Appendix

Here we include the full proofs from sections in the paper.

A. Proofs and additional results from Section 2.1

Proof of Theorem 2

Proof. We show inductively that $F_A(x; W)$ partitions the input space into convex polytopes via hyperplanes. Consider the image of the input space under the first hidden layer. Each neuron $v_i^{(1)}$ defines hyperplane(s) on the input space: letting $W_i^{(0)}$ be the ith row of $W^{(0)}$, $b_i^{(0)}$ the bias, we have the hyperplane $W_i^{(0)}x + b_i = 0$ for a ReLU and hyperplanes $W_i^{(0)}x + b_i = \pm 1$ for a hard-tanh. Considering all such hyperplanes over neurons in the first layer, we get a hyperplane arrangement in the input space, each polytope corresponding to a specific activation pattern in the first hidden layer.

Now, assume we have partitioned our input space into convex polytopes with hyperplanes from layers $\leq d-1$. Consider $v_i^{(d)}$ and a specific polytope R_i. Then the activation pattern on layers $\leq d-1$ is constant on R_i, and so the input to $v_i^{(d)}$ on R_i is a linear function of the inputs $\sum_j \lambda_j x_j + b$ and some constant term, comprising of the bias and the output of saturated units. Setting this expression to zero (for ReLUs) or to ± 1 (for hard-tanh) again gives a hyperplane equation, but this time, the equation is only valid in R_i (as we get a different linear function of the inputs in a different region.) So the defined hyperplane(s) either partition R_i (if they intersect R_i) or the output pattern of $v_i^{(d)}$ is also constant on R_i. The theorem then follows. □

This implies that any one dimensional trajectory $x(t)$, that does not 'double back' on itself (i.e. reenter a polytope it has previously passed through), will not repeat activation patterns. In particular, after seeing a transition (crossing a hyperplane to a different region in input space) we will never return to the region we left. A simple example of such a trajectory is a straight line:

Corollary 1. Transitions and Output Patterns in an Affine Trajectory *For any affine one dimensional trajectory $x(t) = x_0 + t(x_1 - x_0)$ input into a neural network F_W, we partition $\mathbb{R} \ni t$ into intervals every time a neuron transitions. Every interval has a unique network activation pattern on F_W.*

Generalizing from a one dimensional trajectory, we can ask how many regions are achieved over the entire input – i.e. how many distinct activation patterns are seen? We first prove a bound on the number of regions formed by k hyperplanes in $\mathbb{R}^m$ (in a purely elementary fashion, unlike the proof presented in (Stanley, 2011))

Theorem 5. Upper Bound on Regions in a Hyperplane Arrangement *Suppose we have k hyperplanes in $\mathbb{R}^m$ - i.e. k equations of form $\alpha_i x = \beta_i$. for $\alpha_i \in \mathbb{R}^m$, $\beta_i \in \mathbb{R}$. Let the number of regions (connected open sets bounded on some sides by the hyperplanes) be $r(k, m)$. Then*

$$r(k, m) \leq \sum_{i=0}^{m} \binom{k}{i}$$

Proof of Theorem 5

Proof. Let the hyperplane arrangement be denoted $\mathcal{H}$, and let $H \in \mathcal{H}$ be one specific hyperplane. Then the number of regions in $\mathcal{H}$ is precisely the number of regions in $\mathcal{H} - H$ plus the number of regions in $\mathcal{H} \cap H$. (This follows from the fact that H subdivides into two regions exactly all of the regions in $\mathcal{H} \cap H$, and does not affect any of the other regions.)

In particular, we have the recursive formula

$$r(k, m) = r(k-1, m) + r(k-1, m-1)$$

We now induct on $k + m$ to assert the claim. The base cases of $r(1, 0) = r(0, 1) = 1$ are trivial, and assuming the claim for $\leq k + m - 1$ as the induction hypothesis, we have

$$\begin{aligned} r(k-1,m) + r(k-1,m-1) &\leq \sum_{i=0}^{m} \binom{k-1}{i} + \sum_{i=0}^{m-1} \binom{k-1}{i} \\ &\leq \binom{k-1}{0} + \sum_{i=0}^{d-1} \binom{k-1}{i} + \binom{k-1}{i+1} \\ &\leq \binom{k}{0} + \sum_{i=0}^{m-1} \binom{k}{i+1} \end{aligned}$$

where the last equality follows by the well known identity

$$\binom{a}{b} + \binom{a}{b+1} = \binom{a+1}{b+1}$$

This concludes the proof. □

With this result, we can easily prove Theorem 1 as follows:

Proof of Theorem 1

Proof. First consider the ReLU case. Each neuron has one hyperplane associated with it, and so by Theorem 5, the first hidden layer divides up the inputs space into $r(k,m)$ regions, with $r(k,m) \leq O(k^m)$.

Now consider the second hidden layer. For every region in the first hidden layer, there is a different activation pattern in the first layer, and so (as described in the proof of Theorem 2) a different hyperplane arrangement of k hyperplanes in an m dimensional space, contributing at most $r(k,m)$ regions.

In particular, the total number of regions in input space as a result of the first and second hidden layers is $\leq r(k,m) * r(k,m) \leq O(k^2 m)$. Continuing in this way for each of the n hidden layers gives the $O(k^m n)$ bound.

A very similar method works for hard tanh, but here each neuron produces two hyperplanes, resulting in a bound of $O((2k)^{mn})$.

□

B. Proofs and additional results from Section 2.2

Proof of Theorem 3

B.1. Notation and Preliminary Results

Difference of points on trajectory Given $x(t) = x, x(t+dt) = x + \delta x$ in the trajectory, let $\delta z^{(d)} = z^{(d)}(x + \delta x) - z^{(d)}(x)$

Parallel and Perpendicular Components: Given vectors x, y, we can write $y = y_\perp + y_\|$ where $y_\perp$ is the component of y perpendicular to x, and $y_\|$ is the component parallel to x. (Strictly speaking, these components should also have a subscript x, but we suppress it as the direction with respect to which parallel and perpendicular components are being taken will be explicitly stated.)

This notation can also be used with a matrix W, see Lemma 1.

Before stating and proving the main theorem, we need a few preliminary results.

Lemma 1. Matrix Decomposition *Let $x, y \in \mathbb{R}^k$ be fixed non-zero vectors, and let W be a (full rank) matrix. Then, we can write*

$$W = {}^{\|}W_{\|} + {}^{\|}W_{\perp} + {}^{\perp}W_{\|} + {}^{\perp}W_{\perp}$$

such that

$$
\begin{aligned}
{}^{||}W_{\perp}x &= 0 & \quad\quad {}^{\perp}W_{\perp}x &= 0 \\
y^T\,{}^{\perp}W_{||} &= 0 & \quad\quad y^T\,{}^{\perp}W_{\perp} &= 0
\end{aligned}
$$

i.e. the row space of W is decomposed to perpendicular and parallel components with respect to x (subscript on right), and the column space is decomposed to perpendicular and parallel components of y (superscript on left).

