Mathematical
Institute

DNN input domain partitioning and random paths.

Theories of Deep Learning: C6.5, Video 5

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford
Mathematics

Prior to the approximation rate results from Telgarsky 15' and Yarotsky 16', there were qualitative geometric results showing showing potential for exponential expressivity:

- On the number of response regions of deep feedforward networks with piecewise linear activations (Pascanu et al. 14')
  `https://arxiv.org/pdf/1312.6098.pdf`
- On the expressive power of deep neural networks (Raghu et al. 16')
  `https://arxiv.org/abs/1606.05336`
- Trajectory growth lower bounds for random sparse deep ReLU networks (Price et al. 19')
  `https://arxiv.org/abs/1911.10651`

ReLU hyperplane arrangement

Partition of the input domain $\mathbb{R}^{n_0}$: one layer

Mathematical Institute

The action of ReLU to an affine transform is a linearly increasing function orthogonal to hyperplanes; let $W \in \mathbb{R}^{n_1 \times n_0}$ then:

$$H_i := \{x \in \mathbb{R}^{n_0} : \ W_i x + b_i = 0\} \quad \forall i \in [n_1]$$

where W_i is the i^{th} row of W.
The normals to these hyperplanes partition the input dimension n_0, and if W is in general position (all subsets of rows are maximal rank), then the number of partitions is:

$$\sum_{j=0}^{n_0} \binom{n_1}{j}$$

`https://arxiv.org/pdf/1312.6098.pdf`

Mathematical Institute

The number of partitions in one layer is lower bounded by

$$\sum_{j=0}^{n_0} \binom{n_1}{j} \geq n_1^{\min\{n_0, n_1/2\}}$$

and each hidden layers can further subdivide these regions:

Theorem (Pascanu et al. 14')

An L layer DNN with ReLU activation, input $\mathbb{R}^{n_0}$, and hidden layers of width $n_1, n_2, \ldots, n_L$ partitions the input space into at least

$$\Pi_{\ell=0}^{L} n_\ell^{\min\{n_0, n_\ell/2\}}$$

This shows an exponential dependence on depth L.

`https://arxiv.org/pdf/1312.6098.pdf`

ReLU hyperplane arrangement

Partition of the input domain $\mathbb{R}^{n_0}$: plot Pascanu et al. 14'

