Mathematical
Institute

Random DNN: hidden layer output evolution with depth

Theories of Deep Learning: C6.5, Video 6

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford
Mathematics

Random DNNs hidden layer outputs

Exponential geometric complexity

Mathematical Institute

Random DNNs can generate exponentially complex geometric objects: (Raghu et al. 16') and (Poole et al. 16')

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

`https://arxiv.org/pdf/1606.05340.pdf`

respectively derive lower bounds on *length* and *curvature* of simple line inputs. These lower bounds are exponential in depth.

Theorem (Price et al. 19')

Let $f_{NN}(x; \mathcal{P}, \mathcal{Q})$ be a random net with $\mathcal{E}[|u^T w_i|] \geq M\|u\|$, for w_i is the i^{th} row of $W \in \mathcal{P}$, and u and M are constants, then

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

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

UNIVERSITY OF OXFORD
Mathematical Institute

Let $f_{NN}(x)$ denote a random Gaussian DNN

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

which takes as input the vector x, and is parameterised by random weight matrices $W^{(\ell)}$ and bias vectors $b^{(\ell)}$ with entries sampled iid from the Gaussian normal distributions $\mathcal{N}(0, \sigma_w^2)$ and $\mathcal{N}(0, \sigma_b^2)$. Define the ℓ^2 length of the pre-activation hidden layer $h^{(\ell)} \in \mathbb{R}^{n_\ell}$ output as:

$$q^\ell = n_\ell^{-1} \left\| h^{(\ell)} \right\|_{\ell^2}^2 := \frac{1}{n_\ell} \sum_{i=1}^{\ell} \left(h^{(\ell)}(i) \right)^2 .$$

which is the average value of the random entries $\left(h^{(\ell)}(i) \right)^2$.

Random DNN recursion map (Poole et al. 16')

Norm of hidden layer output dependence on prior layer

UNIVERSITY OF OXFORD
Mathematical Institute

The norm of the hidden layer output $q^\ell = n_\ell^{-1} \left\| h^{(\ell)} \right\|_{\ell^2}^2$ has an expected value over the random draws of $W^{(\ell)}$ and $b^{(\ell)}$ which satisfies

$$\mathcal{E}(q^\ell) = \mathcal{E}\left(\left(h^{(\ell)}(i) \right)^2 \right)$$

which as $h^{(\ell)} = W^{(\ell)} \phi(h^{(\ell-1)}) + b^{(\ell)}$ is

$$\mathcal{E}(q^\ell) = \mathcal{E}\left(\left(w_i^{(\ell)} \phi\left(h^{(\ell-1)} \right) \right)^2 \right) + \mathcal{E}\left(\left(b_i^{(\ell)} \right)^2 \right)$$

$$= \sigma_w^2 n_{\ell-1}^{-1} \sum_{i=1}^{n_{\ell-1}} \phi\left(h_i^{(\ell-1)} \right)^2 + \sigma_b^2.$$

`https://arxiv.org/pdf/1606.05340.pdf`

Random DNN recursion map (Poole et al. 16')

Mean field recursion between layers

Mathematical Institute

Approximating $h_i^{(\ell-1)}$ as Gaussian with expected variance $q^{(\ell-1)}$:

$$n_{\ell-1}^{-1} \sum_{i=1}^{n_{\ell-1}} \phi\left(h_i^{(\ell-1)}\right)^2 = \mathcal{E}\left(\phi\left(q^{(\ell-1)}\right)^2\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi\left(\sqrt{q^{(\ell-1)}} z\right)^2 e^{-z^2/2} dz$$

which gives a recursive map of $q^\ell = n_\ell^{-1} \left\| h^{(\ell)} \right\|_{\ell^2}^2$ between layers

$$q^{(\ell)} = \sigma_b^2 + \sigma_w^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi\left(\sqrt{q^{(\ell-1)}} z\right)^2 e^{-z^2/2} dz =: \mathcal{V}(q^{(\ell-1)} | \sigma_w, \sigma_b, \phi(\cdot)).$$

Note that the integral is larger for $\phi(x) = |x|$ than ReLU, which are larger than $\phi(x) = tanh(x)$, indicating smaller σ_w, σ_b needed.

`https://arxiv.org/pdf/1606.05340.pdf`

Example of DNN recursion fixed points (Poole et al. 16')

Dependence on σ_w, σ_b, $\phi(\cdot)$ for $\phi(\cdot) = tanh(\cdot)$.