Proof. Let V, U be rotations such that $Vx = (||x||, 0..., 0)^T$ and $Uy = (||y||, 0...0)^T$. Now let $\tilde{W} = UWV^T$, and let $\tilde{W} = {}^{||}\tilde{W}_{||} + {}^{||}\tilde{W}_{\perp} + {}^{\perp}\tilde{W}_{||} + {}^{\perp}\tilde{W}_{\perp}$, with ${}^{||}\tilde{W}_{||}$ having non-zero term exactly $\tilde{W}_{11}$, ${}^{||}\tilde{W}_{\perp}$ having non-zero entries exactly $\tilde{W}_{1i}$ for $2 \leq i \leq k$. Finally, we let ${}^{\perp}\tilde{W}_{||}$ have non-zero entries exactly $\tilde{W}_{i1}$, with $2 \leq i \leq k$ and ${}^{\perp}\tilde{W}_{\perp}$ have the remaining entries non-zero.

If we define $\tilde{x} = Vx$ and $\tilde{y} = Uy$, then we see that

$$
\begin{aligned}
{}^{||}\tilde{W}_{\perp}\tilde{x} &= 0 & \quad\quad {}^{\perp}\tilde{W}_{\perp}\tilde{x} &= 0 \\
\tilde{y}^T\,{}^{\perp}\tilde{W}_{||} &= 0 & \quad\quad \tilde{y}^T\,{}^{\perp}\tilde{W}_{\perp} &= 0
\end{aligned}
$$

as $\tilde{x}, \tilde{y}$ have only one non-zero term, which does not correspond to a non-zero term in the components of $\tilde{W}$ in the equations.

Then, defining ${}^{||}W_{||} = U^T\,{}^{||}\tilde{W}_{||}V$, and the other components analogously, we get equations of the form

$$
{}^{||}W_{\perp}x = U^T\,{}^{||}\tilde{W}_{\perp}Vx = U^T\,{}^{||}\tilde{W}_{\perp}\tilde{x} = 0
$$

□

Observation 1. Given W, x as before, and considering $W_{||}$, $W_{\perp}$ with respect to x (wlog a unit vector) we can express them directly in terms of W as follows: Letting $W^{(i)}$ be the ith row of W, we have

$$
W_{||} = \begin{pmatrix} ((W^{(0)})^T \cdot x)x \\ \vdots \\ ((W^{(k)})^T \cdot x)x \end{pmatrix}
$$

i.e. the projection of each row in the direction of x. And of course

$$
W_{\perp} = W - W_{||}
$$

The motivation to consider such a decomposition of W is for the resulting independence between different components, as shown in the following lemma.

Lemma 2. Independence of Projections *Let x be a given vector (wlog of unit norm.) If W is a random matrix with $W_{ij} \sim \mathcal{N}(0, \sigma^2)$, then $W_{||}$ and $W_{\perp}$ with respect to x are independent random variables.*

Proof. There are two possible proof methods:

(a) We use the rotational invariance of random Gaussian matrices, i.e. if W is a Gaussian matrix, iid entries $\mathcal{N}(0, \sigma^2)$, and R is a rotation, then RW is also iid Gaussian, entries $\mathcal{N}(0, \sigma^2)$. (This follows easily from affine transformation rules for multivariate Gaussians.)

Let V be a rotation as in Lemma 1. Then $\tilde{W} = WV^T$ is also iid Gaussian, and furthermore, $\tilde{W}_{||}$ and $\tilde{W}_{\perp}$ partition the entries of $\tilde{W}$, so are evidently independent. But then $W_{||} = \tilde{W}_{||}V^T$ and $W_{\perp} = \tilde{W}_{\perp}V^T$ are also independent.

(b) From the observation note that $W_{||}$ and $W_{\perp}$ have a centered multivariate joint Gaussian distribution (both consist of linear combinations of the entries W_{ij} in W.) So it suffices to show that $W_{||}$ and $W_{\perp}$ have covariance 0. Because both are centered Gaussians, this is equivalent to showing $\mathbb{E}(< W_{||}, W_{\perp} >) = 0$. We have that

$$
\mathbb{E}(< W_{||}, W_{\perp} >) = \mathbb{E}(W_{||}W_{\perp}^T) = \mathbb{E}(W_{||}W^T) - \mathbb{E}(W_{||}W_{||}^T)
$$

As any two rows of W are independent, we see from the observation that $\mathbb{E}(W_{||}W^T)$ is a diagonal matrix, with the ith diagonal entry just $((W^{(0)})^T \cdot x)^2$. But similarly, $\mathbb{E}(W_{||}W_{||}^T)$ is also a diagonal matrix, with the same diagonal entries - so the claim follows.

□

In the following two lemmas, we use the rotational invariance of Gaussians as well as the chi distribution to prove results about the expected norm of a random Gaussian vector.

Lemma 3. Norm of a Gaussian vector *Let $X \in \mathbb{R}^k$ be a random Gaussian vector, with X_i iid, $\sim \mathcal{N}(0, \sigma^2)$. Then*

$$\mathbb{E}\left[||X||\right] = \sigma\sqrt{2}\frac{\Gamma((k+1)/2)}{\Gamma(k/2)}$$

Proof. We use the fact that if Y is a random Gaussian, and $Y_i \sim \mathcal{N}(0,1)$ then $||Y||$ follows a chi distribution. This means that $\mathbb{E}(||X/\sigma||) = \sqrt{2}\Gamma((k+1)/2)/\Gamma(k/2)$, the mean of a chi distribution with k degrees of freedom, and the result follows by noting that the expectation in the lemma is σ multiplied by the above expectation. □

We will find it useful to bound ratios of the Gamma function (as appear in Lemma 3) and so introduce the following inequality, from (Kershaw, 1983) that provides an extension of Gautschi's Inequality.

Theorem 6. An Extension of Gautschi's Inequality *For $0 < s < 1$, we have*

$$\left(x + \frac{s}{2}\right)^{1-s} \leq \frac{\Gamma(x+1)}{\Gamma(x+s)} \leq \left(x - \frac{1}{2} + \left(s + \frac{1}{4}\right)^{\frac{1}{2}}\right)^{1-s}$$

We now show:

Lemma 4. Norm of Projections *Let W be a k by k random Gaussian matrix with iid entries $\sim \mathcal{N}(0, \sigma^2)$, and x, y two given vectors. Partition W into components as in Lemma 1 and let $x_\perp$ be a nonzero vector perpendicular to x. Then*

(a)

$$\mathbb{E}\left[\left|\left|{}^{\perp}W_\perp x_\perp\right|\right|\right] = ||x_\perp||\,\sigma\sqrt{2}\frac{\Gamma(k/2)}{\Gamma((k-1)/2} \geq ||x_\perp||\,\sigma\sqrt{2}\left(\frac{k}{2} - \frac{3}{4}\right)^{1/2}$$

(b) If $1_{\mathcal{A}}$ is an identity matrix with non-zeros diagonal entry i iff $i \in \mathcal{A} \subset [k]$, and $|A| > 2$, then

$$\mathbb{E}\left[\left|\left|1_{\mathcal{A}}{}^{\perp}W_\perp x_\perp\right|\right|\right] \geq ||x_\perp||\,\sigma\sqrt{2}\frac{\Gamma(|\mathcal{A}|/2)}{\Gamma((|\mathcal{A}|-1)/2)} \geq ||x_\perp||\,\sigma\sqrt{2}\left(\frac{|\mathcal{A}|}{2} - \frac{3}{4}\right)^{1/2}$$

