Appendices

A Comment on batch normalization

A crucial component of practically used ResNets is batch normalization (Ioffe and Szegedy, 2015). When it is used on pre-activations, between each layer, the propagation of the information in the network is described by:

$$\boldsymbol{x}^l = \phi(\boldsymbol{y}^l) + a\boldsymbol{x}^{l-1}, \quad \boldsymbol{h}^l = \boldsymbol{W}^l\boldsymbol{x}^{l-1} + \boldsymbol{b}^l, \tag{28}$$

with

$$y_i^l = \frac{h_i^l - \mu_i^l}{\sigma_i^l}\gamma_i^l + \beta_i^l, \tag{29}$$

where μ_i^l is the mean and σ_i^l (regularized with some small ϵ) is the standard deviation of the k'th mini batch inputs i'th coefficient in layer l. γ_i^l and β_i^l are parameters optimized during the learning process. In this case, the formula for the Jacobian reads:

$$\boldsymbol{J} = \prod_{l=1}^{L}\left(\boldsymbol{D}^l\boldsymbol{H}^l\boldsymbol{W}^l + \mathbf{1}a\right), \tag{30}$$

where $\boldsymbol{H}^l$ is a diagonal matrix such that $H_{ij}^l = \delta_{ij}\gamma_i^l/\sigma_i^l$. Therefore, the only difference in the spectral statistics derivation from the previous section, is that (21) becomes

$$c_{2\mathrm{BN}}^l = \sigma_W^2\left\langle\left(\frac{\gamma^l}{\sigma^l}\phi'(y^l)\right)^2\right\rangle_l . \tag{31}$$

Thus, the universal, large L limit equation for the Green's function of the Jacobian (25) holds also when batch normalization is included. Again, σ_i^l and y_i^l can be treated as random variables. Unfortunately, the evolution of their probability density functions across the layers is more complicated and beyond the scope of this paper.

B Spectrum of the Jacobian

To make the characteristics of the spectrum more explicit, we shall calculate the positions of the spectral edges of the probability density of squared singular values. Their locations, $z*$, can be determined from the condition $\frac{1}{G'(z*)} = 0$. In this case we assume $a = 1$ for simplicity and take a derivative of (25), obtaining

$$G' = (zG' + G)e^{c(1-2zG)} - (zG-1)e^{c(1-2zG)}2c(G + zG'). \tag{32}$$

The exponent can be eliminated with the help of (25), leading to

$$1 = \left(z + \frac{G}{G'}\right)\frac{G}{zG-1} - 2cG\left(\frac{G}{G'} + z\right). \tag{33}$$

Taking $\frac{1}{G'} = 0$, we arrive to the quadratic equation

$$2cz^2G^2 - 2czG - 1 = 0, \tag{34}$$

which, together with (25), determine the location of the edges and the value of the Green's function at these points. Solving, we obtain

$$z_\pm = \left(1 + c \pm \sqrt{c(2+c)}\right)e^{\pm\sqrt{c(2+c)}}. \tag{35}$$

Note that the perfect isometry (i.e. all eigenvalues are 1) is achieved for $c = 0$, as independently proposed by (Zhang et al., 2019), while for small c the size of the support grows sublinearly $z_\pm \approx 1 \pm 2\sqrt{2c}$. Moreover, for large c, that is far from dynamical isometry, the largest eigenvalue is exponentially large, while the smallest is exponentially small. This fact underscores importance of a proper initialization.

C Detailed derivation of the signal propagation

With the scalings of from section 3.1 made explicit, we have $\left\langle W^l_{ij} W^l_{km} \right\rangle_{wbl} = \left\langle W^l_{ij} W^l_{im} \right\rangle_l = \delta_{ik}\delta_{jm}\frac{(\sigma_W)^2}{LN}$. Now, based on (1) and the above considerations, we have

$$q^l = \frac{(\sigma_W)^2}{L}\left\langle x^2 \right\rangle_{l-1} + (\sigma_b)^2 \tag{36}$$

and

$$\left\langle x^2 \right\rangle_{l-1} = \left\langle \phi\left(h^{l-1}\right)^2 \right\rangle_{l-1} + \frac{2a}{N}\sum_{i=1}^{N} \phi\left(h_i^{l-1}\right) x_i^{l-2} + a^2 \left\langle x^2 \right\rangle_{l-2} \tag{37}$$

