Lecture 1: Introduction, criticality in linear networks

Based on Roberts and Yaida, Principles of Deep Learning Theory, 2106.10165

Why study neural networks? Flexible, differentiable class of functions with which to perform tasks like regression.

$$\text{Data} \to \underbrace{\text{affine transformation} \to \text{nonlinearity}}_{\text{repeat many times}} \to \text{output}$$

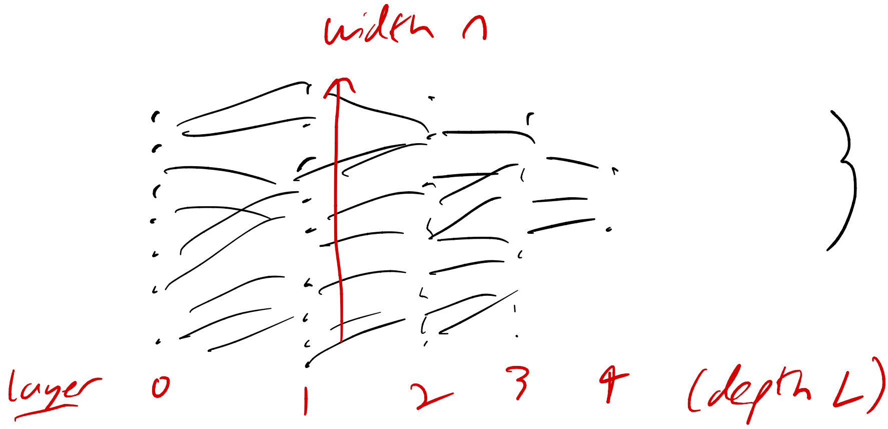
In equations: $z_i^{(1)} = b_i^{(1)} + W_{ij}^{(1)} x_j$

$\uparrow$ biases $\quad \uparrow$ weights $\quad \uparrow$ data

preactivation

$$z_i^{(\ell+1)} = b_i^{(\ell+1)} + W_{ij}^{(\ell+1)} \sigma(z_j^{(\ell)})$$

$\uparrow$ (nonlinear) activation function, applied element-wise



output is a function $f(x;\theta)$ parameterized by $\theta = \{b^{(\ell)}, W^{(\ell)}\}$

In supervised learning, neural networks are trained by comparing the NN output to a desired output and updating the parameters by some form of gradient descent.

Key fact: these networks are massively overparameterized.

A standard benchmark dataset has 50k elements.

A standard "architecture" has width 256, depth 5 => 300k params.

How can you possibly avoid "overfitting" your data with so many parameters? This is one of the alluring mysteries of neural networks: somehow, they work!

3 physics analogies to keep in mind:

1. the initial choice of parameters is random, so NN's should be seen as elements of an ensemble. The statistics of this ensemble simplify as # of params → ∞, just like in stat mech.
2. the flow of information from input to output is like RG flow from UV to IR. The NN equations are recursions, so look for fixed points => criticality (lack of exponential behavior of correlation functions) (lectures 1-2)
3. there is an object called the NTK which acts like a Hamiltonian, governing updates of observables after one step of training. We can connect statistics at initialization to statistics at end of training if we can find a derivative of the NTK which is frozen during training (lecture 3)

NN ensemble

In principle, 3 sources of randomness: initialization, data (drawn from some data distribution), training (e.g. stochastic gradient descent).

To simplify the analysis, we consider only initialization as random. Goal is to compute $p(f(x;\theta^*)|\mathcal{D})$, the distribution over trained networks where θ^* are the optimal parameters given training data $\mathcal{D}$.

We'll warm up by first computing $p(f(x;\theta)|\mathcal{D})$ before any training.

Formally, since the output is given by $z^{(L)}$ for a network of depth L, we have

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

↑ deterministic, given by iteration eqn.

But because NN's have an iterative layer-to-layer structure, will be easiest to marginalize over layers one at a time:

$$p(z^{(\ell+1)}|\mathcal{D}) = \int \prod dz_i^{(\ell)} \, p(z^{(\ell+1)}|z^{(\ell)}) \, p(z^{(\ell)}|\mathcal{D}),$$

which is a recursion with initial condition

$$p(z^{(1)}|\mathcal{D}) = \int \prod db_i^{(1)} \, dW_{ij}^{(1)} \, p(b^{(1)}) \, p(W^{(1)}) \, \delta(z_i^{(1)} - b_i^{(1)} - W_{ij}^{(1)} x_j)$$

For simple $p(b)$, $p(W)$ (i.e. Gaussian distributions) it turns out the marginal distributions, and hence $p(z)$, are perturbatively Gaussian, with non-Gaussianities scaling as $\frac{1}{n}$. Can therefore borrow all of the tools from stat. field theory to compute expectations!

Toy model: linear network

Take $b_i^{(\ell)} = 0$, $\sigma(z) = z$. This is literally successive multiplication by a random matrix. Output is always a linear function of input, but a highly nonlinear function of parameters (matrix entries), so there is still rich structure at initialization that will carry over to the nonlinear case.

