Lecture 2: Universality Classes of Nonlinear Networks

Let's use the same strategy as last time to compute moments for a nonlinear network. New things:

- allow for Gaussian biases, $\mathbb{E}[b_i] = 0$, $\mathbb{E}[b_i b_j] = C_b \delta_{ij}$
- two common examples for $\sigma(z)$ are
  - ReLU ("rectified linear unit") $\sigma(z) = \max(0, z)$
  - $\sigma(z) = \tanh(z)$

Will see that these belong to different universality classes of network behavior

Recall the NN forward equations:

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

We start by computing the 2-point function:

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

[cross-terms vanish since W and b are independent]

$$= \Big(C_b + \underbrace{\frac{C_W}{n} \vec{x}_{\alpha_1} \cdot \vec{x}_{\alpha_2}}_{G^{(1)}_{\alpha_1 \alpha_2}}\Big) \delta_{i_1 i_2}$$

$G^{(1)}_{\alpha_1\alpha_2}$: call this the first-layer metric

A similar calculation yields $\mathbb{E}[z^{(1)} z^{(1)} z^{(1)} z^{(1)}]_{\text{con.}} = 0$: first-layer distribution $p(z^{(1)} | \mathcal{D})$ is Gaussian to $\mathcal{O}(\frac{1}{n})$. If we wanted, we could write this distribution as an action,

$$p(z^{(1)} | \mathcal{D}) = \frac{1}{[\det(2\pi G^{(1)})]^{n/2}} \exp\left(-\frac{1}{2} G_{(1)}^{\alpha_1 \alpha_2} \vec{z}^{(1)}_{\alpha_1} \cdot \vec{z}^{(1)}_{\alpha_2}\right)$$

inverse metric

which is correct to $\mathcal{O}(\frac{1}{n})$.

An identical computation gives the layer-to-layer marginal distribution with which we build the recursion,

$$p(z^{(\ell+1)} \mid z^{(\ell)}) = \frac{1}{\sqrt{\det(2\pi \hat{G}^{(\ell+1)})^{n}}} \exp\left(-\frac{1}{2} \hat{G}^{\alpha_1\alpha_2}_{(\ell+1)} \vec{z}^{(\ell+1)}_{\alpha_1} \cdot \vec{z}^{(\ell+1)}_{\alpha_2}\right)$$

where $\hat{G}^{(\ell+1)}_{\alpha_1\alpha_2} = C_b + \frac{C_W}{n} \vec{\sigma}^{(\ell)}_{\alpha_1} \cdot \vec{\sigma}^{(\ell)}_{\alpha_2}$ is a stochastic variable

$\vec{\sigma}^{(\ell)}_{\alpha_1} \equiv \sigma(\vec{z}_{\alpha_1})$

since it depends on the random variables $z^{(\ell)}, z^{(\ell-1)}, \ldots z^{(1)}$

In other words, marginal distributions are Gaussian with a stochastic covariance matrix, which results in accumulated non-Gaussianities in $p(z^{(\ell+1)} \mid \theta)$

To determine the recursion we integrate out the ℓ^{th} layer preactivations; this is analogous to one step of RG flow.

$$p(z^{(\ell+1)} \mid \theta) = \int \prod_i dz^{(\ell)}_i \; p(z^{(\ell+1)} \mid z^{(\ell)})\, p(z^{(\ell)} \mid \theta)$$

We can now feel free to build up $p(z^{(\ell+1)})$ from its moments.

Starting with $\ell = 2$: $\underbrace{\mathbb{E}}_{\text{expectation taken over } z^{(1)} \text{ and } z^{(2)}}[z^{(2)}_{i_1;\alpha_1} z^{(2)}_{i_2;\alpha_2}] = \delta_{i_1 i_2} \underbrace{\mathbb{E}}_{\text{only over } z^{(1)}}[\hat{G}^{(2)}_{\alpha_1\alpha_2}] = \delta_{i_1 i_2}(C_b + C_W \underbrace{\mathbb{E}[\sigma^{(1)}_{\alpha_1}\sigma^{(1)}_{\alpha_2}]}_{n \text{ copies of same single-neuron expectation}})$

$$= \delta_{i_1 i_2}(C_b + C_W \underbrace{\langle \sigma^{(1)}_{\alpha_1}\sigma^{(1)}_{\alpha_2} \rangle_{G^{(1)}}}_{\text{Gaussian expectation w.r.t. covariance } G^{(1)}})$$

If we let $G^{(\ell)}_{\alpha_1\alpha_2} \equiv \mathbb{E}[\hat{G}^{(\ell)}_{\alpha_1\alpha_2}]$ and take the ansatz $\mathbb{E}[z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_2;\alpha_2}] = \delta_{i_1 i_2} G^{(\ell)}_{\alpha_1\alpha_2}$, we have the recursion

$$\boxed{G^{(\ell+1)}_{\alpha_1\alpha_2} = C_b + C_W \langle \sigma^{(\ell)}_{\alpha_1}\sigma^{(\ell)}_{\alpha_2} \rangle_{G^{(\ell)}} + \mathcal{O}\left(\frac{1}{n}\right)}$$

for $\ell > 2$, $\mathbb{E}[\sigma^{(\ell)}\sigma^{(\ell)}]$ is Gaussian to $\mathcal{O}(\frac{1}{n})$

We can derive criticality conditions in the $n \to \infty$ limit.