We assume that the factorization $\frac{1}{N}\sum_k x_k^{l-1}\phi(h_k^l) = \langle x\rangle_{l-1}\int \mathcal{D}u\phi(u\sqrt{q^l})$ holds in the large N limit. This is justified, as the input to h_k^l comes from all the many elements of x^{l-1}. We can rewrite (37) as

$$\left\langle x^2 \right\rangle_{l-1} = \int \mathcal{D}z\phi^2\left(\sqrt{q^{l-1}}z\right) + 2a\langle x\rangle_{l-2}\int \mathcal{D}z\phi\left(\sqrt{q^{l-1}}z\right) + a^2\left\langle x^2 \right\rangle_{l-2} \tag{38}$$

Turning to $\langle x\rangle_{l-2}$, based on (1), we have:

$$\langle x\rangle_l = a\langle x\rangle_{l-1} + \int \mathcal{D}z\phi\left(\sqrt{q^l}z\right). \tag{39}$$

For $\langle x\rangle_0 = 0$ the recurrence yields

$$\langle x\rangle_l = \sum_{k=0}^{l-1} a^k \int \mathcal{D}z\phi\left(\sqrt{q^{l-k}}z\right). \tag{40}$$

Thus, (38), with a shift in l, turns into

$$\left\langle x^2 \right\rangle_l = \int \mathcal{D}z\phi^2\left(\sqrt{q^l}z\right) + 2\left[\sum_{k=1}^{l-1} a^k \int \mathcal{D}z\phi\left(\sqrt{q^{l-k}}z\right)\right]\int \mathcal{D}z\phi\left(\sqrt{q^l}z\right) + a^2\left\langle x^2 \right\rangle_{l-1}. \tag{41}$$

Finally, we use (36) to obtain

$$q^{l+1} = a^2 q^l + \left(1-a^2\right)\sigma_b^2 + \frac{(\sigma_W)^2}{L}\int \mathcal{D}z\phi^2\left(\sqrt{q^l}z\right) + 2\frac{(\sigma_W)^2}{L}\left[\sum_{k=1}^{l-1} a^k \int \mathcal{D}z\phi\left(\sqrt{q^{l-k}}z\right)\right]\int \mathcal{D}z\phi\left(\sqrt{q^l}z\right), \tag{42}$$

which is a closed recursive equation for q^l. We note that for $a = 0$, the known, feed-forward network recursion relation is recovered. Furthermore, for the case of $a = 1$, in contrast to the feed-forward architecture, the biases do not influence the statistical properties of the pre-activations. Moreover, for ResNets, this recursive relation is iteratively additive, namely each q^{l+1} is a result of adding some terms to the previous q^l. In all the examples studied below, the first term is positive and the second term is non-negative. This in turn means that the variance of pre-activations grows with the networks depth and there are no non-trivial fixed points of this recursion equation. Finally, here we can see the importance of the $\frac{1}{N}$ scaling of $(\sigma_W)^2$, without which, q^l would grow uncontrollably with l.

D Results for various activation functions

We now investigate particular examples of activation functions. For simplicity, we consider purely residual networks (we set $a = 1$). The numerical verifications of the results presented here will follow in the next subsection.

1. Linear

   In the case of the linear activation function $\phi'(x) = 1$ and there is no need to consider the way the pre-activations change across the network. Thus we can proceed to calculating the cumulant which yields $c = c_2 = \sigma_W^2$.

2. Rectified Linear Unit

   The example of ReLU is only slightly more involved. Now we have $\phi'(x) = \theta(x)$, where $\theta(x)$ is the Heaviside theta function, and thus

$$c_2^l = \sigma_W^2 \int \mathcal{D}u \phi'^2 \left(u\sqrt{q^l}\right) = \int_0^\infty \mathcal{D}u = \frac{1}{2}\sigma_W^2. \tag{43}$$

3. Leaky ReLU

   The activation function interpolating between the first two examples is $\phi(x) = \max(\alpha x, x)$ with $0 < \alpha < 1$. In this case

$$c_2^l = \sigma_W^2 \left(\int_{-\infty}^0 \alpha^2 \mathcal{D}u + \int_0^\infty \mathcal{D}u\right) = \frac{\sigma_W^2}{2}(1+\alpha^2). \tag{44}$$

All together, this leads to the following equation for the Green's function

$$G(z) = (zG(z) - 1)\, e^{\frac{1}{2}\sigma_W^2 (1+\alpha^2)(1-2zG(z))}, \tag{45}$$

where $\alpha = 1$ corresponds to the linear activation function and $\alpha = 0$ to ReLU.