Mathematical Institute

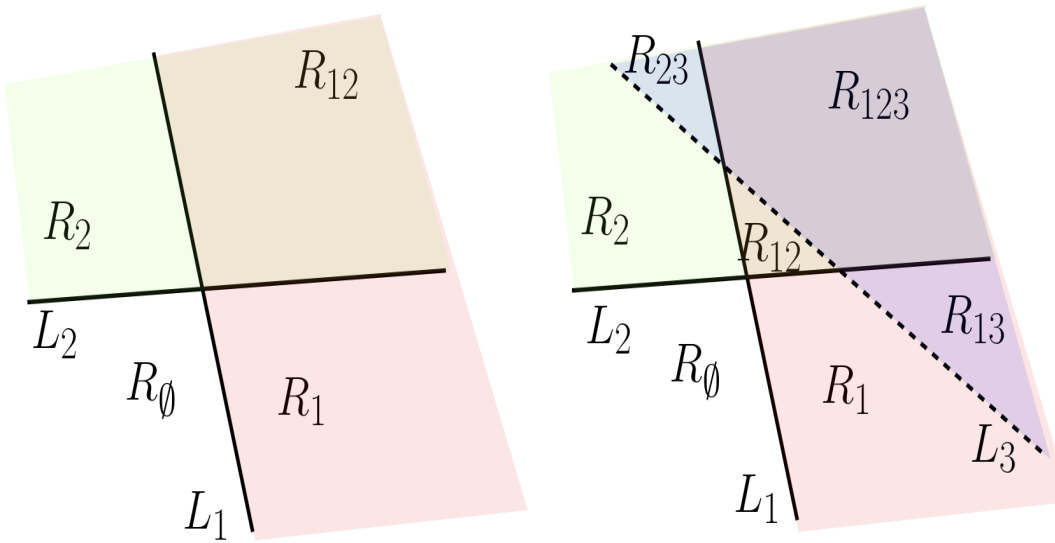


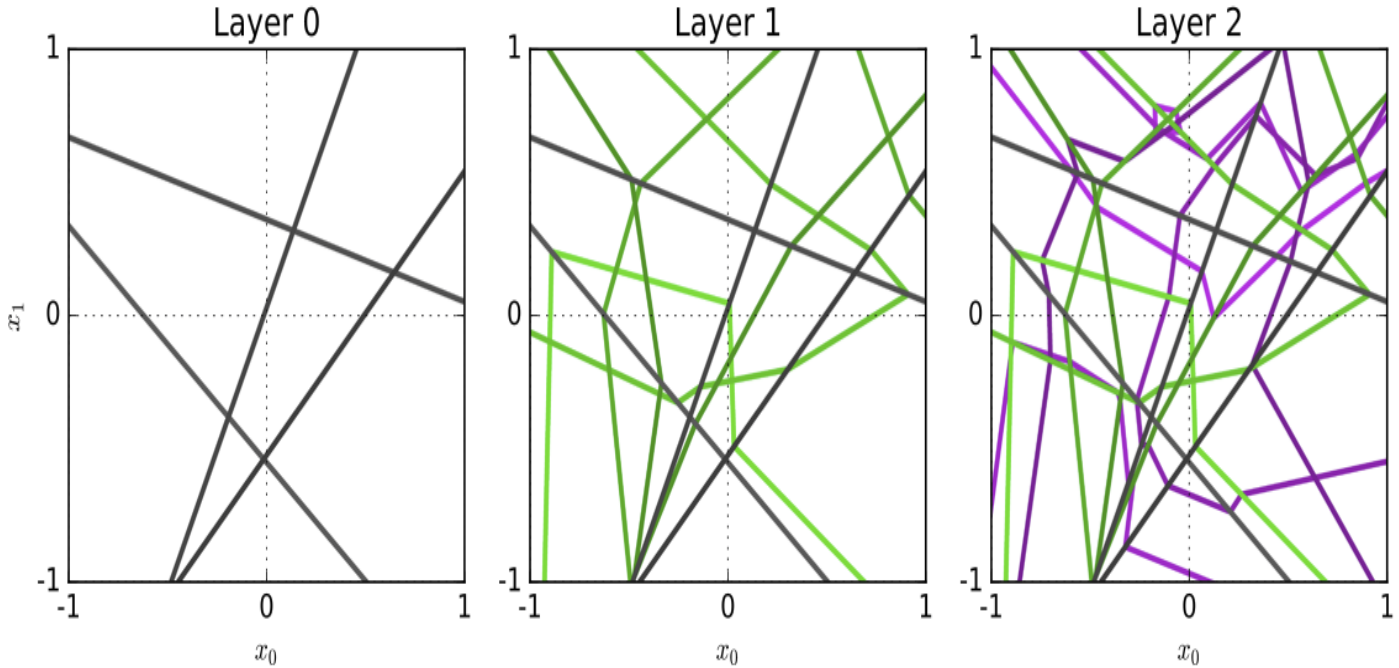
Figure 2: Induction step of the hyperplane sweep method for counting the regions of line arrangements in the plane.

`https://arxiv.org/pdf/1312.6098.pdf`

ReLU hyperplane arrangement

Partition of the input domain $\mathbb{R}^{n_0}$: plot Raghu et al. 16'

Mathematical Institute



https://arxiv.org/abs/1606.05336

This "activation region" perspective is a useful intuition for ReLU, but lacks the quantitative convergence rates we observed in more recent approximation theory results of Yarotsky 16'.

UNIVERSITY OF OXFORD
Mathematical Institute

A random network $f_{NN}(x; \mathcal{P}, \mathcal{Q})$ denotes a deep neural network:

$$h^{(d)} = W^{(d)} z^{(d)} + b^{(d)}, \qquad z^{(d+1)} = \phi(h^{d}), \qquad d = 0, \ldots, L-1,$$

which takes as input the vector x, and is parameterised by random weight matrices $W^{(d)}$ with entries sampled iid from the distribution $\mathcal{P}$, and bias vectors $b^{(d)}$ with entries drawn iid from distribution $\mathcal{Q}$.

While our goal is always to train a network, DNNs typically start as random networks which influence their ability to be trained.

Popular choices are Gaussian, $\mathcal{P} = \mathcal{N}(0, \sigma_w^2)$, or uniform, $\mathcal{P} = \mathcal{U}(-C_w, C_w)$ initialisations.
(*Note, for random networks we use $\phi(\cdot)$ as the nonlinear activation and σ to denote variance.)