Let $\lim_{n\to\infty} G = K$, so our recursion is $K^{(\ell+1)}_{\alpha\beta} = C_b + C_w \langle \sigma_\alpha \sigma_\beta \rangle_{K^{(\ell)}}$

If σ is nonlinear, e.g. $\sigma(z) = z^2$, we have

$$K^{(\ell+1)}_{\alpha\beta} = C_b + C_w \langle z_\alpha^2 z_\beta^2 \rangle_{K^{(\ell)}} = C_b + C_w \left(\underbrace{K_{\alpha\alpha}^{(\ell)} K_{\beta\beta}^{(\ell)}}_{} + 2 (K_{\alpha\beta}^{(\ell)})^2 \right)$$

Operator mixing under RG!

So unlike in linear networks, we have to consider the whole 2×2 K matrix, rather than just a single input.

The diagonal component is easy because it decouples:

$$K_{00}^{(\ell+1)} = C_b + C_w \langle \sigma^2 \rangle_{K^{(\ell)}} = C_b + C_w \Big[\underbrace{\frac{1}{\sqrt{2\pi K^{(\ell)}}} \int_{-\infty}^{\infty} dz \, e^{\frac{-z^2}{2K^{(\ell)}}} \sigma(z)^2}_{} \Big]$$

$g(K)$: single-variable Gaussian expectation

Find a fixed point K_{00}^* by linearizing: $K_{00}^{(\ell)} = K_{00}^* + \Delta K_{00}^{(\ell)}$

To first order in Δ: $\Delta K_{00}^{(\ell+1)} = \chi_{11}(K_{00}^*) \times \Delta K_{00}^{(\ell)}$

$$\chi_{11} = C_w g'(K) = \frac{C_w}{2K^2} \langle \sigma(z)\sigma(z)(z^2 - K) \rangle_K$$

To ensure we don't move away from the fixed point, our first criticality condition is $\boxed{\chi_{11}(K_{00}^*) = 1}$

Note: when $\sigma(z) = z$, $g'(K) = 1$, so we recover $C_w = 1$ from last time.

If $C_b \neq 0$, we have $K_{00}^{(\ell+1)} = C_b + K_{00}^{(\ell)} \Rightarrow K_{00}$ grows linearly

↳ Semi-critical: fixed point at infinity, so should rather choose $C_b = 0$.

The constancy of K_{00} ($\Leftrightarrow \chi_{\parallel} = 1$) ensures that norms of preactivations don't change exponentially. The off-diagonal components of K measure changes to <u>nearby</u> inputs. Requiring that these also not change exponentially gives a second criticality condition.

Derivation is long but straightforward: find a decomposition of $K_{\alpha\beta}$ that diagonalizes the RG evolution, and linearize about a fixed point.

$$\Rightarrow \text{ require } \chi_{\perp}(K^*) = 1, \text{ where } \quad \chi_{\perp}(K) = C_w \langle \sigma'(z)^2 \rangle_K$$

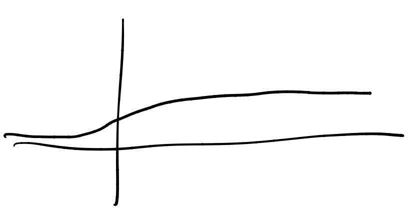
To summarize:

$$\text{criticality} \Rightarrow \begin{matrix} \chi_{\parallel} = 1 \\ \chi_{\perp} = 1 \end{matrix} \Rightarrow \frac{\chi_{\perp}}{\chi_{\parallel}} = 1, \text{ solve for } C_b \text{ and } C_w$$

$$\boxed{\left[\frac{2K^2 \langle \sigma'(z)^2 \rangle_K}{\langle \sigma(z)^2 (z^2 - K) \rangle_K} \right]_{K=K^*} = 1, \qquad \begin{matrix} C_w = \dfrac{1}{\langle \sigma'(z)^2 \rangle_{K^*}} \\[2ex] C_b = K^* - \dfrac{\langle \sigma(z)^2 \rangle_{K^*}}{\langle \sigma'(z)^2 \rangle_{K^*}} \end{matrix}}$$

Note: not every activation function has a consistent solution!

