{q^*}}\epsilon^l\left[\int \mathcal{D}z_1\mathcal{D}z_2 z_1\phi(u_1^*)\phi'(u_2^*) - \frac{c^*}{\sqrt{1-(c^*)^2}}\int \mathcal{D}z_1\mathcal{D}z_2 z_2\phi(u_1^*)\phi'(u_2^*)\right] \tag{37}$$

$$= c^* + \sigma_w^2\epsilon^l\left[\int \mathcal{D}z_1\mathcal{D}z_2(\phi'(u_1^*)\phi'(u_2^*) + c^*\phi(u_1^*)\phi''(u_2^*)) - c^*\int \mathcal{D}z_1\mathcal{D}z_2\phi(u_1^*)\phi''(u_2^*)\right] \tag{38}$$

$$= c^* + \sigma_w^2\epsilon^l\int \mathcal{D}z_1\mathcal{D}z_2\phi'(u_1^*)\phi'(u_2^*). \tag{39}$$

where u_1^* and u_2^* are appropriately defined asymptotic random variables. This leads to the asymptotic recurrence relation,

$$\epsilon^{l+1} = \sigma_w^2\epsilon^l\int \mathcal{D}z_1\mathcal{D}z_2\phi'(u_1^*)\phi'(u_2^*) \tag{40}$$

as required.

We now consider the case where $c^* = 1$ and $c_{ab}^l = 1 - \epsilon^l$. In this case the expansion of u_2^l will become,

$$u_2^l = \sqrt{q^*}z_1 + \sqrt{2q^*\epsilon^l}z_2 - \sqrt{q^*}\epsilon^l z_1 + \mathcal{O}(\epsilon^{3/2}) \tag{41}$$

and so the lowest order correction is of order $\mathcal{O}(\sqrt{\epsilon^l})$ as opposed to $\mathcal{O}(\epsilon^l)$. As usual we now expand the recurrence relation, noting that $u_2^* = u_1^*$ is independent of z_2 when $c^* = 1$ to find,

$$c_{ab}^{l+1} = \frac{\sigma_w^2}{q^*}\int \mathcal{D}z_1\mathcal{D}z_2\phi(u_1^*)\phi(u_2^l) + \sigma_b^2 \tag{42}$$

$$\approx \frac{\sigma_w^2}{q^*}\int \mathcal{D}z_1\mathcal{D}z_2\phi(u_1^*)\left[\phi(u_2^*) + \left(\sqrt{2q^*\epsilon^l}z_2 - \sqrt{q^*}\epsilon^l z_1\right)\phi'(u_2^*) + q^*\epsilon^l z_2^2\phi''(u_2^*)\right] + \sigma_b^2 \tag{43}$$

$$= c^* + \sigma_w^2\epsilon^l\int \mathcal{D}z\phi(\sqrt{q^*}z)\left[\phi''(\sqrt{q^*}z) - \frac{1}{\sqrt{q^*}}z\phi'(\sqrt{q^*}z)\right] \tag{44}$$

$$= c^* + \sigma_w^2\epsilon^l\left[\int \mathcal{D}z\phi(\sqrt{q^*}z)\phi''(\sqrt{q^*}z) - \frac{1}{\sqrt{q^*}}\int \mathcal{D}z z\phi(\sqrt{q^*}z)\phi'(\sqrt{q^*}z)\right] \tag{45}$$

$$= c^* - \sigma_w^2\epsilon^l\int \mathcal{D}z\left[\phi'(\sqrt{q^*}z)\right]^2 \tag{46}$$

It follows that the asymptotic recurrence relation in this case will be,

$$\epsilon^{l+1} = -\epsilon^l\sigma_w^2\int \mathcal{D}z\left[\phi'(\sqrt{q^*}z)\right]^2 = -\epsilon^l\chi_1. \tag{47}$$

where χ_1 is the stability condition for the ordered phase. We note that although the approximations were somewhat different the asymptotic recurrence relation for $c^* < 1$ reduces eq. 47 result for $c^* = 1$. We may therefore use 4 for all c^*.