Equation (45) can be easily solved numerically for the spectral probability density of the Jacobian. For completeness, we write down the recursion relation (42) in these three cases:

$$q^l = q^{l-1} + \frac{\sigma_W^2}{2L}\left(\alpha^2 + 1\right) q^{l-1} + \frac{\sigma_W^2}{\pi L}(1-\alpha)^2 \sqrt{q^{l-1}} \left(\sum_{k=1}^{l-2} \sqrt{q^k}\right). \tag{46}$$

For the linear activation function ($\alpha = 1$), its solution is readily available and reads

$$q^l = q^1 \left(1 + \frac{\sigma_W^2}{L}\right)^{l-1} \simeq q^1 e^{(l-1)\frac{\sigma_W^2}{L}}, \tag{47}$$

which explicitly shows the importance of the $1/L$ rescaling introduced earlier.

4. Hard hyperbolic tangent

   The hard tanh activation function is defined by $\phi(x) = x$ for $|x| \leq 1$ and $\phi(x) = \text{sgn(x)}$ elsewhere. Thus:

$$c_2^l = \sigma_W^2 \int_{-1}^1 \mathcal{D}u = \sigma_W^2 \operatorname{erf}\left(\frac{1}{\sqrt{2}}\right) = c. \tag{48}$$

   The resulting recurrence equation for the variance of the preactivations reads:

$$q^{l+1} = q^l \left[1 + \frac{\sigma_W^2}{L}\left(\operatorname{erf}\left(\frac{1}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi e}}\right)\right] + \frac{\sigma_W^2}{L}\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}\right)\right) \tag{49}$$

   and can be easily solved.

In the preceding examples we dealt with piecewise linear activation functions. Note that in these cases the parameter c does not depend on the variance of biases and linearly increases with the variance of weights. For other nonlinear activation functions to obtain the cumulants, we need to use the recurrence relation describing the signal propagation in the network.

5. Hyperbolic tangent

   When $\phi(x) = \tanh(x)$, the activation function is antisymmetric and the last term of (42) vanishes. Thus the recurrence takes the form:

$$q^{l+1} = q^l + \frac{(\sigma_W)^2}{L} \int \mathcal{D}z \phi^2 \left(\sqrt{q^l} z\right) \tag{50}$$

   In the large L limit, we can write

$$q^{l+1} = q^l + \frac{(\sigma_W)^2}{L} \int \mathcal{D}z \phi^2 \left(\sqrt{q^{l-1} + \Delta^l} z\right), \tag{51}$$

where we assume $\Delta^l \sim 1/L$. Expanding this recursively around q^{l-1} for decreasing l and keeping only the leading term, as long as for any k, $q^k \gg 1/L^2$, we obtain :

$$q^{l+1} \approx q^l + \frac{(\sigma_W)^2}{L} \int \mathcal{D}z\phi^2\left(\sqrt{q^1}z\right). \tag{52}$$

Therefore, the solution of the recursion is:

$$q^l \approx q^1 + (l-1)\frac{(\sigma_W)^2}{L} \int \mathcal{D}z\phi^2\left(\sqrt{q^1}z\right). \tag{53}$$

The variance grows linearly with l (this is verified in appendix E, see Fig. 5). Cumulants c_2^l and thus c are obtained with numerical integration.

In fact, in the above calculations we have only used the antisymmetry property of the hyperbolic tangent activation function and the properties of its behavior near q^1. Therefore, these results are valid for other antisymmetric activation functions like $\phi(\mathrm{x}) = \arctan(\mathrm{x})$.

6. Sigmoid

The sigmoid activation function, $\phi(x) = \frac{1}{1+e^{-x}}$ is the first example we encounter, for which $\langle\phi(h)\rangle \neq 0$, thus it deserves special attention. In particular, in this case one needs to additionally address the last term in (42). It turns out, that

$$\int \mathcal{D}z\phi\left(\sqrt{q^l}z\right) = \frac{1}{2} \tag{54}$$

irrespective of l. Therefore, the recurrence relation becomes:

$$q^{l+1} = q^l + \frac{(\sigma_W)^2}{L} \int \mathcal{D}z\phi^2\left(\sqrt{q^l}z\right) + \frac{(\sigma_W)^2}{2L}(l-1). \tag{55}$$

Thus, we can see, that due to the non-zero first moment of the activation function (54) the mean and the second moment of post-activations grow with depth. Similarly, the variance of pre-activations increases as the signal propagates, which causes quick saturation of the sigmoid nonlinearity. This in turn precludes training of deep networks (Glorot and Bengio, 2010).

Analogically to the previous case, one can derive an approximation to the solution of the recursion relation. In this case it becomes:

$$q^{l+1} \approx q^1 + \frac{(\sigma_W)^2 l}{L} \int \mathcal{D}z\phi^2\left(\sqrt{q^1}z\right) + \frac{(\sigma_W)^2}{4L}l(l-1). \tag{56}$$

We verify this result in Fig. 5 in Appendix E

7. Scaled Exponential Linear Units

Our final example is the SELU activation function, one introduced recently in (Klambauer et al., 2017) with the intent to bypass the batch normalization procedure. In this case, we have $\phi(x) = \lambda x$ for $x > 0$ and $\phi(x) = \lambda\beta\left(e^x - 1\right)$ for $x \leq 0$. Thus, it is not antisymmetric and is nonlinear for negative arguments. It turns out, that

$$c_2^l = \frac{(\sigma_W)^2\lambda^2}{2}\left[1 + \beta^2 e^{2q^l} \operatorname{erfc}\left(\sqrt{2q^l}\right)\right]. \tag{57}$$

Moreover:

$$\int \mathcal{D}z\phi^2\left(\sqrt{q^l}z\right) = \frac{\lambda^2 q^l}{2} + \frac{\beta^2\lambda^2}{2}\left[1 + e^{2q^l} \operatorname{erfc}\left(\sqrt{2q^l}\right) - 2e^{q^l/2} \operatorname{erfc}\left(\sqrt{\frac{q^l}{2}}\right)\right] \tag{58}$$

and

$$\int \mathcal{D}z\phi\left(\sqrt{q^l}z\right) = \lambda\sqrt{\frac{q^l}{2\pi}} + \frac{\lambda\beta}{2}\left(e^{q^l/2}\operatorname{erfc}\left(\sqrt{\frac{q^l}{2}}\right) - 1\right). \tag{59}$$

These yield the recursion relation for q^l via (42). One can check that for $\beta = 0$ and $\lambda = 1$, the results for ReLU are recovered.

These theoretical predictions for the recursion relations are tested with numerical simulations using Mathematica. The results are relegated to Appendix E.

E Numerical verification of the recurrence relations

To validate the assumptions made and corroborate the theoretical results obtained in subsections 3.2 and D, we simulate signal propagation in the studied residual neural networks with different activation functions. The outcomes of these experiments are depicted in Figure 5. Numerical solution to the recurrence relations allows us to numerically calculate the parameter c, as a function of variances of weights and biases, which is presented in Figure 6.

F Baseline

We advocate for setting the same value of the effective cumulant (and hence keeping the same spectrum of the input-output Jacobian) when comparing the effects of using different activation function on the learning process. Thus, here, in Fig. 7, for comparison, we showcase the learning accuracy when instead of the effective cumulant, the weight matrices entries' variances (equal to $1/NL$) are kept the same across the networks.