Proof. (a) Let $U, V, \tilde{W}$ be as in Lemma 1. As U, V are rotations, $\tilde{W}$ is also iid Gaussian. Furthermore for any fixed W, with $\tilde{a} = Va$, by taking inner products, and square-rooting, we see that $\left|\left|\tilde{W}\tilde{a}\right|\right| = ||Wa||$. So in particular

$$\mathbb{E}\left[\left|\left|{}^{\perp}W_\perp x_\perp\right|\right|\right] = \mathbb{E}\left[\left|\left|{}^{\perp}\tilde{W}_\perp \tilde{x}_\perp\right|\right|\right]$$

But from the definition of non-zero entries of ${}^{\perp}\tilde{W}_\perp$, and the form of $\tilde{x}_\perp$ (a zero entry in the first coordinate), it follows that ${}^{\perp}\tilde{W}_\perp\tilde{x}_\perp$ has exactly $k-1$ non zero entries, each a centered Gaussian with variance $(k-1)\sigma^2\,||x_\perp||^2$. By Lemma 3, the expected norm is as in the statement. We then apply Theorem 6 to get the lower bound.

(b) First note we can view $1_{\mathcal{A}}{}^{\perp}W_\perp = {}^{\perp}1_{\mathcal{A}}W_\perp$. (Projecting down to a random (as W is random) subspace of fixed size $|\mathcal{A}| = m$ and then making perpendicular commutes with making perpendicular and then projecting everything down to the subspace.)

So we can view W as a random m by k matrix, and for x, y as in Lemma 1 (with y projected down onto m dimensions), we can again define U, V as k by k and m by m rotation matrices respectively, and $\tilde{W} = UWV^T$, with analogous properties to Lemma 1. Now we can finish as in part (a), except that ${}^{\perp}\tilde{W}_{\perp}\tilde{x}$ may have only $m-1$ entries, (depending on whether y is annihilated by projecting down by$1_{\mathcal{A}}$) each of variance $(k-1)\sigma^2 \left|\left|x_{\perp}\right|\right|^2$.

□

Lemma 5. Norm and Translation *Let X be a centered multivariate Gaussian, with diagonal covariance matrix, and μ a constant vector.*

$$\mathbb{E}(||X-\mu||) \geq \mathbb{E}(||X||)$$

Proof. The inequality can be seen intuitively geometrically: as X has diagonal covariance matrix, the contours of the pdf of $||X||$ are circular centered at 0, decreasing radially. However, the contours of the pdf of $||X-\mu||$ are shifted to be centered around $||\mu||$, and so shifting back μ to 0 reduces the norm.

A more formal proof can be seen as follows: let the pdf of X be $f_X(\cdot)$. Then we wish to show

$$\int_x ||x-\mu||\, f_X(x)dx \geq \int_x ||x||\, f_X(x)dx$$

Now we can pair points $x, -x$, using the fact that $f_X(x) = f_X(-x)$ and the triangle inequality on the integrand to get

$$\int_{|x|} \left(||x-\mu|| + ||-x-\mu||\right) f_X(x)dx \geq \int_{|x|} ||2x||\, f_X(x)dx = \int_{|x|} \left(||x|| + ||-x||\right) f_X(x)dx$$

□

B.2. Proof of Theorem

We use $v_i^{(d)}$ to denote the i^{th} neuron in hidden layer d. We also let $x = z^{(0)}$ be an input, $h^{(d)}$ be the hidden representation at layer d, and ϕ the non-linearity. The weights and bias are called $W^{(d)}$ and $b^{(d)}$ respectively. So we have the relations

$$h^{(d)} = W^{(d)}z^{(d)} + b^{(d)}, \qquad\qquad z^{(d+1)} = \phi(h^{(d)}). \tag{1}$$

Proof. We first prove the zero bias case. To do so, it is sufficient to prove that

$$\mathbb{E}\left[\left|\left|\delta z^{(d+1)}(t)\right|\right|\right] \geq O\left(\left(\frac{\sqrt{\sigma k}}{\sqrt{\sigma + k}}\right)^{d+1}\right)\left|\left|\delta z^{(0)}(t)\right|\right| \tag{**}$$

as integrating over t gives us the statement of the theorem.

For ease of notation, we will suppress the t in $z^{(d)}(t)$.

We first write

$$W^{(d)} = W_{\perp}^{(d)} + W_{||}^{(d)}$$

where the division is done with respect to $z^{(d)}$. Note that this means $h^{(d+1)} = W_{||}^{(d)}z^{(d)}$ as the other component annihilates (maps to 0) $z^{(d)}$.

We can also define $\mathcal{A}_{W_{||}^{(d)}} = \{i : i \in [k], |h_i^{(d+1)}| < 1\}$ i.e. the set of indices for which the hidden representation is not saturated. Letting W_i denote the ith row of matrix W, we now claim that:

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z^{(d+1)}\right|\right|\right] = \mathbb{E}_{W_{||}^{(d)}}\mathbb{E}_{W_{\perp}^{(d)}}\left[\left(\sum_{i\in\mathcal{A}_{W_{||}^{(d)}}}\left((W_{\perp}^{(d)})_i\delta z^{(d)} + (W_{||}^{(d)})_i\delta z^{(d)}\right)^2\right)^{1/2}\right] \tag{*}$$

Indeed, by Lemma 2 we first split the expectation over $W^{(d)}$ into a tower of expectations over the two independent parts of W to get

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z^{(d+1)}\right|\right|\right] = \mathbb{E}_{W_{||}^{(d)}}\mathbb{E}_{W_{\perp}^{(d)}}\left[\left|\left|\phi(W^{(d)}\delta z^{(d)})\right|\right|\right]$$

But conditioning on $W_{\|}^{(d)}$ in the inner expectation gives us $h^{(d+1)}$ and $\mathcal{A}_{W_{\|}^{(d)}}$, allowing us to replace the norm over $\phi(W^{(d)}\delta z^{(d)})$ with the sum in the term on the right hand side of the claim.

Till now, we have mostly focused on partitioning the matrix $W^{(d)}$. But we can also set $\delta z^{(d)} = \delta z_{\|}^{(d)} + \delta z_{\perp}^{(d)}$ where the perpendicular and parallel are with respect to $z^{(d)}$. In fact, to get the expression in (**), we derive a recurrence as below:

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z_{\perp}^{(d+1)}\right|\right|\right] \geq O\left(\frac{\sqrt{\sigma k}}{\sqrt{\sigma + k}}\right)\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]$$

To get this, we first need to define $\tilde{z}^{(d+1)} = 1_{\mathcal{A}_{W_{\|}^{(d)}}} h^{(d+1)}$ - the latent vector $h^{(d+1)}$ with all saturated units zeroed out.