7.3 VARIANCE OF AN INPUT WITH DROPOUT

Result:

In the presence of dropout with rate ρ, the variance of a single input as it is passed through the network is described by the recurrence relation,

$$\bar{q}^l_{aa} = \frac{\sigma^2_w}{\rho} \int \mathcal{D}z \phi^2 \left(\sqrt{\bar{q}^{l-1}_{aa}} z \right) + \sigma^2_b. \tag{48}$$

Derivation:

Recall that the recurrence relation for the pre-activations is given by,

$$z^l_i = \frac{1}{\rho} \sum_j W^l_{ij} p^l_j y^l_j + b^l_i \tag{49}$$

where $p^l_j \sim \text{Bernoulli}(\rho)$. It follows that the variance will be given by,

$$\begin{aligned} \bar{q}^l_{aa} &= \mathbb{E}[(z^l_i)^2] && (50) \\ &= \frac{1}{\rho^2} \sum_j \mathbb{E}[(W^l_{ij})^2] \mathbb{E}[(\rho^l_j)^2] \mathbb{E}[(y^l_j)^2] + \mathbb{E}[(b^l_i)^2] && (51) \\ &= \frac{\sigma^2_w}{\rho} \int \mathcal{D}z \phi^2 \left(\sqrt{\bar{q}^{l-1}_{aa}} z \right) + \sigma^2_b. && (52) \end{aligned}$$

where we have used the fact that $\mathbb{E}[(p^l_j)^2] = \rho$.

7.4 Covariance of two inputs with dropout

Result:

The co-variance between two signals, $z^l_{i;a}$ and $z^l_{i;b}$, with separate i.i.d. dropout masks $p^l_{i;a}$ and $p^l_{i;b}$ is given by,

$$\bar{q}^l_{ab} = \sigma^2_w \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(\bar{u}_1) \phi(\bar{u}_2) + \sigma^2_b. \tag{53}$$

where, in analogy to eq. 4, $\bar{u}_1 = \sqrt{\bar{q}^l_{aa}} z_1$ and $\bar{u}_2 = \sqrt{\bar{q}^l_{bb}} \left(\bar{c}^l_{ab} z_1 + \sqrt{1 - (\bar{c}^l_{ab})^2} z_2 \right)$.

Derivation:

Proceeding directly we find that,

$$\begin{aligned} \mathbb{E}[z^l_{i;a} z^l_{i;b}] &= \frac{1}{\rho^2} \sum_j \mathbb{E}[(W^l_{ij})^2] \mathbb{E}[p^l_{j;a}] \mathbb{E}[p^l_{j;b}] \mathbb{E}[y^l_{j;a} y^l_{j;b}] + \mathbb{E}[b^l_i] && (54) \\ &= \sigma^2_w \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(\bar{u}_1) \phi(\bar{u}_2) + \sigma^2_b && (55) \end{aligned}$$

where we have used the fact that $\mathbb{E}[p^l_{i;a}] = \mathbb{E}[p^l_{i;b}] = \rho$. We have also used the same substitution for $\mathbb{E}[y^l_{j;a} y^l_{j;b}]$ used in the original mean field calculation with the appropriate substitution.

7.5 The lack of a $c^* = 1$ fixed point with dropout

Result:

If $c^l_{ab} = 1$ then it follows that,

$$\bar{c}^{l+1}_{ab} = 1 - \frac{1 - \rho}{\rho \bar{q}^*} \sigma^2_w \int \mathcal{D}z \phi^2 \left(\sqrt{\bar{q}^*} z \right) \tag{56}$$

subject to the approximation, $q^l_{aa} \approx q^l_{bb} \approx q^*$. This implies that $c^{l+1}_{ab} < 1$.