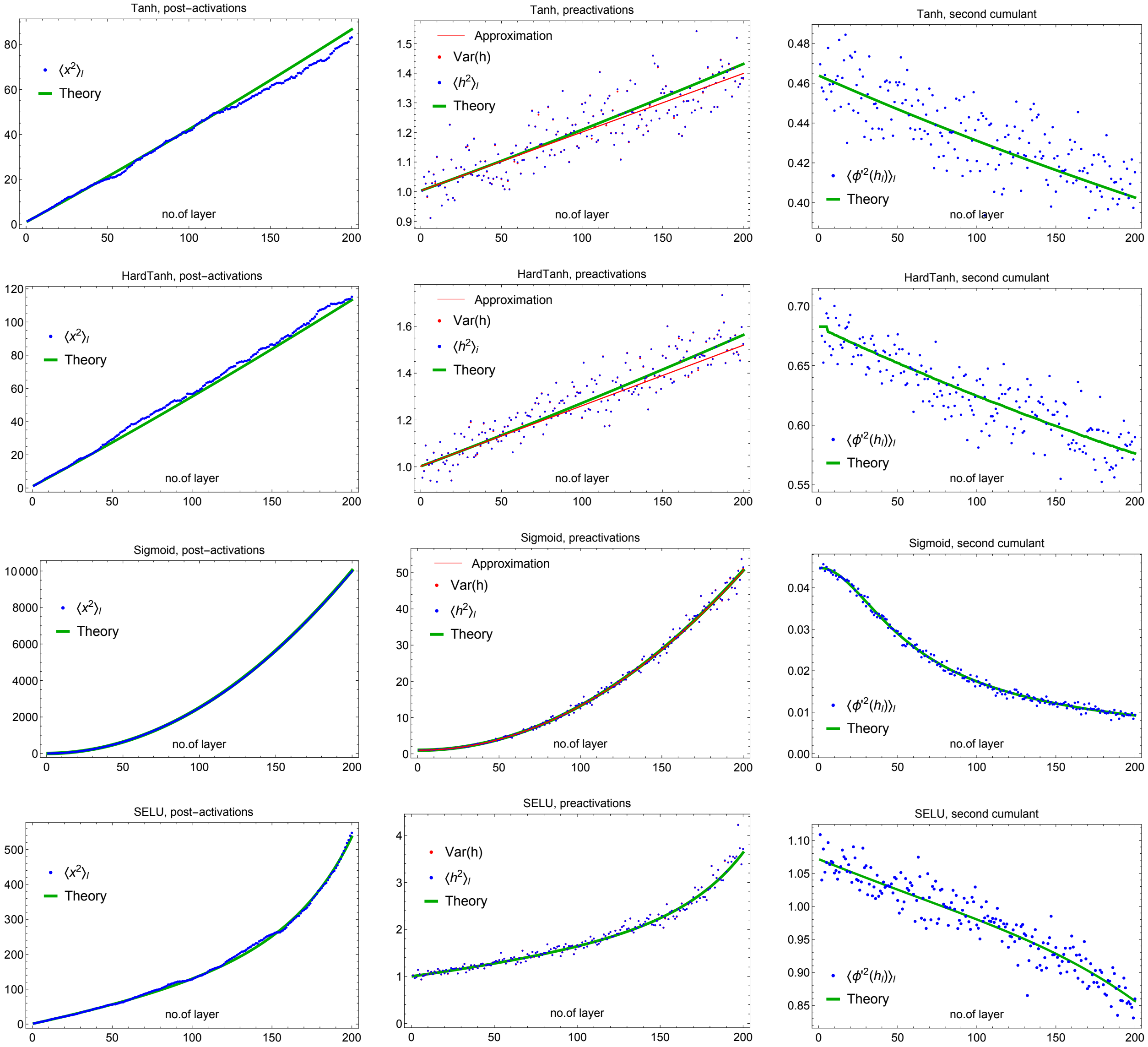

Figure 5: Verification of the numerical solution to the recurrence equations for post-activations (38) (left column) and preactivations (42) (middle). Based on this signal propagation the effective cumulant c_l for each layer was calculated (right column). The solid red lines represent the approximation (53) for tanh and hard tanh nonlinearity and (56) for sigmoid. Solutions of recurrences (solid) are confronted with the numerical simulation (dots) of residual fully connected networks with $L = 200$ layers of width $N = 800$. Data points represent a single run of simulations. Weights are independently sampled from Gaussian distribution of zero mean and variance equal to $\frac{1}{NL}$. Biases and network input are sampled from standard normal distribution. A small variability of c^l across the network justifies the assumption made for derivation of (22).