Anticipating the nearly-Gaussian statistics, instead of computing $p(z|\mathcal{D})$ directly, we will compute its first few connected moments.

Take our weight matrix entries to be i.i.d. Gaussian;

$$\mathbb{E}[W_{ij}] = 0, \quad \mathbb{E}[W_{i_1j_1} W_{i_2j_2}] = \frac{C_W}{n}\delta_{i_1i_2}\delta_{j_1j_2}$$

(for simplicity, same width n and same weight distribution in each layer)

Let $\mathcal{D} = \{x_{i;\alpha}\}$ where α is a sample index. Then

$$p(z^{(\ell)}|\mathcal{D}) = p(z^{(\ell)}_{i;\alpha_1}, z^{(\ell)}_{i;\alpha_2}, \dots z^{(\ell)}_{i;\alpha_{N_\mathcal{D}}}) \quad \text{w} \quad z_\alpha \equiv z(x_\alpha)$$

Easy things first: all odd moments vanish

e.g. $\mathbb{E}[z_\alpha^{(\ell)}] = \mathbb{E}[W^{(\ell)} W^{(\ell-1)} \cdots W^{(1)} x_\alpha]$ (dropping indices for clarity)

$= \mathbb{E}[W^{(\ell)}]\mathbb{E}[W^{(\ell-1)}]\cdots\mathbb{E}[W^{(1)}]\, x_\alpha$ (each layer's weights are independent)

$= 0$ (weights are mean-zero)

First nontrivial moment is

$$\mathbb{E}[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_2;\alpha_2}] = \mathbb{E}[W^{(1)}_{i_1j_1} x_{j_1;\alpha_1} W^{(1)}_{i_2j_2} x_{j_2;\alpha_2}]$$

$$= \mathbb{E}[W^{(1)}_{i_1j_1} W^{(1)}_{i_2j_2}]\, x_{j_1;\alpha_1} x_{j_2;\alpha_2}$$

$$= \frac{C_W}{n}\delta_{i_1i_2}\delta_{j_1j_2}\, x_{j_1;\alpha_1} x_{j_2;\alpha_2}$$

$$= C_W \Big(\underbrace{\frac{1}{n}\vec{x}_{\alpha_1}\cdot\vec{x}_{\alpha_2}}_{G^{(0)}_{\alpha_1\alpha_2}\ \equiv\ \text{covariance btw. two inputs}}\Big)\delta_{i_1i_2}$$

Note this is diagonal in neural indices: no covariance btw. neurons 1 and 2

Now write a recursion for 2-point correlator in layer ℓ, with the ansatz $\mathbb{E}[z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_2;\alpha_2}] = G^{(\ell)}_{\alpha_1\alpha_2}\delta_{i_1i_2}$. Our initial condition is $G^{(1)}_{\alpha_1\alpha_2} = C_W G^{(0)}_{\alpha_1\alpha_2}$.

Continuing, $\mathbb{E}[z^{(\ell+1)}_{i_1;\alpha_1} z^{(\ell+1)}_{i_2;\alpha_2}] = \mathbb{E}[W^{(\ell+1)}_{i_1 j_1} z^{(\ell)}_{j_1;\alpha_1} W^{(\ell+1)}_{i_2 j_2} z^{(\ell)}_{j_2;\alpha_2}]$

Key point: $z^{(\ell)}$ only depends on $W^{(\ell)}, W^{(\ell-1)}, \dots$, so is statistically independent from deeper layers, including $W^{(\ell+1)}$

$$= \mathbb{E}[W^{(\ell+1)}_{i_1 j_1} W^{(\ell+1)}_{i_2 j_2}]\,\mathbb{E}[z^{(\ell)}_{j_1;\alpha_1} z^{(\ell)}_{j_2;\alpha_2}]$$

$$= \frac{C_W}{n}\delta_{i_1 i_2}\underbrace{\delta_{j_1 j_2}\, G^{(\ell)}_{\alpha_1\alpha_2}\,\delta_{j_1 j_2}}_{\mathrm{Tr}(\mathbb{1}_{n\times n})=n} \qquad \text{(ansatz)}$$

$$= C_W G^{(\ell)}_{\alpha_1\alpha_2}\delta_{i_1 i_2}$$

So our ansatz is consistent, and the 2-point recursion is

$$G^{(\ell+1)}_{\alpha_1\alpha_2} = C_W G^{(\ell)}_{\alpha_1\alpha_2} \qquad (\ell = 0, 1, \dots)$$

Solution: $G^{(\ell)}_{\alpha_1\alpha_2} = C_W^{\,\ell}\, G^{(0)}_{\alpha_1\alpha_2}$

We can also sum over i_1, i_2 in our ansatz to find $G^{(\ell)}_{\alpha_1\alpha_2} = \frac{1}{n}\mathbb{E}[\vec{z}_{\alpha_1}\cdot\vec{z}_{\alpha_2}]$

$\Rightarrow G^{(\ell)}_{\alpha_1\alpha_2}$ is a covariance at layer ℓ, and blows up exponentially with depth if $C_W > 1$, or shrinks exponentially if $C_W < 1$.