Derivation:

Plugging in $c^l_{ab} = 1$ with $q^l_{aa} \approx q^l_{bb} \approx q^*$ we find that $\bar{u}_1 = \bar{u}_2 = \sqrt{q^*}z_1$. It follows that,

$$c^{l+1}_{ab} = \frac{q^{l+1}_{ab}}{q^*} \tag{57}$$

$$= \frac{1}{q^*}\left[\sigma_w^2 \int \mathcal{D}z\phi^2\left(\sqrt{q^*}z\right) + \sigma_b^2\right] \tag{58}$$

$$= \frac{1}{q^*}\left[\sigma_w^2(1-\rho^{-1}+\rho^{-1}) \int \mathcal{D}z\phi^2\left(\sqrt{q^*}z\right) + \sigma_b^2\right] \tag{59}$$

$$= \frac{1}{q^*}\left[\frac{\sigma_w^2}{\rho} \int \mathcal{D}z\phi^2\left(\sqrt{q^*}z\right) + \sigma_b^2\right] + \frac{\sigma_w^2}{q^*}(1-\rho^{-1}) \int \mathcal{D}z\phi^2\left(\sqrt{q^*}z\right) \tag{60}$$

$$= 1 - \frac{1-\rho}{\rho\bar{q}^*}\sigma_w^2 \int \mathcal{D}z\phi^2\left(\sqrt{\bar{q}^*}z\right) \tag{61}$$

as required. Here we have integrated out z_2 since nether $\bar{u}_1$ nor $\bar{u}_2$ depend on it.

7.6 Mean field gradient scaling

Result:

In mean field theory the expected magnitude of the gradient $||\nabla_{W^l_{ij}} E||^2$ will be proportional to $\mathbb{E}[(\delta^l_i)^2]$.

Derivation:

We first note that since the W^l_{ij} are i.i.d. it follows that,

$$||\nabla_{W^l_{ij}} E||^2 = \sum_{ij}\left(\frac{\partial E}{\partial W^l_{ij}}\right)^2 \tag{62}$$

$$\approx N_l N_{l+1}\mathbb{E}\left[\left(\frac{\partial E}{\partial W^l_{ij}}\right)^2\right] \tag{63}$$

where we have used the fact that the first line is related to the sample expectation over the different realizations of the W^l_{ij} to approximate it by the analytic expectation in the second line. In mean field theory since the pre-activations in each layer are assumed to be i.i.d. Gaussian it follows that,

$$\mathbb{E}\left[\left(\frac{\partial E}{\partial W^l_{ij}}\right)^2\right] = \mathbb{E}[(\delta^l_i)^2]\mathbb{E}[\phi^2(z^{l-1}_j)] \tag{64}$$

and the result follows.

7.7 Mean field backpropagation

Result:

In mean field theory the recursion relation for the variance of the errors, $\tilde{q}^{\,l} = \mathbb{E}[(\delta^l_i)^2]$ is given by,

$$\tilde{q}^{\,l}_{aa} = \tilde{q}^{\,l+1}_{aa}\frac{N_{l+1}}{N_{l+2}}\chi_1(q^l_{aa}). \tag{65}$$

Derivation:

Computing the variance directly and using mean field approximation,

$$\tilde{q}_{aa}^{\ l} = \mathbb{E}[(\delta_{i;a}^{l})^2] = \mathbb{E}[(\phi'(z_{i;a}^{l}))^2] \sum_j \mathbb{E}[(\delta_{j;a}^{l+1})^2]\mathbb{E}[(W_{ji}^{l+1})^2] \tag{66}$$

$$= \mathbb{E}[(\phi'(z_{i;a}^{l}))^2] \frac{\sigma_w^2}{N_{l+1}} \sum_j \mathbb{E}[(\delta_{j;a}^{l+1})^2] \tag{67}$$

$$= \mathbb{E}[(\phi'(z_{i;a}^{l}))^2] \frac{N_{l+1}}{N_{l+2}} \sigma_w^2 \tilde{q}_{aa}^{\ l+1} \tag{68}$$

$$= \sigma_w^2 \tilde{q}_{aa}^{\ l+1} \frac{N_{l+1}}{N_{l+2}} \int \mathcal{D}z \left[\phi'\left(\sqrt{q_{aa}^{l}} z\right)\right]^2 \tag{69}$$

$$\approx \tilde{q}_{aa}^{\ l+1} \frac{N_{l+1}}{N_{l+2}} \chi_1 \tag{70}$$

as required. In the last step we have made the approximation that $q_{aa}^{l} \approx q^*$ since the depth scale for the variance is short ranged.

