

Title: Lecture 1: Introduction: Criticality in Linear Networks

Speakers: Yonatan Kahn

Collection/Series: Special Seminars

Subject: Other

Date: October 08, 2025 - 3:00 PM

URL: <https://pirsa.org/25100143>

(Main reference; Roberts, Yaida, Hanin 2106.10165)

I. Criticality in linear networks

Neural network: flexible, differentiable class
of functions w/ repeating structure

data \rightarrow affine trans. \rightarrow nonlinearity \rightarrow output
repeat k times

repeat # times

$$z_i^{(1)} = b_i^{(1)} + w_{ij}^{(1)} x_j$$

↑ pre-activation $\in \mathbb{R}^n$ ↑ bias $\in \mathbb{R}^n$ ↑ weights $\in \mathbb{R}^{1 \times 1}$ ↑ data $\in \mathbb{R}^n$

$$z_i^{(l+1)} = b_i^{(l+1)} + w_{ij}^{(l+1)} \sigma(z_j^{(l)})$$

← activation function



width n_l

(compare to some target $y(x)$)

Output $f(x; \theta)$

$$\theta = \{b^{(l)}, w^{(l)}\}$$

$l=1, \dots, l_{\max}$

Typically massively overparameterized
 Dataset could have 50k elements
 width 256, depth 5 \Rightarrow 300k params.

width 256, depth 5 \Rightarrow 300K params.

Physics analogies

1. Initial choice of params θ is random

\Rightarrow NNs should be seen as elements of an ensemble
Simplifies as # params $\rightarrow \infty$

2. Flow of "information" from input \rightarrow output
is like RG flow from UV \rightarrow IR

3. There is an object called the NTK which
acts like a Hamiltonian
 \Rightarrow governs updates of θ for each step
of gradient descent
lets us connect statistics at initialization
to statistics at end of training

NN ensembles

3 sources of stochasticity:

1. Init.

~~2. Data~~ ← fixed \mathcal{D}

~~3. Training~~

$$p(f(x; \theta) | \mathcal{D})$$

$$f(x) \equiv z^{(L)} \quad (\text{layers } 1, 2, \dots, L)$$

$$p(f) = p(z^{(L)} | \mathcal{D}) = \int \prod_{n=1}^L d\theta_n p(\theta) p(z^{(L)} | \theta, \mathcal{D})$$

Look at $p(z^{(k+1)} | \mathcal{D}) = \int \prod dz_i^{(k)} p(z^{(k+1)} | z^{(k)}) p(z^{(k)} | \mathcal{D})$

Initial condition $p(z^{(1)} | \mathcal{D}) = \int \prod db^{(1)} dw^{(1)} p(b^{(1)}) p(w^{(1)})$
 $\delta(z_i^{(1)} - b_i^{(1)} - w_{i,j}^{(1)} x_j)$

$$O(z_i - b_i - w_{ij} x_j)$$

For "good" choices of $p(b)$, $p(w)$, all of these distributions (and hence $p(z)$) are perturbatively Gaussian (non-Gaussianity $\propto \frac{1}{\epsilon}$)

Toy model: linear network w/ zero biases

Take $b_i^{(0)} = 0$, $\sigma(z) = z$

$$z^{(l+1)} = W^{(l+1)} z^{(l)}$$

$$z^{(L)} = \prod_{l=1}^L W^{(l)} x$$

Anticipating the near Gaussianity, instead of computing $p(z^{(l)} | \theta)$, compute its first few moments

Take $p(w)$ to be Gaussian: $E[W_{ij}] = 0$

$$E[W_{i_1 j_1}^{(l)} W_{i_2 j_2}^{(l)}] = \frac{C_w}{\lambda} \delta_{i_1 i_2} \delta_{j_1 j_2}$$

Let $\mathcal{D} = \{x_{i\alpha}\}$ α labels elements of \mathcal{D}

$$p(z^{(L)} | \mathcal{D}) = p(z_{i\alpha_1}^{(L)}, z_{i\alpha_2}^{(L)}, \dots, z_{i\alpha_{N_{\mathcal{D}}}}^{(L)})$$

First, odd moments vanish

$$E[z_{\alpha}^{(L)}] = E[\cancel{w^{(L)} w^{(L-1)}} \cdot \cancel{w^{(L)}} x_{\alpha}] = 0$$

~~3. Training~~

$$E[z_{i_1 \alpha_1}^{(1)} z_{i_2 \alpha_2}^{(1)}] = E[W_{i_1 j_1}^{(1)} x_{i_1 \alpha_1} W_{i_2 j_2}^{(1)} x_{i_2 \alpha_2}]$$

$$i_1, i_2 = 1, \dots, n$$

$$\alpha_1, \alpha_2 = 1, \dots, N_\theta$$

$$= E[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}] x_{i_1 \alpha_1} x_{i_2 \alpha_2}$$

$$= \frac{C_w}{n} \delta_{i_1 i_2} \delta_{j_1 j_2} x_{i_1 \alpha_1} x_{i_2 \alpha_2}$$

$$= C_w \left(\frac{1}{n} \vec{x}_{\alpha_1} \vec{x}_{\alpha_2} \right) \delta_{i_1 i_2}$$

$$E[z_{i_1 \alpha_1}^{(l+1)} z_{i_2 \alpha_2}^{(l+1)}] = E[w_{i_1 j_1}^{(l+1)} z_{j_1 \alpha_1}^{(l)} w_{i_2 j_2}^{(l+1)} z_{j_2 \alpha_2}^{(l)}]$$