We then split the column space of $W^{(d)} = {}^{\perp}W^{(d)} + {}^{\|}W^{(d)}$, where the split is with respect to $\tilde{z}^{(d+1)}$. Letting $\delta z_{\perp}^{(d+1)}$ be the part perpendicular to $z^{(d+1)}$, and $\mathcal{A}$ the set of units that are unsaturated, we have an important relation:

Claim

$$\left|\left|\delta z_{\perp}^{(d+1)}\right|\right| \geq \left|\left|{}^{\perp}W^{(d)}\delta z^{(d)} 1_{\mathcal{A}}\right|\right|$$

(where the indicator in the right hand side zeros out coordinates not in the active set.)

To see this, first note, by definition,

$$\delta z_{\perp}^{(d+1)} = W^{(d)}\delta z^{(d)} \cdot 1_{\mathcal{A}} - \langle W^{(d)}\delta z^{(d)} \cdot 1_{\mathcal{A}}, \hat{z}^{(d+1)}\rangle \hat{z}^{(d+1)} \tag{1}$$

where the $\hat{\cdot}$ indicates a unit vector.

Similarly

$${}^{\perp}W^{(d)}\delta z^{(d)} = W^{(d)}\delta z^{(d)} - \langle W^{(d)}\delta z^{(d)}, \hat{\tilde{z}}^{(d+1)}\rangle \hat{\tilde{z}}^{(d+1)} \tag{2}$$

Now note that for any index $i \in \mathcal{A}$, the right hand sides of (1) and (2) are identical, and so the vectors on the left hand side agree for all $i \in \mathcal{A}$. In particular,

$$\delta z_{\perp}^{(d+1)} \cdot 1_{\mathcal{A}} = {}^{\perp}W^{(d)}\delta z^{(d)} \cdot 1_{\mathcal{A}}$$

Now the claim follows easily by noting that $\left|\left|\delta z_{\perp}^{(d+1)}\right|\right| \geq \left|\left|\delta z_{\perp}^{(d+1)} \cdot 1_{\mathcal{A}}\right|\right|$.

Returning to (*), we split $\delta z^{(d)} = \delta z_{\perp}^{(d)} + \delta z_{\|}^{(d)}$, $W_{\perp}^{(d)} = {}^{\|}W_{\perp}^{(d)} + {}^{\perp}W_{\perp}^{(d)}$ (and $W_{\|}^{(d)}$ analogously), and after some cancellation, we have

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z^{(d+1)}\right|\right|\right] = \mathbb{E}_{W_{\|}^{(d)}}\mathbb{E}_{W_{\perp}^{(d)}}\left[\left(\sum_{i \in \mathcal{A}_{W_{\|}^{(d)}}}\left(({}^{\perp}W_{\perp}^{(d)} + {}^{\|}W_{\perp}^{(d)})_i \delta z_{\perp}^{(d)} + ({}^{\perp}W_{\|}^{(d)} + {}^{\|}W_{\|}^{(d)})_i \delta z_{\|}^{(d)}\right)^2\right)^{1/2}\right]$$

We would like a recurrence in terms of only perpendicular components however, so we first drop the ${}^{\|}W_{\perp}^{(d)}$, ${}^{\|}W_{\|}^{(d)}$ (which can be done without decreasing the norm as they are perpendicular to the remaining terms) and using the above claim, have

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z_{\perp}^{(d+1)}\right|\right|\right] \geq \mathbb{E}_{W_{\|}^{(d)}}\mathbb{E}_{W_{\perp}^{(d)}}\left[\left(\sum_{i \in \mathcal{A}_{W_{\|}^{(d)}}}\left(({}^{\perp}W_{\perp}^{(d)})_i \delta z_{\perp}^{(d)} + ({}^{\perp}W_{\|}^{(d)})_i \delta z_{\|}^{(d)}\right)^2\right)^{1/2}\right]$$

But in the inner expectation, the term ${}^{\perp}W_{\parallel}^{(d)}\delta z_{\parallel}^{(d)}$ is just a constant, as we are conditioning on $W_{\parallel}^{(d)}$. So using Lemma 5 we have

$$\mathbb{E}_{W_{\perp}^{(d)}}\left[\left(\sum_{i\in\mathcal{A}_{W_{\parallel}^{(d)}}}\left(({}^{\perp}W_{\perp}^{(d)})_i\delta z_{\perp}^{(d)}+({}^{\perp}W_{\parallel}^{(d)})_i\delta z_{\parallel}^{(d)}\right)^2\right)^{1/2}\right]\geq\mathbb{E}_{W_{\perp}^{(d)}}\left[\left(\sum_{i\in\mathcal{A}_{W_{\parallel}^{(d)}}}\left(({}^{\perp}W_{\perp}^{(d)})_i\delta z_{\perp}^{(d)}\right)^2\right)^{1/2}\right]$$

We can then apply Lemma 4 to get

$$\mathbb{E}_{W_{\perp}^{(d)}}\left[\left(\sum_{i\in\mathcal{A}_{W_{\parallel}^{(d)}}}\left(({}^{\perp}W_{\perp}^{(d)})_i\delta z_{\perp}^{(d)}\right)^2\right)^{1/2}\right]\geq\frac{\sigma}{\sqrt{k}}\sqrt{2}\frac{\sqrt{2|\mathcal{A}_{W_{\parallel}^{(d)}}|-3}}{2}\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]$$

The outer expectation on the right hand side only affects the term in the expectation through the size of the active set of units. For ReLUs, $p=\mathbb{P}(h_i^{(d+1)}>0)$ and for hard tanh, we have $p=\mathbb{P}(|h_i^{(d+1)}|<1)$, and noting that we get a non-zero norm only if $|\mathcal{A}_{W_{\parallel}^{(d)}}|\geq 2$ (else we cannot project down a dimension), and for $|\mathcal{A}_{W_{\parallel}^{(d)}}|\geq 2$,

$$\sqrt{2}\frac{\sqrt{2|\mathcal{A}_{W_{\parallel}^{(d)}}|-3}}{2}\geq\frac{1}{\sqrt{2}}\sqrt{|\mathcal{A}_{W_{\parallel}^{(d)}}|}$$

we get

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z_{\perp}^{(d+1)}\right|\right|\right]\geq\frac{1}{\sqrt{2}}\left(\sum_{j=2}^{k}\binom{k}{j}p^j(1-p)^{k-j}\frac{\sigma}{\sqrt{k}}\sqrt{j}\right)\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]$$

We use the fact that we have the probability mass function for an (k,p) binomial random variable to bound the $\sqrt{j}$ term:

$$\begin{aligned}\sum_{j=2}^{k}\binom{k}{j}p^j(1-p)^{k-j}\frac{\sigma}{\sqrt{k}}\sqrt{j}&=-\binom{k}{1}p(1-p)^{k-1}\frac{\sigma}{\sqrt{k}}+\sum_{j=0}^{k}\binom{k}{j}p^j(1-p)^{k-j}\frac{\sigma}{\sqrt{k}}\sqrt{j}\\&=-\sigma\sqrt{k}p(1-p)^{k-1}+kp\cdot\frac{\sigma}{\sqrt{k}}\sum_{j=1}^{k}\frac{1}{\sqrt{j}}\binom{k-1}{j-1}p^{j-1}(1-p)^{k-j}\end{aligned}$$

But by using Jensen's inequality with $1/\sqrt{x}$, we get

$$\sum_{j=1}^{k}\frac{1}{\sqrt{j}}\binom{k-1}{j-1}p^{j-1}(1-p)^{k-j}\geq\frac{1}{\sqrt{\sum_{j=1}^{k}j\binom{k-1}{j-1}p^{j-1}(1-p)^{k-j}}}=\frac{1}{\sqrt{(k-1)p+1}}$$

where the last equality follows by recognising the expectation of a binomial$(k-1,p)$ random variable. So putting together, we get

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z_{\perp}^{(d+1)}\right|\right|\right]\geq\frac{1}{\sqrt{2}}\left(-\sigma\sqrt{k}p(1-p)^{k-1}+\sigma\cdot\frac{\sqrt{k}p}{\sqrt{1+(k-1)p}}\right)\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]\qquad\text{(a)}$$

From here, we must analyse the hard tanh and ReLU cases separately. First considering the hard tanh case:

To lower bound p, we first note that as $h_i^{(d+1)}$ is a normal random variable with variance $\leq\sigma^2$, if $A\sim\mathcal{N}(0,\sigma^2)$

$$\mathbb{P}(|h_i^{(d+1)}|<1)\geq\mathbb{P}(|A|<1)\geq\frac{1}{\sigma\sqrt{2\pi}}\qquad\text{(b)}$$

where the last inequality holds for $\sigma \geq 1$ and follows by Taylor expanding $e^{-x^2/2}$ around 0. Similarly, we can also show that $p \leq \frac{1}{\sigma}$.

So this becomes

$$\mathbb{E}\left[\left|\left|\delta z^{(d+1)}\right|\right|\right] \geq \left(\frac{1}{\sqrt{2}}\left(\frac{1}{(2\pi)^{1/4}}\frac{\sqrt{\sigma k}}{\sqrt{\sigma\sqrt{2\pi}+(k-1)}} - \sqrt{k}\left(1-\frac{1}{\sigma}\right)^{k-1}\right)\right)\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]$$
$$= O\left(\frac{\sqrt{\sigma k}}{\sqrt{\sigma + k}}\right)\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]$$

Finally, we can compose this, to get

$$\mathbb{E}\left[\left|\left|\delta z^{(d+1)}\right|\right|\right] \geq \left(\frac{1}{\sqrt{2}}\left(\frac{1}{(2\pi)^{1/4}}\frac{\sqrt{\sigma k}}{\sqrt{\sigma\sqrt{2\pi}+(k-1)}} - \sqrt{k}\left(1-\frac{1}{\sigma}\right)^{k-1}\right)\right)^{d+1} c \cdot ||\delta x(t)|| \qquad \text{(c)}$$

with the constant c being the ratio of $||\delta x(t)_{\perp}||$ to $||\delta x(t)||$. So if our trajectory direction is almost orthogonal to $x(t)$ (which will be the case for e.g. random circular arcs, c can be seen to be ≈ 1 by splitting into components as in Lemma 1, and using Lemmas 3, 4.)

The ReLU case (with no bias) is even easier. Noting that for random weights, $p = 1/2$, and plugging in to equation (a), we get

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z_{\perp}^{(d+1)}\right|\right|\right] \geq \frac{1}{\sqrt{2}}\left(\frac{-\sigma\sqrt{k}}{2^k} + \sigma \cdot \frac{\sqrt{k}}{\sqrt{2(k+1)}}\right)\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right] \qquad \text{(d)}$$

But the expression on the right hand side has exactly the asymptotic form $O(\sigma\sqrt{k}/\sqrt{k+1})$, and we finish as in (c).

Result for non-zero bias In fact, we can easily extend the above result to the case of non-zero bias. The insight is to note that because $\delta z^{(d+1)}$ involves taking a *difference* between $z^{(d+1)}(t+dt)$ and $z^{(d+1)}(t)$, the bias term does not enter at all into the expression for $\delta z^{(d+1)}$. So the computations above hold, and equation (a) becomes

$$\mathbb{E}_{W^{(d)}}\left[\left|\left|\delta z_{\perp}^{(d+1)}\right|\right|\right] \geq \frac{1}{\sqrt{2}}\left(-\sigma_w\sqrt{k}p(1-p)^{k-1} + \sigma_w \cdot \frac{\sqrt{k}p}{\sqrt{1+(k-1)p}}\right)\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]$$

For ReLUs, we require $h_i^{(d+1)} = w_i^{(d+1)} z_i^{(d)} + b_i^{(d+1)} > 0$ where the bias and weight are drawn from $\mathcal{N}(0, \sigma_b^2)$ and $\mathcal{N}(0, \sigma_w^2)$ respectively. But with $p \geq 1/4$, this holds as the signs for w, b are purely random. Substituting in and working through results in the *same* asymptotic behavior as without bias.

For hard tanh, not that as $h_i^{(d+1)}$ is a normal random variable with variance $\leq \sigma_w^2 + \sigma_b^2$ (as equation (b) becomes

$$\mathbb{P}(|h_i^{(d+1)}| < 1) \geq \frac{1}{\sqrt{(\sigma_w^2+\sigma_b^2)}\sqrt{2\pi}}$$

This gives Theorem 3

$$\mathbb{E}\left[\left|\left|\delta z^{(d+1)}\right|\right|\right] \geq O\left(\frac{\sigma_w}{(\sigma_w^2+\sigma_b^2)^{1/4}} \cdot \frac{\sqrt{k}}{\sqrt{\sqrt{\sigma_w^2+\sigma_b^2}+k}}\right)\mathbb{E}\left[\left|\left|\delta z_{\perp}^{(d)}\right|\right|\right]$$

□

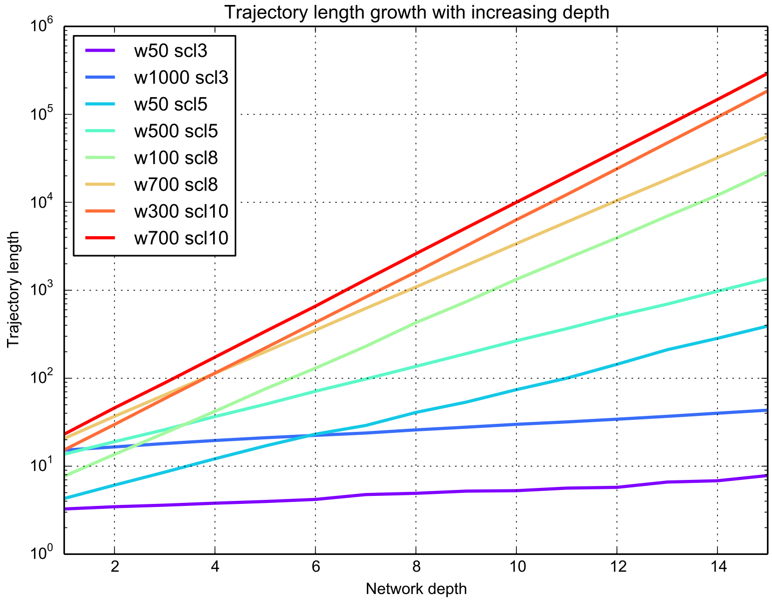


Figure 12. The figure above shows trajectory growth with different initialization scales as a trajectory is propagated through a fully connected network for MNIST, with Relu activations. Note that as described by the bound in Theorem 3 we see that trajectory growth is 1) exponential in depth 2) increases with initialization scale and width, 3) increases faster with scale over width, as expected from σ_w compared to $\sqrt{k/(k+1)}$ in the Theorem.