7.8 Mean field gradient covariance scaling

Result:

In mean field theory we expect the covariance between the gradients of two different inputs to scale as,

$$(\nabla_{W_{ij}^{l}} E_a) \cdot (\nabla_{W_{ij}^{l}} E_b) \sim \mathbb{E}[\delta_{i;a}\delta_{i;b}]. \tag{71}$$

Derivation:

We proceed in a manner analogous to Appendix 7.6. Note that in mean field theory since the weights are i.i.d. it follows that

$$(\nabla_{W_{ij}^{l}} E_a) \cdot (\nabla_{W_{ij}^{l}} E_b) = \sum_{ij} \frac{\partial E_a}{\partial W_{ij}^{l}} \frac{\partial E_b}{\partial W_{ij}^{l}} \tag{72}$$

$$\approx N_l N_{l+1} \mathbb{E}\left[\frac{\partial E_a}{\partial W_{ij}^{l}} \frac{\partial E_b}{\partial W_{ij}^{l}}\right] \tag{73}$$

where, as before, the final term is approximating the sample expectation. Since the weights in the forward and backwards passes are chosen independently it follows that we can factor the expectation as,

$$\mathbb{E}\left[\frac{\partial E_a}{\partial W_{ij}^{l}} \frac{\partial E_b}{\partial W_{ij}^{l}}\right] = \mathbb{E}[\delta_{i;a}^{l}\delta_{i;b}^{l}]\mathbb{E}[\phi(z_{i;a}^{l})\phi(z_{i;b}^{l})] \tag{74}$$

and the result follows.

7.9 Mean field backpropagation of covariance

Result:

The covariance between the gradients due to two inputs scales as,

$$\tilde{q}_{ab}^{\ l} = \tilde{q}_{ab}^{\ l+1} \frac{N_{l+1}}{N_{l+2}} \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1)\phi'(u_2) \tag{75}$$

under backpropagation.

Derivation

As in the analogous derivation for the variance, we compute directly,

$$\tilde{q}_{ab}^{\ l} = \mathbb{E}[\delta_{i;a}^{l}\delta_{i;b}^{l}] = \mathbb{E}\left[\phi'(z_{i;a})\phi'(z_{i;b})\right] \sum_j \mathbb{E}[\delta_{j;a}^{l+1}\delta_{j;b}^{l+1}]\mathbb{E}[(W_{ji}^{l+1})^2] \tag{76}$$

$$= \tilde{q}_{ab}^{\ l+1} \frac{N_{l+1}}{N_{l+2}} \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1)\phi'(u_2) \tag{77}$$

as required.

7.10 FURTHER EXPERIMENTAL RESULTS

Here we include some more experimental figures that investigate the effects of training time, minimizer, and dataset more closely.