Can tune the network to criticality by choosing $\boxed{C_W = 1}$, in which case $G^{*}_{\alpha_1\alpha_2} \equiv G^{(0)}_{\alpha_1\alpha_2}$ is a fixed point of the G recursion. Input covariance is preserved during propagation through the network.

(Very) surprising fact about NN's: this tuning is sufficient to prevent exponential behavior of all higher-point correlators, and hence the full $p(z^{(\ell)}|\theta)$

4-point recursion (briefly!): start w/ layer 1,
set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 \equiv \alpha$

$$\mathbb{E}[z^{(1)}_{i_1} z^{(1)}_{i_2} z^{(1)}_{i_3} z^{(1)}_{i_4}] = \mathbb{E}[W_{i_1 j_1} W_{i_2 j_2} W_{i_3 j_3} W_{i_4 j_4}] x_{j_1} x_{j_2} x_{j_3} x_{j_4}$$

Wick's Theorem gives this in terms of C_W:

$$\frac{C_W^2}{n^2}\left(\delta_{i_1 i_2}\delta_{j_1 j_2}\delta_{i_3 i_4}\delta_{j_3 j_4} + (2 \text{ more contractions})\right)$$

$$= C_W^2\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)(G_2^{(0)})^2$$

$\equiv G^{(0)}_{\alpha\alpha}$

This is precisely the structure of a Wick contraction, so can subtract off 2-point correlators to find

$$\mathbb{E}[z^{(1)}_{i_1} z^{(1)}_{i_2} z^{(1)}_{i_3} z^{(1)}_{i_4}]_{\text{conn.}} = 0 \leftarrow$$

Gaussian up to 4th moment, in first layer

However, non-Gaussianities get generated in deeper layers. Taking the ansatz $\mathbb{E}[z^{(\ell)}_{i_1} z^{(\ell)}_{i_2} z^{(\ell)}_{i_3} z^{(\ell)}_{i_4}] = G_4^{(\ell)}(\delta_{i_1 i_2}\delta_{i_3 i_4} + (2 \text{ sim.}))$, compute $\mathbb{E}[z^{(\ell+1)} z^{(\ell+1)} z^{(\ell+1)} z^{(\ell+1)}]$ to derive the recursion

$G_4^{(\ell+1)} = C_W^2\left(1 + \frac{2}{n}\right) G_4^{(\ell)}$. Has a closed-form solution, but instead let's expand perturbatively in $\frac{1}{n}$:

$$\mathbb{E}[z^{4(\ell)}]_{\text{conn.}} \propto G_4^{(\ell)} - (G_2^{(\ell)})^2 = \frac{2(\ell - 1)}{n}(G_2^{(\ell)})^2$$

At criticality ($C_W = 1$), connected 4-pt. grows linearly with depth:

$$\mathbb{E}[z^{4(\ell)}]_{\text{conn.}} \propto \frac{2\ell}{n}(G_2^{*})^2 \leftarrow$$

marginally relevant

Some final comments:

- as $n \to \infty$, 4th and higher cumulants vanish: $p(z^{(l)}|\mathcal{D})$ is purely Gaussian.
- as $L \to \infty$, combinatorial factors eventually grow and spoil criticality. "Infinite size" limit is $n \to \infty$, not $L \to \infty$
- for Gaussian inits, $\frac{L}{n}$ appears as the cutoff of the effective theory. (Not true for other inits in general!)
- there is a closed-form expression for single-input $p(z|x_\alpha)$:

$$p(z^{(l)}|x) \propto G^{l,0}_{0,l}\left(\frac{\vec{z}^{(l)}\cdot\vec{z}^{(l)}}{2^l n C_W^l G_2^{(0)}}\,\middle|\,\begin{matrix}\overline{}\\ 0,0,\dots 0\end{matrix}\right) \qquad [2104.11734]$$

$\uparrow$

depth dependence encoded in Meijer G-function

- if we take our weight matrices to be orthogonal, distributed under the Haar measure on $O(n)$,

$$p(z^{(l)}|x) \propto \delta\left(\vec{z}^{(l)}\cdot z^{(l)} - \vec{x}\cdot\vec{x}\right) \qquad [\text{YK, Hannah Day, D. Roberts}]$$

Duh. $O(n)$ preserves norm, so just a random rotation on the $(n-1)$-sphere. But expand perturbatively:

$$\mathbb{E}[z^{4(l)}]_{\text{conn.}} = -\frac{2}{n}\left(G_2^{(0)}\right)^2$$

independent of l: no $\frac{L}{n}$ cutoff! Exactly marginal.