$l+1$ uncorrelated w/l $\Rightarrow E[w_{i_1 j_1}^{(l+1)} w_{i_2 j_2}^{(l+1)}] E[z_{j_1 \alpha_1}^{(l)} z_{j_2 \alpha_2}^{(l)}]$

$$= \frac{C_w}{n} \delta_{i_1 j_1} \delta_{i_2 j_2} C_{\alpha_1 \alpha_2} \delta_{j_1 j_2} +$$

Recursion. $G_{\alpha_1, \alpha_2}^{(l)} = c_w^l G_{\alpha_1, \alpha_2}^{(0)}$

If $c_w \neq 1$, 2-point function blows up/shrinks exponentially w. th depth

Take $c_w = 1 \Rightarrow G_{\alpha_1, \alpha_2}^* \equiv G_{\alpha_1, \alpha_2}^{(0)}$

Recursion $G_{\alpha_1, \alpha_2}^{(l)} = C_w^l G_{\alpha_1, \alpha_2}^{(0)}$

If $C_w \neq 1$, 2-point function blows up/shrinks exponentially w. the depth

Take $C_w = 1 \Rightarrow G_{\alpha_1, \alpha_2}^* \equiv G_{\alpha_1, \alpha_2}^{(0)}$
(turns out this covers all exponential behavior)

4-point recursion:

(for now, set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 \equiv \alpha$)

$$\begin{aligned}
 E[z_{i_1}^{(1)} z_{i_2}^{(1)} z_{i_3}^{(1)} z_{i_4}^{(1)}] &= E[w_{i_1 i_1}^{(1)} w_{i_2 i_2}^{(1)} w_{i_3 i_3}^{(1)} w_{i_4 i_4}^{(1)}] x_{i_1} x_{i_2} x_{i_3} x_{i_4} \\
 &= \frac{C_w^2}{n^2} (\delta_{i_1 i_2} \delta_{i_3 i_4} + (2 \text{ s.i.m.})) \\
 &= C_w^2 \left(\frac{1}{n} + \frac{1}{n} + \frac{1}{n} + \frac{1}{n} \right) (G_2^{(0)})^2
 \end{aligned}$$

$$E[z_{i1}^{(l)} z_{i2}^{(l)} z_{i3}^{(l)} z_{i4}^{(l)}]_{\text{conn.}} = 0$$

Non-Gaussianity in deeper layers

$$E[z_{i1}^{(l)} z_{i2}^{(l)} z_{i3}^{(l)} z_{i4}^{(l)}] = G_4^{(l)} (\sigma_{i1i2} \sigma_{i3i4} + (2 \text{ sim.}))$$

$$\Rightarrow G_4^{(l+1)} = C_w^2 \left(1 + \frac{2}{n}\right) G_4^{(l)}$$

Expand in $\frac{1}{n}$:

$$E[z^{q(l)}]_{\text{conn.}} \propto G_{\alpha}^{(l)} - (G_{\gamma}^{(l)})^2 = \frac{2(l-1)}{n} (G_{\gamma}^{(l)})^2$$

At criticality ($n=1$), $E[z^{q(l)}]_{\text{conn.}} \propto \frac{2l}{n} (G_{\gamma}^{(l)})^2$

$$E[z^{2n}] \propto \frac{l^{n-1}}{n^{n-1}}$$

(marginally relevant)

$$E[z^{2n}] \propto \frac{L^{n-1}}{n-1}$$

(marginally relevant)

Comments:

• $n \rightarrow \infty$, 4th and higher cumulants vanish

$p(z^{(4)} | \infty)$ is Gaussian in infinite-width limit

• Cutoff scale is $\frac{L}{n}$ (not for other limits!)

If instead $W \sim \text{Haar}(O(n))$

$$\Rightarrow p(z^{(k)} | x) = \delta(\vec{z}^{(k)} \cdot \vec{z}^{(k)} - \vec{x} \cdot \vec{x})$$

$$\Rightarrow \mathbb{E}[z^{(k)}]_{\text{corr}} = -\frac{2}{n} (G_v^{(k)})^2$$

independent of k !

If instead $W \sim \text{Haar}(O(n))$

$$\Rightarrow p(z^{(l)} | x) = \delta(\bar{z}^{(l)} \cdot \bar{z}^{(l)} - \bar{x} \cdot \bar{x})$$

$$\Rightarrow \mathbb{E}[z^{4(l)}]_{\text{conn}} = -\frac{2}{n} (G_V^{(0)})^2$$

independent of l

Pehlevan et al: Gaussian $p(w)$

$$p(z^{(l)} | x) \propto G_{0,l}^{l,0}(\dots)$$

exactly marginal

Take $w=1 \Rightarrow G_{d,d}^* \equiv G_{d,d}^{(0)}$
(turns out this covers all exponential behavior)

$$p(z^{(0)}|x) \propto G_{0,l}^{l,0}(\dots)$$

exactly marginal

performance

$C_w < 1$



training time

- total

at

inc