Raghu et al. 16' introduced the notion of trajectory length

$$l(x(t)) = \int_t \left|\left| \frac{dx(t)}{dt} \right|\right| dt.$$

as a measure of expressivity of a DNN. In particular, they considered passing a simple geometric object $x(t)$, such as a line $x(t) = tx_0 + (1-t)x_1$ for $x_0, x_1 \in \mathbb{R}^k$ and measure the expected length of the output of the random DNN at layer d:

$$\frac{\mathcal{E}\left[\ell(z^{(d)})\right]}{\ell(x(t))}$$

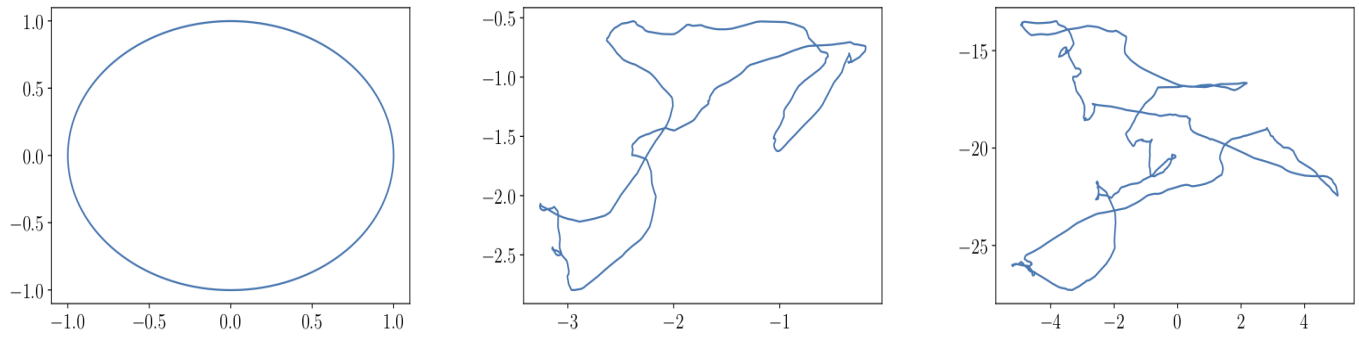
`https://arxiv.org/abs/1606.05336`

Example of circle passed through a random DNN

Complexity of output increasing with depth

UNIVERSITY OF OXFORD
Mathematical Institute

A circle passed through a random DNN and the pre-activation output $h^{(d)}$ at layers $d = 6$ and 12.



(a) Input (b) Layer 6 (c) Layer 12

Figure 1: A circular trajectory, passed through a ReLU network with $\sigma_w = 2$. The plots show the pre-activation trajectory at different layers projected down onto 2 dimensions.

DNNs can be used to *generative* data, GANs, and there one might consider the complexity of the manifold the GAN can generate as a measure of expressivity.

Random DNN: expected path length lower bound

Path length bound dependence on σ_w.

Mathematical Institute

Consider random DNNs of width n and depth L with weights and bias are drawn i.i.d. $W^{(\ell)}(i,j) \sim \mathcal{N}(0, \sigma_w^2/n)$, $b^{(\ell)}(j) \sim \mathcal{N}(0, \sigma_b^2)$

Theorem (Raghu et al. 16')

Consider as input a one dimensional trajectory $x(t)$ with arc-length $\ell(x(t)) = \int_t \left\| \frac{dx(t)}{dt} \right\| dt$ and let $z^{(L)}(t)$ be the output of the Gaussian random feedforward network with ReLu activations, then

$$\frac{\mathcal{E}\left[\ell(z^{(L)})\right]}{\ell(x(t))} \geq \mathcal{O}\left(\left(\frac{\sigma_w}{(\sigma_w^2+\sigma_b^2)^{1/4}} \cdot \frac{n^{1/2}}{(n+(\sigma_w^2+\sigma_b^2)^{1/2})^{1/2}}\right)^L\right).$$

https://arxiv.org/abs/1606.05336

Exponential growth of path length with depth.