Statement and Proof of Upper Bound for Trajectory Growth for Hard Tanh Replace hard-tanh with a linear coordinate-wise identity map, $h_i^{(d+1)} = (W^{(d)}z^{(d)})_i + b_i$. This provides an upper bound on the norm. We also then recover a chi distribution with k terms, each with standard deviation $\frac{\sigma_w}{k^{\frac{1}{2}}}$,

$$\mathbb{E}\left[\left|\left|\delta z^{(d+1)}\right|\right|\right] \leq \sqrt{2}\frac{\Gamma\left((k+1)/2\right)}{\Gamma\left(k/2\right)}\frac{\sigma_w}{k^{\frac{1}{2}}}\left|\left|\delta z^{(d)}\right|\right| \tag{2}$$

$$\leq \sigma_w\left(\frac{k+1}{k}\right)^{\frac{1}{2}}\left|\left|\delta z^{(d)}\right|\right|, \tag{3}$$

where the second step follows from (Laforgia and Natalini, 2013), and holds for $k > 1$.

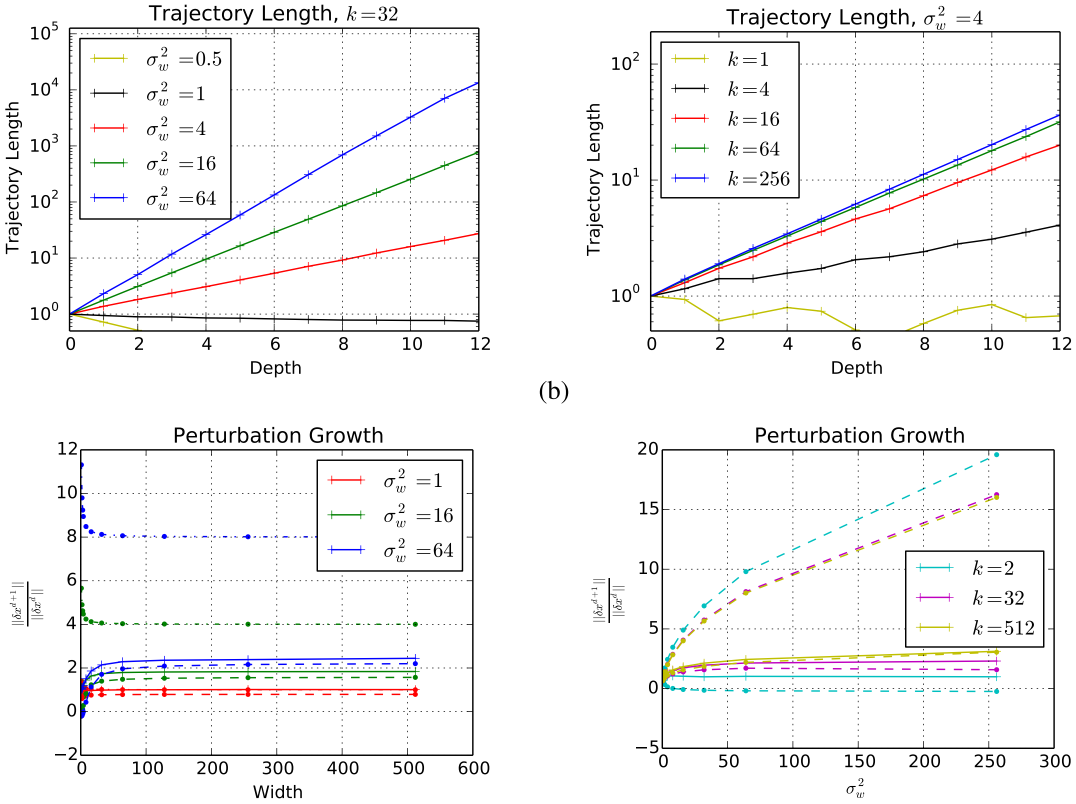
Proof of Theorem 4

Proof. **For $\sigma_b = 0$:**

For hidden layer $d < n$, consider neuron $v_1^{(d)}$. This has as input $\sum_{i=1}^k W_{i1}^{(d-1)} z_i^{(d-1)}$. As we are in the large σ case, we assume that $|z_i^{(d-1)}| = 1$. Furthermore, as signs for $z_i^{(d-1)}$ and $W_{i1}^{(d-1)}$ are both completely random, we can also assume wlog that $z_i^{(d-1)} = 1$. For a particular input, we can define $v_1^{(d)}$ as *sensitive* to $v_i^{(d-1)}$ if $v_i^{(d-1)}$ transitioning (to wlog -1) will induce a transition in node $v_1^{(d)}$. A sufficient condition for this to happen is if $|W_{i1}| \geq |\sum_{j\neq i} W_{j1}|$. But $X = W_{i1} \sim \mathcal{N}(0, \sigma^2/k)$ and $\sum_{j\neq i} W_{j1} = Y' \sim \mathcal{N}(0, (k-1)\sigma^2/k)$. So we want to compute $\mathbb{P}(|X| > |Y'|)$. For ease of computation, we instead look at $\mathbb{P}(|X| > |Y|)$, where $Y \sim \mathcal{N}(0, \sigma^2)$.

But this is the same as computing $\mathbb{P}(|X|/|Y| > 1) = \mathbb{P}(X/Y < -1) + \mathbb{P}(X/Y > 1)$. But the ratio of two centered independent normals with variances σ_1^2, σ_2^2 follows a Cauchy distribution, with parameter σ_1/σ_2, which in this case is $1/\sqrt{k}$. Substituting this in to the cdf of the Cauchy distribution, we get that

$$\mathbb{P}\left(\frac{|X|}{|Y|} > 1\right) = 1 - \frac{2}{\pi}\arctan(\sqrt{k})$$



(a)

(b)

(c)

(d)

Figure 13. The exponential growth of trajectory length with depth, in a random deep network with hard-tanh nonlinearities. A circular trajectory is chosen between two random vectors. The image of that trajectory is taken at each layer of the network, and its length measured. *(a,b)* The trajectory length vs. layer, in terms of the network width k and weight variance σ_w^2, both of which determine its growth rate. *(c,d)* The average ratio of a trajectory's length in layer $d+1$ relative to its length in layer d. The solid line shows simulated data, while the dashed lines show upper and lower bounds (Theorem 3). Growth rate is a function of layer width k, and weight variance σ_w^2.