Mathematical Institute

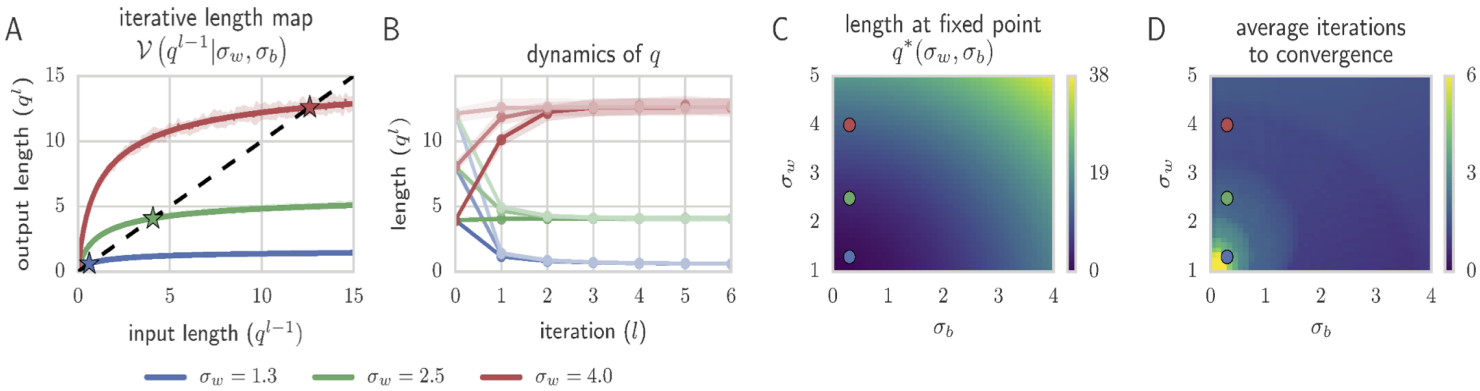

Figure 1: Dynamics of the squared length q^l for a sigmoidal network ($\phi(h) = \tanh(h)$) with 1000 hidden units. (A) The iterative length map in (3) for 3 different σ_w at $\sigma_b = 0.3$. Theoretical predictions (solid lines) match well with individual network simulations (dots). Stars reflect fixed points q^* of the map. (B) The iterative dynamics of the length map yields rapid convergence of q^l to its fixed point q^* , independent of initial condition (lines=theory; dots=simulation). (C) q^* as a function of σ_w and σ_b. (D) Number of iterations required to achieve $\leq 1\%$ fractional deviation off the fixed point. The (σ_b, σ_w) pairs in (A,B) are marked with color matched circles in (C,D).

Note that the fixed points here are all stable.

`https://arxiv.org/pdf/1606.05340.pdf`

A single input x has hidden pre-activation output converging to a fixed expected length.
Consider the map governing the angle between the hidden layer pre-activations of two distinct inputs $x^{(0,a)}$ and $x^{(0,b)}$:

$$q_{ab}^{(\ell)} = n_\ell^{-1} \sum_{i=1}^{n_l} h_i^{(\ell)}(x^{(0,a)}) h_i^{(\ell)}(x^{(0,b)}).$$

Similar to the analysis before, use the relation $h^{(\ell)} = W^{(\ell)}\phi(h^{(\ell-1)}) + b^{(\ell)}$ to show the relation between layers.
`https://arxiv.org/pdf/1606.05340.pdf`

Replacing the average in the sum with the expected value gives

$$q_{ab}^{(\ell)} = n_\ell^{-1} \sum_{i=1}^{n_l} h_i^{(\ell)}(x^{(0,a)}) h_i^{(\ell)}(x^{(0,b)})$$

$$= \sigma_b^2 + \sigma_w^2 \mathcal{E}\left(\phi(h^{(\ell-1)}(x^{(0,a)}))\phi(h^{(\ell-1)}(x^{(0,b)}))\right)$$

where as before, $h_i^{(\ell-1)}$ are well modelled as being Gaussian with expected length $q^{(\ell-1)}$ which converge to fixed points q^*.
`https://arxiv.org/pdf/1606.05340.pdf`

Random DNN correlations between inputs (Poole et al. 16')

Recursion correlation map.

Mathematical Institute

Defining the angle between the hidden layers as $c^{(\ell)} = q_{12}^{(\ell)}/q^{(*)}$ and writing the expectation as integrals we have

$$c^{(\ell)} = \mathcal{C}\left(c^{(\ell-1)}|\sigma_w, \sigma_b, \phi(\cdot)\right) := \sigma_b^2 + \sigma_w^2 \int Dz_1 Dz_2 \phi(u_1)\phi(u_2)$$

where the double integral is with respect to the measure $Dz = (2\pi)^{-1/2} e^{-z^2/2} dz$ where $u_1 = \sqrt{q^*} z_1$ and $u_2 = \sqrt{q^*}[c^{(\ell-1)} z_1 + \sqrt{1-(c^{(\ell-1)})^2} z_2]$ are a change of variables for the integrals.

