Trajectory growth lower bounds for random sparse deep ReLU networks

Ilan Price & Jared Tanner
University of Oxford

Trajectory growth in deep networks

We lower-bound the growth in length of a 1D trajectory as it is passed through a random ReLU network, in expectation.

The bounds are derived in terms of network hyperparamters: width k, depth d, connection density α, and parameters of the non-zero-weights distributions.

- This is one measure of neural network expressivity (see Raghu et al. (2017) whose work we extend.)
- Mathematically, the properties of the manifolds generated by random neural networks are of interest in their own right.

Definitions

A random deep neural network is one in which the weights and biases are random variables.

A trajectory $x(t)$ is a 1-dimensional path in $\mathbb{R}^n$ from x_0 to x_1 which is parameterised by $t \in [0, 1]$, with $x_0 = x(0)$ and $x_1 = x(1)$.

A sparse deep neural network is one whose weight matrices are sparse.

Main theorem (informal):

Let f_{NN} be a random sparse network with layers of width k with sparsity $(1 - \alpha)$. If the the non-zero weights and bias distributions are 'jointly even', and $\mathbb{E}[|\boldsymbol{u}^\top w|] \geq M\|\boldsymbol{u}\|$ for any row of a weight matrix w and any constant vector $\boldsymbol{u}$, then for $x(t)$ a 1-dimensional trajectory in input space,

$$\mathbb{E}[\text{length at layer d}] \geq \left(\frac{\alpha M\sqrt{k}}{2}\right)^d \cdot \text{length of } x(t). \quad (1)$$

The base of the exponent is the growth factor from layer to layer, which we can denote G.

Trajectory growth in sparsely-connected, deep ReLU networks is exponential in depth, for many parameter distributions.

Contributions of this work:

- An alternative, simpler, and **more general method** for lower bounding expected trajectory growth.
- Bounds for **a broad class of parameter distributions**, including sparse-Gaussian, sparse-uniform, and sparse-discrete nets - all of which are exponential in depth.
- Experiments showing that expected length growth factor is strikingly similar across these three distributions, suggesting a **universal dependence** of the expected growth in length **on the weight variance and sparsity**.

Take a picture to download the full paper

The Alan Turing Institute

UNIVERSITY OF OXFORD

Results for common distributions:

We derive corollaries for sparse networks with the following non-zero-parameter distributions:

- Gaussian: $w_{ij} \sim \mathcal{N}(0, \sigma_w^2)$
- Uniform: $w_{ij} \sim \mathcal{U}(-C_w, C_w)$
- Discrete: w_{ij} from finite, discrete, symmetric set $\mathcal{W}$

Different distribution $\implies$ **different** M **in Main Theorem**
$\implies$ **different** G.

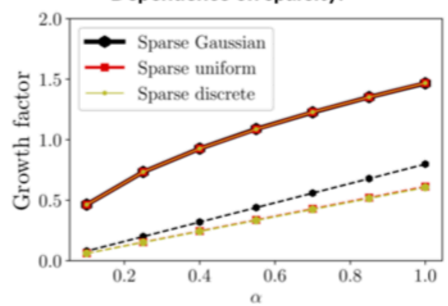
The growth factors in the resulting lower bounds exhibit the following relationships with network hyperparameters:

- Sparse-Gaussian nets: $G \propto \alpha\sigma_w\sqrt{k}$
- Sparse-uniform nets: $G \propto \alpha C_w\sqrt{k}$
- Sparse-discrete nets: $G \propto \alpha\sqrt{k} \cdot \frac{\sum_{w \in \mathcal{W}} |w|}{N_w}$

In practice: a universal dependence on σ_w and α?

Experimental setup: A line connecting MNIST datapoints is passed through untrained feedforward nets, with variance scaled by network width. Results are averaged over many sampled nets.
Dotted lines show the lower bounds.
Solid lines show the observed results, **which overlap exactly across distributions.**

Dependence on sparsity:



Dependence on weight scale: