around a fixed value of L regardless of ρ, which shows that indeed gradient dynamics determines the initialization quality in tanh resnets. There is also a minor increase in performance with increasing ρ regardless of L; this is counterintuitive as increasing ρ means "decreasing expressivity." It is currently not clear what accounts for this effect.

ReLU, vary σ_w We train a grid of ReLU FRN on MNIST, varying $\sigma_w^2 \in [0, 1.5]$ while fixing $\sigma_v^2 = 1, \sigma_a^2 = \sigma_b^2 = \frac{1}{2}$. The resulting test set accuracies are shown in Fig. 3(d). The dark upper region signifies failure of training caused by numerical issues with exploding activation and gradient norms: This corresponds to the region where $\mathbf{p}^{(L)}$, which is a measure of the mean magnitude of an neuronal activation in layer L, becomes too big. We see that the best test accuracies are given by depths just below where these numerical issues occur. However, if we were to predict that the optimal init is the one minimizing $\boldsymbol{\chi}^{(0)}/\boldsymbol{\chi}^{(L)} \geq 1$, then we would be wrong — in fact it is exactly the opposite. In this case, the dynamics of $\mathbf{s}^{(l)}$, $\mathbf{p}^{(l)}$, and $\boldsymbol{\chi}^{(0)}/\boldsymbol{\chi}^{(l)}$ are approximately the same (all $\exp(\Theta(l))$ with the same hidden constants), and optimal performance corresponds to the highest $\mathbf{s}^{(L)}$, $\mathbf{p}^{(L)}$, and $\boldsymbol{\chi}^{(0)}/\boldsymbol{\chi}^{(L)}$ without running into `infs`.

α-ReLU, vary α. We similarly trained a grid of α-ReLU FRN on MNIST, varying only α and the depth, fixing all $\sigma_\bullet$. Fig. 3(e) shows their test accuracies. We see similar behavior to ReLU, where when the net is too deep, numerical issues doom the training (black upper right corner), but the best performance is given by L just below where this problem occurs. In this case, if we were to predict optimality based on minimizing gradient explosion, we would be again wrong, and furthermore, the contour plot of $\boldsymbol{\chi}^{(0)}/\boldsymbol{\chi}^{(L)}$ (white dashed line) now gives no information at all on the test set accuracy. In contrast, the contours for $\mathbf{s}^{(l)}$ succeeds remarkably well at this prediction (yellow/green lines).[6] By interpolation, this suggests that indeed in the ReLU case, it is expressivity, not trainability, which determines performance at test time.

In all of our experiments, we did not find $\mathbf{e}$ dynamics to be predictive of neural network performance.

7 Conclusion

In this paper, we have extended the mean field formalism developed by [9, 10, 11] to residual networks, a class of models closer to practice than classical feedforward neural networks as were investigated earlier. We proved and verified that in both the forward and backward passes, most of the residual networks discussed here do not collapse their input space geometry or the gradient information exponentially. We found our theory incredibly predictive of test time performance despite saying nothing about the dynamics of training. In addition, we overwhelmingly find, through theory and experiments, that an optimal initialization scheme must take into account the depth of the residual network. The reason that Xavier [4] or He [5] scheme are not the best for residual networks is in fact not that their statistical assumptions are fragile — theirs are similar to our mean field theoretic assumptions, and they hold up in experiments for large width — but rather that their structural assumptions on the network break very badly on residual nets.

Open Problems. Our work thus have shown that optimality of initialization schemes can be very unstable with respect to architecture. We hope this work will form a foundation toward a mathematically grounded initialization scheme for state-of-the-art architectures like the original He et al. residual network. To do so, there are still two major components left to study out of the following three: 1. Residual/skip connection 2. Batchnorm 3. Convolutional layers. Recurrent architectures and attention mechanisms are also still mostly unexplored in terms of mean field theory. Furthermore, many theoretical questions still yet to be resolved; the most important with regard to mean field theory is: why can we make Axioms 3.1 and 3.2 and still be able to make accurate predictions? We hope to make progress on these problems in the future and encourage readers to take part in this effort.