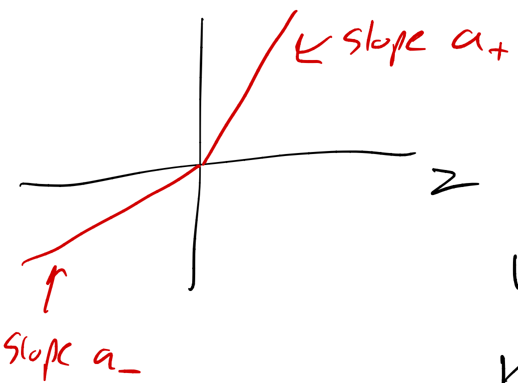
e.g. $\sigma(z) = \dfrac{1}{1+e^{-z}}$



Used historically... but criticality implies $C_b = -\left(\frac{\sigma(0)^2}{\sigma'(0)^2}\right) < 0$ and a negative variance is unphysical. The problem is $\sigma(0) \neq 0$!

The criticality equations can be solved numerically, or by inspection. Different values of K^* correspond to different universality classes:

- Scale-invariant (i.e. piecewise-linear w/ $\sigma(0)=0$)



$$\Rightarrow K^* = \frac{2}{a_+^2+a_-^2}\left[\frac{1}{n}\vec{x}\cdot\vec{x}\right]$$

when $C_b=0$ and $C_W = \frac{2}{a_+^2+a_-^2}$,

K^* is constant as a function of depth: a line of nontrivial fixed points corresponding to different input norms.

Linear activation has $a_+=a_-=1 \Rightarrow C_W=1$, as we found before.

ReLU has $a_-=0$, $a_+=1 \Rightarrow C_W=2$, to compensate for the fact that $\max(0,z)$ knocks out half the activations on average.

- $K^*=0$: consider a smooth activation function

$$\sigma(z) = \sum_{p=0}^{\infty} \frac{\sigma_p}{p!} z^p.$$

$$\Rightarrow \langle\sigma^2(z)\rangle_K = \sigma_0^2 + (\sigma_1^2 + 2\sigma_0\sigma_2)K + \mathcal{O}(K^2)$$

$\Rightarrow K^* = C_b + C_W\langle\sigma^2\rangle_{K^*}$ has a solution, $K^*=0$, iff $\sigma_0=0$ and $C_b=0$

Linearizing $\chi_\parallel$ and $\chi_\perp$ about $z=0$ and expanding in K, we find $C_W = \frac{1}{\sigma_1^2}$. For $\sigma(z)=\tanh z$, $\sigma_1=1$, and $C_W=1$.

In the $K^* = 0$ universality class, K decays to 0, but only like a power law. Linearizing the recursions gives $K^{(\ell)} = \frac{1}{2\ell}$ for tanh (and $\# \times \frac{1}{\ell}$ for other activations), up to $\mathcal{O}(\frac{1}{\ell^2})$ corrections. In other words, K is marginally irrelevant.

Fluctuations

Using the same techniques as before, we can derive recursions for the 4-point correlator. Non-Gaussianity comes from both $\mathbb{E}[(\hat{G}-G)^2]$ and $\det(2\pi\hat{G})$, this is a long calculation.

For a single input, let's call the coefficient of the wick tensor structure V (meant to remind us of a 4-point vertex)

Recursion is $V^{(\ell)} = \chi_{\parallel}^2(K^{(\ell)})V^{(\ell)} + C_w^2[\langle\sigma^4\rangle_{K^{(\ell)}} - \langle\sigma^2\rangle^2_{K^{(\ell)}}]$

Amazingly, tuning C_w to criticality tames exponential behavior of $V^{(\ell)}$ too!

- Scale-invariant: $V^{(\ell)} = (\ell-1)(\#)(K^*)^2$ (marginally relevant)
- $K^* = 0$: $V^{(\ell)} = (\#)\left(\frac{1}{\ell}\right)$ (marginally irrelevant)

Note that we can assign a power-counting dimension to z. $[K] = 2$ and $[V] = 4$, so the dimensionless correlator is

$$\frac{V}{K^2} \overset{\mathcal{O}(\frac{1}{n})}{\sim} \frac{\ell}{n}$$ for both universality classes!

This is the same linear growth of fluctuations we saw in a linear network, but now we know it also characterizes nonlinear networks.

Caveat: with orthogonal initializations $\frac{V}{K^2}$ is constant with ℓ for $K^* = 0$, but scales as $\frac{\ell}{n}$ for scale-inv., except linear activations which give $\frac{V}{K^2}$ constant with depth.

$\Rightarrow$ $K^* = 0$ is linearizing an activation function at large depths, so nonlinear networks sort of behave linearly.