Finally, using the identity $\arctan(x) + \arctan(1/x)$ and the Laurent series for $\arctan(1/x)$, we can evaluate the right hand side to be $O(1/\sqrt{k})$. In particular

$$\mathbb{P}\left(\frac{|X|}{|Y|} > 1\right) \geq O\left(\frac{1}{\sqrt{k}}\right) \tag{c}$$

This means that in expectation, any neuron in layer d will be sensitive to the transitions of $\sqrt{k}$ neurons in the layer below. Using this, and the fact the while $v_i^{(d-1)}$ might flip very quickly from say -1 to 1, the gradation in the transition ensures that neurons in layer d sensitive to $v_i^{(d-1)}$ will transition at distinct times, we get the desired growth rate in expectation as follows:

Let $T^{(d)}$ be a random variable denoting the number of transitions in layer d. And let $T_i^{(d)}$ be a random variable denoting the number of transitions of neuron i in layer d. Note that by linearity of expectation and symmetry, $\mathbb{E}\left[T^{(d)}\right] = \sum_i \mathbb{E}\left[T_i^{(d)}\right] = k\mathbb{E}\left[T_1^{(d)}\right]$

Now, $\mathbb{E}\left[T_1^{(d+1)}\right] \geq \mathbb{E}\left[\sum_i 1_{(1,i)} T_i^{(d)}\right] = k\mathbb{E}\left[1_{(1,1)} T_1^{(d)}\right]$ where $1_{(1,i)}$ is the indicator function of neuron 1 in layer $d+1$ being sensitive to neuron i in layer d.

But by the independence of these two events, $\mathbb{E}\left[1_{(1,1)} T_1^{(d)}\right] = \mathbb{E}\left[1_{(1,1)}\right] \cdot \mathbb{E}\left[T_1^{(d)}\right]$. But the firt time on the right hand side is $O(1/\sqrt{k})$ by (c), so putting it all together, $\mathbb{E}\left[T_1^{(d+1)}\right] \geq \sqrt{k}\mathbb{E}\left[T_1^{(d)}\right]$.

Written in terms of the entire layer, we have $\mathbb{E}\left[T^{(d+1)}\right] \geq \sqrt{k}\mathbb{E}\left[T^{(d)}\right]$ as desired.

For $\sigma_b > 0$:

We replace $\sqrt{k}$ with $\sqrt{k(1 + \sigma_b^2/\sigma_w^2)}$, by noting that $Y \sim \mathcal{N}(0, \sigma_w^2 + \sigma_b^2)$. This results in a growth rate of form $O(\sqrt{k}/\sqrt{1 + \frac{\sigma_b^2}{\sigma_w^2}})$. □

B.3. Dichotomies: a natural dual

Our measures of expressivity have mostly concentrated on sweeping the input along a trajectory $x(t)$ and taking measures of $F_A(x(t); W)$. Instead, we can also sweep the weights W along a trajectory $W(t)$, and look at the consequences (e.g. binary labels – i.e. *dichotomies*), say for a fixed set of inputs $x_1, ..., x_s$.

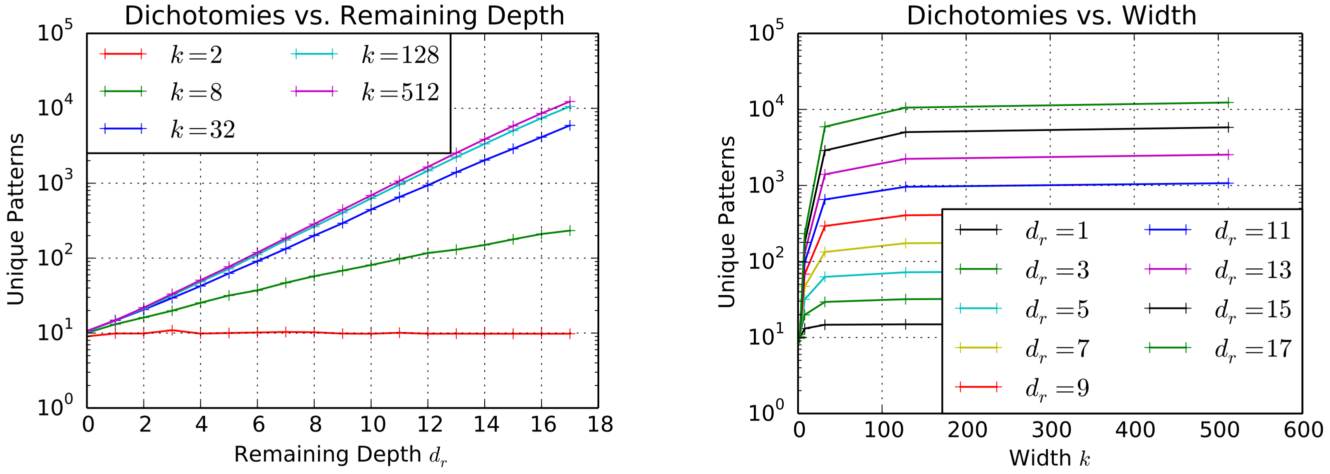
In fact, after random initialization, sweeping the first layer weights is statistically very similar to sweeping the input along a trajectory $x(t)$. In particular, letting W' denote the first layer weights, for a particular input x_0, $x_0 W'$ is a vector, each coordinate is iid, $\sim \mathcal{N}(0, ||x_0||^2 \sigma_w^2)$. Extending this observation, we see that (providing norms are chosen appropriately), $x_0 W' \cos(t) + x_1 W' \sin(t)$ (fixed x_0, x_1, W) has the same distribution as $x_0 W_0' \cos(t) + x_0 W_1' \sin(t)$ (fixed x_0, W_0', W_1').

So we expect that there will be similarities between results for sweeping weights and for sweeping input trajectories, which we explore through some synthetic experiments, primarily for hard tanh, in Figures 15, 16. We find that the proportionality of transitions to trajectory length extends to dichotomies, as do results on the expressive power afforded by remaining depth.

For non-random inputs and non-random functions, this is a well known question upper bounded by the Sauer-Shelah lemma (Sauer, 1972). We discuss this further in Appendix **??**. In the random setting, the statistical duality of weight sweeping and input sweeping suggests a direct proportion to transitions and trajectory length for a fixed input. Furthermore, if the $x_i \in S$ are sufficiently uncorrelated (e.g. random) class label transitions should occur independently for each x_i Indeed, we show this in Figure 14.

C. Addtional Experiments from Section 3

Here we include additional experiments from Section 3



(a) (b)

Figure 14. We sweep the weights W of a layer through a trajectory $W(t)$ and count the number of labellings over a set of datapoints. When W is the first layer, this is statistically identical to sweeping the input through $x(t)$ (see Appendix). Thus, similar results are observed, with exponential increase with the depth of an architecture, and much slower increase with width. Here we plot the number of classification dichotomies over $s = 15$ input vectors achieved by sweeping the first layer weights in a hard-tanh network along a one-dimensional great circle trajectory. We show this *(a)* as a function of depth for several widths, and *(b)* as a function of width for several depths. All networks were generated with weight variance $\sigma_w^2 = 8$, and bias variance $\sigma_b^2 = 0$.

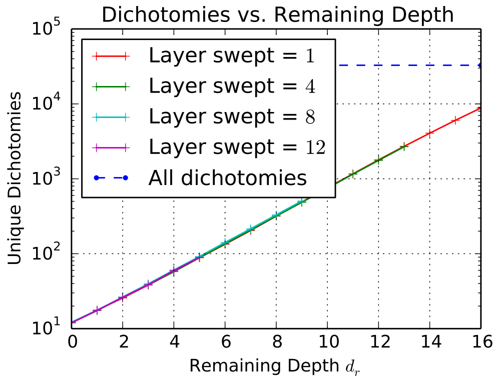


Figure 15. Expressive power depends only on remaining network depth. Here we plot the number of dichotomies achieved by sweeping the weights in different network layers through a 1-dimensional great circle trajectory, as a function of the remaining network depth. The number of achievable dichotomies does not depend on the total network depth, only on the number of layers above the layer swept. All networks had width $k = 128$, weight variance $\sigma_w^2 = 8$, number of datapoints $s = 15$, and hard-tanh nonlinearities. The blue dashed line indicates all 2^s possible dichotomies for this random dataset.

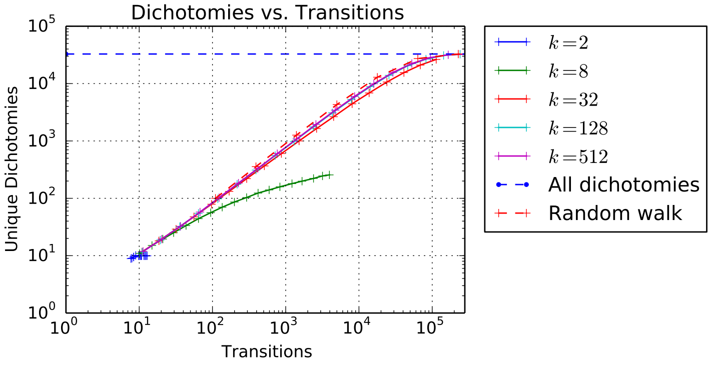


Figure 16. Here we plot the number of unique dichotomies that have been observed as a function of the number of transitions the network has undergone. Each datapoint corresponds to the number of transitions and dichotomies for a hard-tanh network of a different depth, with the weights in the first layer undergoing interpolation along a great circle trajectory $W^{(0)}(t)$. We compare these plots to a random walk simulation, where at each transition a single class label is flipped uniformly at random. Dichotomies are measured over a dataset consisting of $s = 15$ random samples, and all networks had weight variance $\sigma_w^2 = 16$. The blue dashed line indicates all 2^s possible dichotomies.

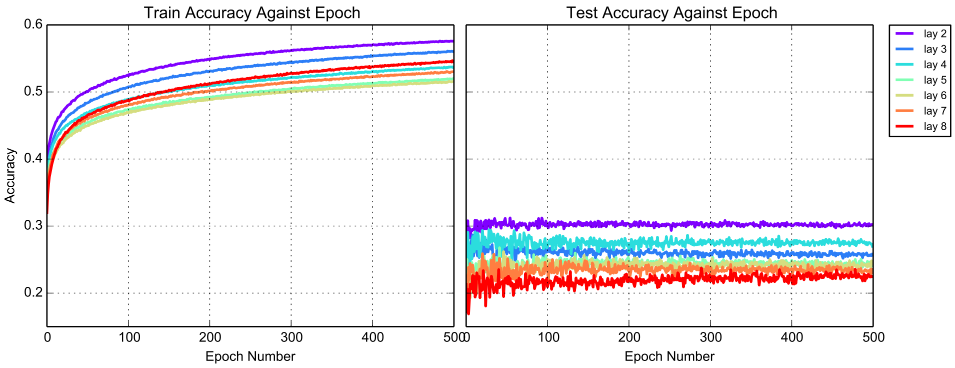


Figure 17. We repeat a similar experiment in Figure 7 with a fully connected network on CIFAR-10, and mostly observe that training lower layers again leads to better performance, although, as expected, overall performance is impacted by training only a single layer. The networks had width $k = 200$, weight variance $\sigma_w^2 = 1$, and hard-tanh nonlinearities. We again only train from the second hidden layer on so that the number of parameters remains fixed. The theory only applies to training error (the ability to fit a function), and generalisation accuracy remains low in this very constrained setting.

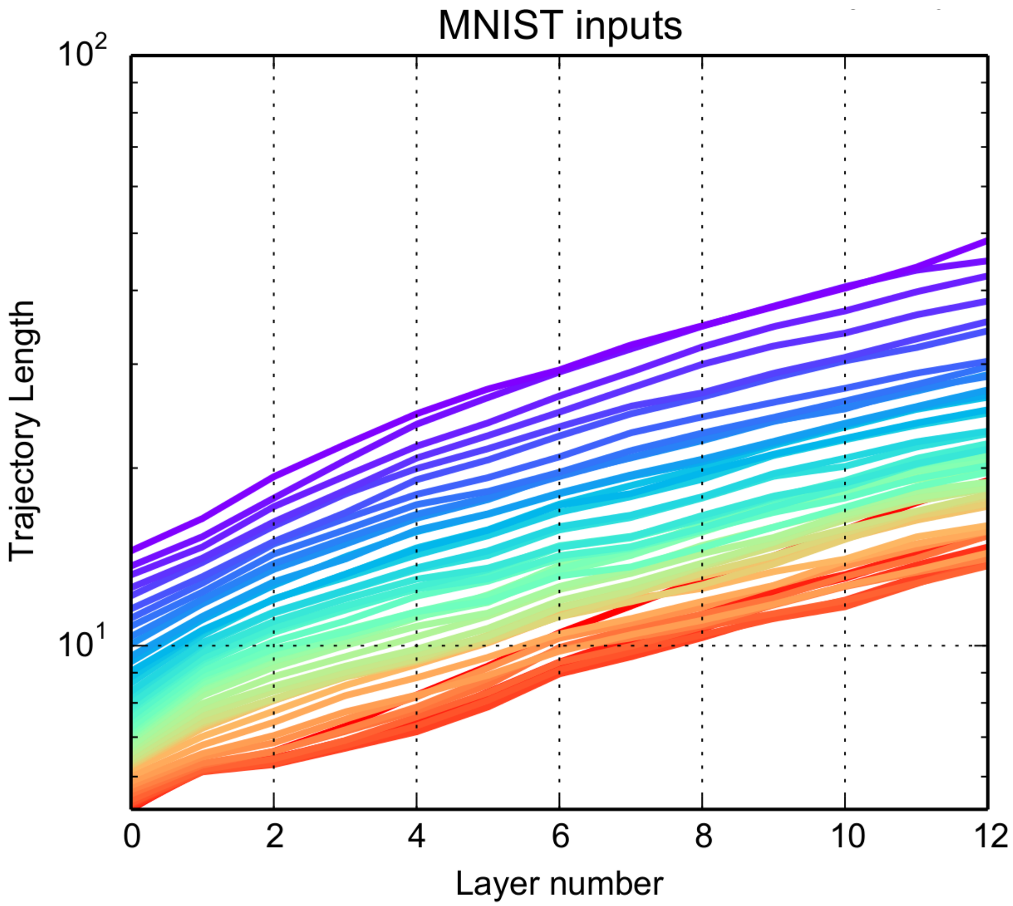


Figure 18. Training increases the trajectory length for smaller initialization values of σ_w. This experiment plots the growth of trajectory length as a circular interpolation between two MNIST datapoints is propagated through the network, at different train steps. Red indicates the start of training, with purple the end of training. We see that the training process *increases* trajectory length, likely to increase the expressivity of the input-output map to enable greater accuracy.