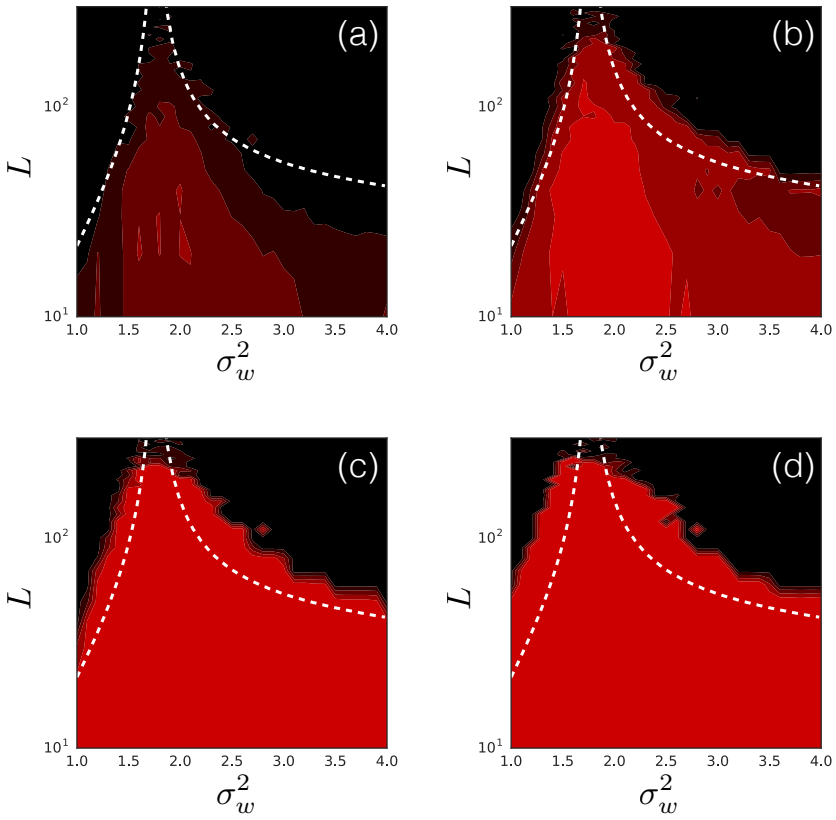

Figure 7: Training accuracy on MNIST after (a) 45 (b) 304 (c) 2048 and (d) 13780 steps of SGD with learning rate 10^{-3}.

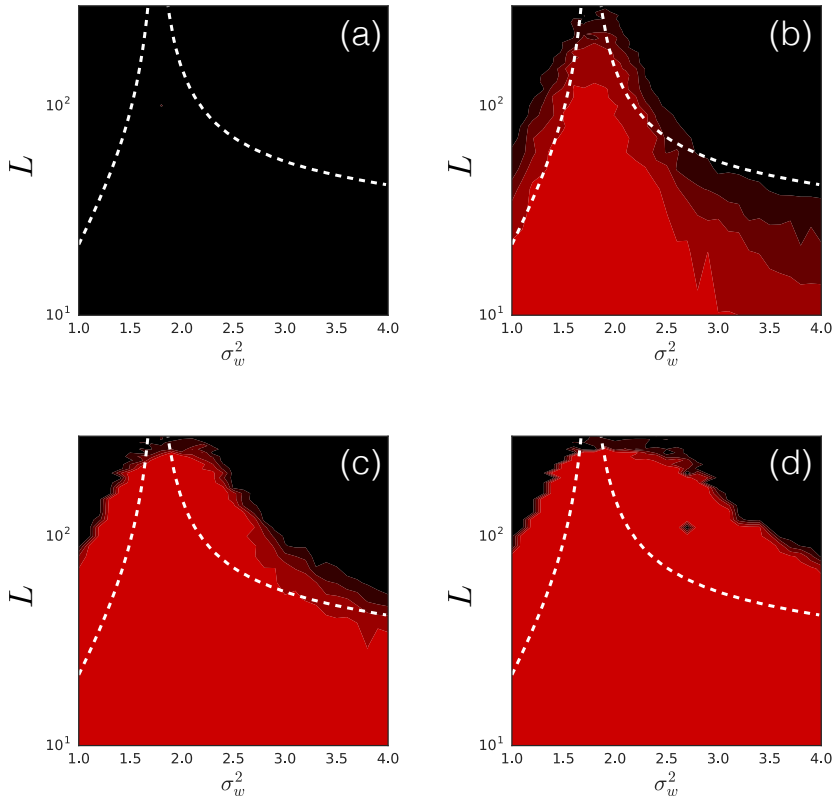

Figure 8: Training accuracy on MNIST after (a) 45 (b) 304 (c) 2048 and (d) 13780 steps of RMSProp with learning rate 10^{-5}.