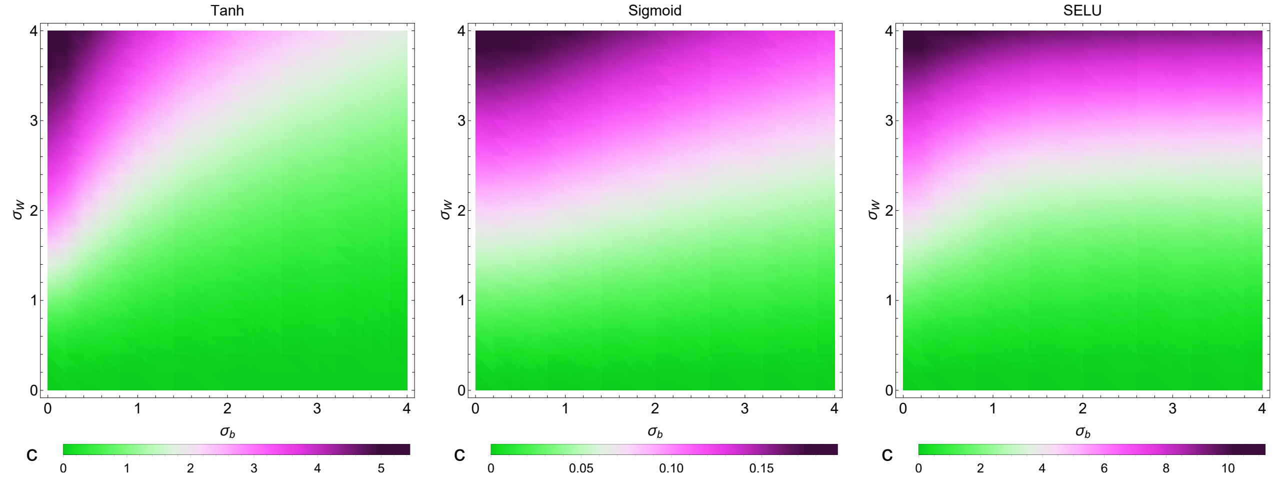

Figure 6: Dependence of the parameter c (which determines the shape of the spectrum) on the variances of biases and weights. The smaller c, the closer to the perfect dynamical isometry. Note different scales on each plot. Low value of c for sigmoid is a consequence of saturation of the nonlinearity.

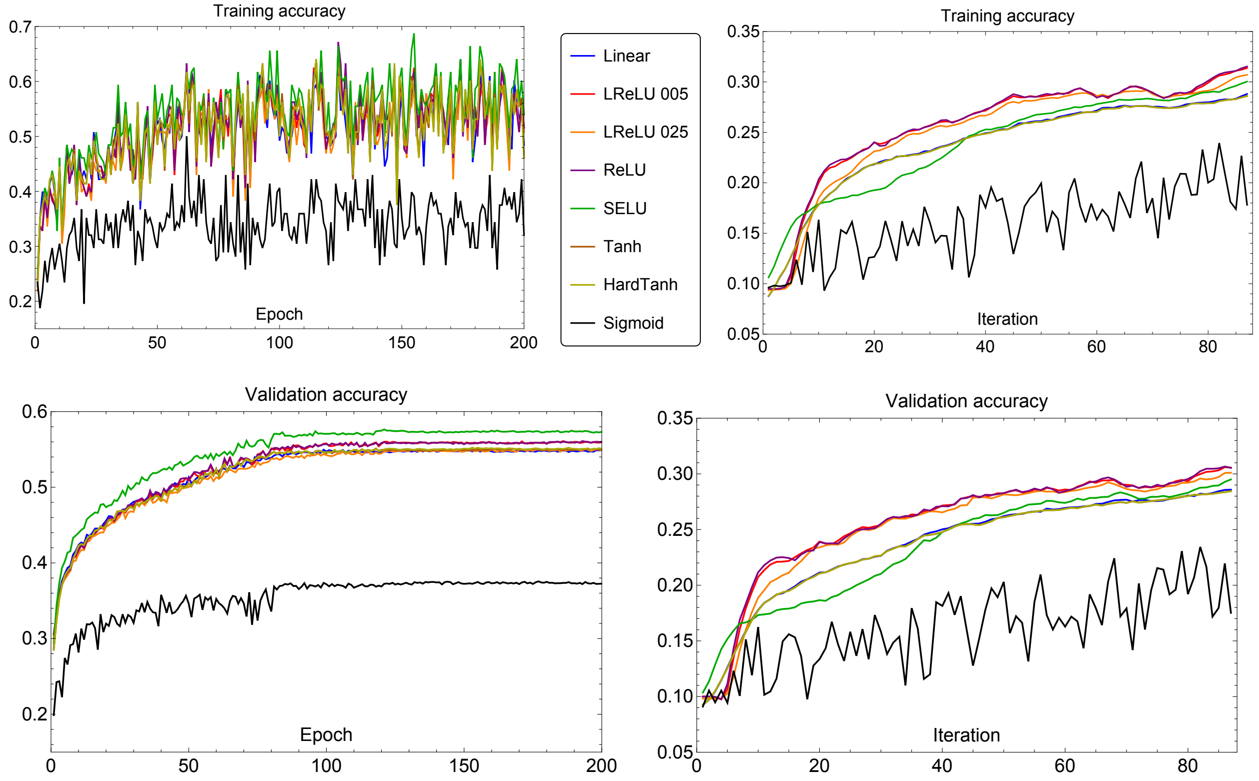

Figure 7: Training (top) and validation (bottom) accuracy during first 200 epochs (left) and first 100 iterations (right) of residual networks with various activation functions. The weight initialization was Gaussian with zero mean and $1/NL$ variance. We set $\alpha = 0.05$ and $\alpha = 0.25$ for leaky ReLU (LReLU).