`https://arxiv.org/pdf/1606.05340.pdf`

Random DNN correlations between inputs (Poole et al. 16')

University of Oxford
Mathematical Institute

The map has a fixed point at $c = 1$. For $\sigma_b^2 > 0$ the map starts at a positive value, even if $x^{(0,a)}$ and $x^{(0,b)}$ are orthogonal.
Of particular note is the slope of $\mathcal{C}$ at $c = 1$

$$\chi := \frac{\partial c^{(\ell)}}{\partial c^{(\ell-1)}}|_{c=1} = \sigma_w^2 \int Dz[\phi'(\sqrt{q^*}z)^2].$$

Stability of the fixed point at $c = 1$ is determined by χ:

- $\chi < 1$: $c = 1$is locally stable and points which are sufficiently correlated all converge, with depth, to the same point.
- $\chi > 1$: $c = 1$ is unstable and nearby points become uncorrelated with depth.
- Preferable to choose $\chi = 1$ if possible for $\sigma_w, \sigma_b, \phi(\cdot)$.

`https://arxiv.org/pdf/1606.05340.pdf`

Example of DNN correlation fixed points (Poole et al. 16')

Dependence on σ_w, σ_b, $\phi(\cdot)$ for $\phi(\cdot) = tanh(\cdot)$.

Mathematical Institute

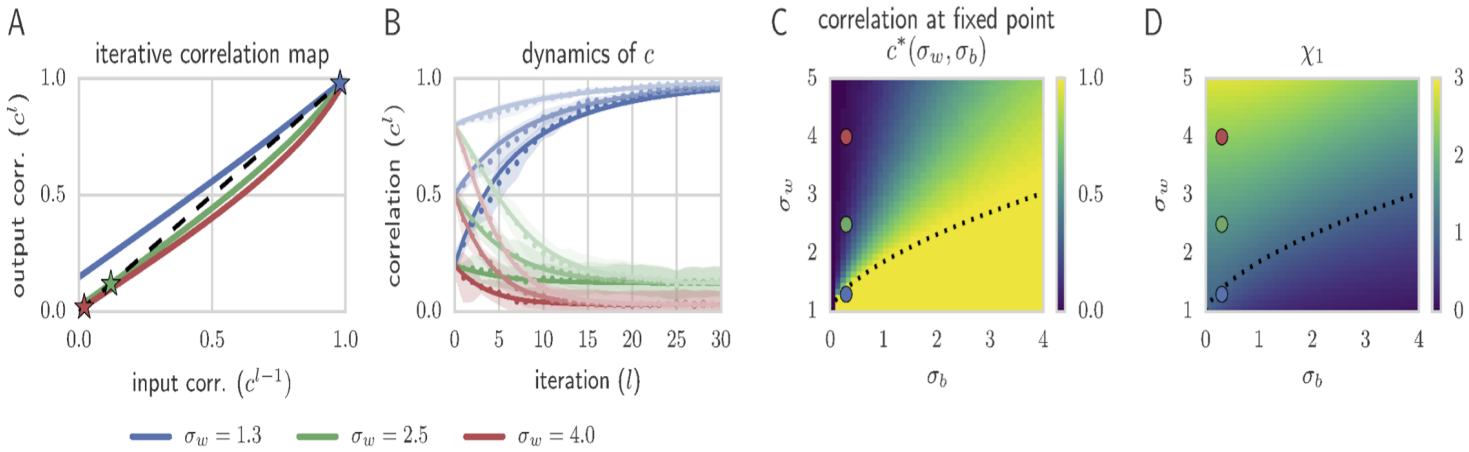

Figure 2: Dynamics of correlations, c_{12}^l, in a sigmoidal network with $\phi(h) = \tanh(h)$. (A) The $\mathcal{C}$-map in (6) for the same σ_w and $\sigma_b = 0.3$ as in Fig. 1A. (B) The $\mathcal{C}$-map dynamics, derived from both theory, through (6) (solid lines) and numerical simulations of (1) with $N_l = 1000$ (dots) (C) Fixed points c^* of the $\mathcal{C}$-map. (D) The slope of the $\mathcal{C}$-map at 1, χ_1, partitions the space (black dotted line at $\chi_1 = 1$) into chaotic ($\chi_1 > 1, c^* < 1$) and ordered ($\chi_1 < 1, c^* = 1$) regions.

Note three respective stable fixed points, determined in part by χ.

`https://arxiv.org/pdf/1606.05340.pdf`

DNN random initialisation summary

Dependence on $\sigma_w, \sigma_b, \phi(\cdot)$.

Mathematical Institute

- The hidden layers converge to fixed expected length.
- All inputs converge to either one another or to be orthogonal, independent of the class the data is in, typically happening at a rate which is exponential with depth.
- The rate with which these phenomenon occur, and values which they take, are determined by the choice of $\sigma_w, \sigma_b, \phi(\cdot)$.
- Very DNNs can be especially hard to train for activations with unfavourable initialisations; e.g. ReLU with $\chi = 1$ requires $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$.