Empirical experiments for htanh activation

Mathematical Institute

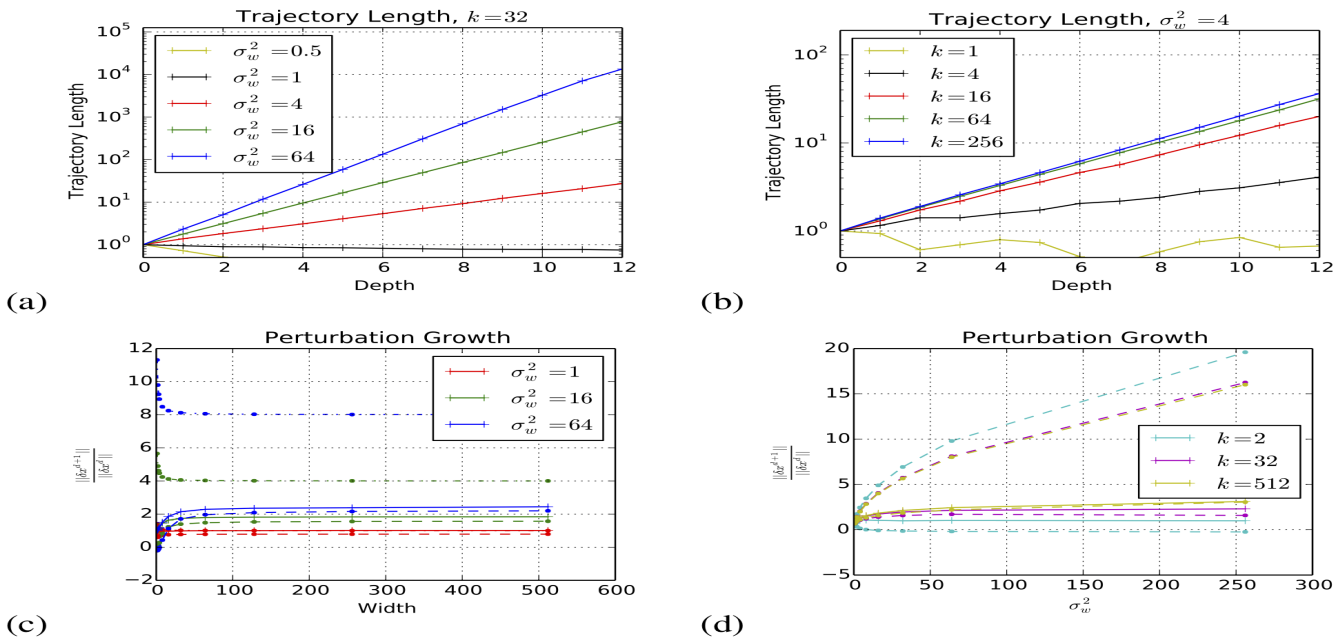


Figure 2: The exponential growth of trajectory length with depth, in a random deep network with hard-tanh nonlinearities. A circular trajectory is chosen between two random vectors. The image of that trajectory is taken at each layer of the network, and its length measured. *(a,b)* The trajectory length 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 1). Growth rate is a function of layer width k, and weight variance σ_w^2.

https://arxiv.org/pdf/1611.08083.pdf

Mathematical Institute

Theorem (Price et al. 19')

Let $f_{NN}(x; \alpha, \mathcal{P}, \mathcal{Q})$ be a random sparse net with layers of width n. Then, if $\mathbb{E}[|u^T w_i|] \geq M\|u\|$, where w_i is the i^{th} row of $W \in \mathcal{P}$, and u and M are constants, then

$$\mathbb{E}[l(z^{(L)}(t))] \geq \left(\frac{M}{2}\right)^L \cdot l(x(t))$$

for $x(t)$ a 1-dimensional trajectory in input space.

Exponential growth with depth for random initialisations such as Gaussian, uniform, and discrete; e.g. for Gaussian $M = \sigma_w\sqrt{2/\pi}$.
`https://arxiv.org/abs/1911.10651`

Observed growth rate (solid) and bounds (dashed)

Empirical experiments showing dependence on σ_w and sparsity α

University of Oxford
Mathematical Institute

Price et al. 19' also extended the results to have all but α fraction of the entries in W equal to 0.

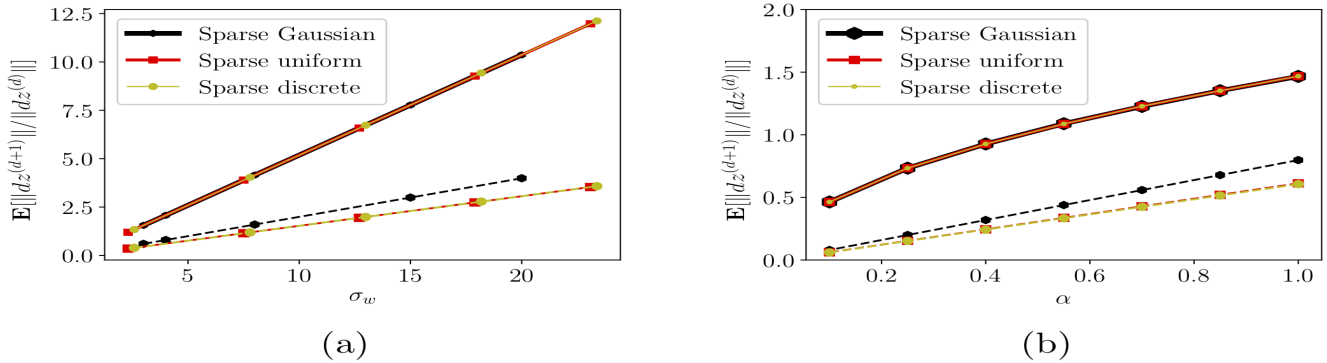


Figure 3: Expected growth factor, that is, the expected ratio of the length of any very small line segment in layer $d+1$ to its length in layer d. Figure 3a shows the dependence on the variance of the weights' distribution, and Figure 3b shows the dependence on sparsity.

Unless σ_w or α small enough at initialisation the pre-activiation output is exponentially complex.

`https://arxiv.org/abs/1911.10651`