Deep Linear Networks

By considering networks with many layers and no activation function we can introduce the machinery that will be used to analyse more general networks later. In particular this chapter shows how layer-to-layer recursions determine the statistics (M-point (connected) correlators) of neural networks at initialisation, and solves the recursion exactly for selected correlators. These correlators are measurements of NN behaviour and ways of controlling these behaviors are discussed.

In this chapter we consider NNs with linear transformations at each layer with no biases:

$$z_{i;\alpha}^{(\ell)} = \sum_{j_0=1}^{n_0} \sum_{j_1=1}^{n_1} \cdots \sum_{j_{\ell-1}=1}^{n_{\ell-1}} W_{ij_{\ell-1}}^{(\ell)} W_{j_{\ell-1}j_{\ell-2}}^{(\ell-1)} \cdots W_{j_1 j_0}^{(1)} x_{j_0;\alpha} \equiv \sum_{j=1}^{n_0} \mathcal{W}_{ij}^{(\ell)} x_{j;\alpha} \,. \tag{3.2}$$

Here we have introduced an n_ℓ-by-n_0 matrix

$$\mathcal{W}_{ij}^{(\ell)} = \sum_{j_1=1}^{n_1} \cdots \sum_{j_{\ell-1}=1}^{n_{\ell-1}} W_{ij_{\ell-1}}^{(\ell)} W_{j_{\ell-1}j_{\ell-2}}^{(\ell-1)} \cdots W_{j_1 j}^{(1)} \,, \tag{3.3}$$

These layers are initialised independently according to a normal distribution with mean 0:

$$\mathbb{E}\left[W_{ij}^{(\ell)}\right] = 0 \,, \qquad \mathbb{E}\left[W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)}\right] = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_W}{n_{\ell-1}} \,. \tag{3.4}$$

Deep linear networks represent a smaller set of functions than generic linear transformations: consider a 2 layer neural network with inputs n_0, outputs n_2 and a hidden layer with one neuron. This bottleneck means the network can represent only a subset of the transformations given by all $n_2 \times n_0$ matrices.

Similarly, the statistics of a deep network differ from the statistics of a one-layer network: each layer has gaussian $W_{ij}^{(l)}$, but the product $\mathcal{W}_{ij}^{(l)}$ is in general non-gaussian.

In this chapter we want to determine:

$$p\left(z^{(\ell)} \middle| \mathcal{D}\right) \equiv p\left(z^{(\ell)}(x_1), \ldots, z^{(\ell)}(x_{N_\mathcal{D}})\right) \,,$$

Which (like any distribution) is determined by its M-point correlators.

It is trivial to show that $\mathbb{E}[z^{(l)}] = 0$ by taking the expectation of (3.2). Similarly its possible to show that the odd M-point correlators is also 0.

Criticality

Recursion for the two point correlator:

$$\begin{aligned}\mathbb{E}\left[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_2;\alpha_2}\right] &= \sum_{j_1,j_2=1}^{n_0} \mathbb{E}\left[W^{(1)}_{i_1 j_1} x_{j_1;\alpha_1} W^{(1)}_{i_2 j_2} x_{j_2;\alpha_2}\right] \\ &= \sum_{j_1,j_2=1}^{n_0} \mathbb{E}\left[W^{(1)}_{i_1 j_1} W^{(1)}_{i_2 j_2}\right] x_{j_1;\alpha_1} x_{j_2;\alpha_2} \\ &= \sum_{j_1,j_2=1}^{n_0} \frac{C_W}{n_0} \delta_{i_1 i_2} \delta_{j_1 j_2} x_{j_1;\alpha_1} x_{j_2;\alpha_2} = \delta_{i_1 i_2} C_W \frac{1}{n_0} \sum_{j=1}^{n_0} x_{j;\alpha_1} x_{j;\alpha_2} \,,\end{aligned} \tag{3.8}$$

Writing the inner product of $\mathbf{x}_{\alpha_1}$ and $\mathbf{x}_{\alpha_2}$ as:

$$G^{(0)}_{\alpha_1\alpha_2} \equiv \frac{1}{n_0} \sum_{i=1}^{n_0} x_{i;\alpha_1} x_{i;\alpha_2} \,,$$

We get:

$$\mathbb{E}\left[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_2;\alpha_2}\right] = \delta_{i_1 i_2} C_W G^{(0)}_{\alpha_1\alpha_2} \,.$$

Next to evaluate the recursion for an arbitrary layer:

$$\begin{aligned}\mathbb{E}\left[z^{(\ell+1)}_{i_1;\alpha_1} z^{(\ell+1)}_{i_2;\alpha_2}\right] &= \sum_{j_1,j_2=1}^{n_\ell} \mathbb{E}\left[W^{(\ell+1)}_{i_1 j_1} W^{(\ell+1)}_{i_2 j_2} z^{(\ell)}_{j_1;\alpha_1} z^{(\ell)}_{j_2;\alpha_2}\right] \\ &= \sum_{j_1,j_2=1}^{n_\ell} \mathbb{E}\left[W^{(\ell+1)}_{i_1 j_1} W^{(\ell+1)}_{i_2 j_2}\right] \mathbb{E}\left[z^{(\ell)}_{j_1;\alpha_1} z^{(\ell)}_{j_2;\alpha_2}\right] \\ &= \delta_{i_1 i_2} C_W \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{E}\left[z^{(\ell)}_{j;\alpha_1} z^{(\ell)}_{j;\alpha_2}\right] \,,\end{aligned}$$

Notice that for any layer the two-point correlator vanishes unless the neural indices i_1, i_2 are the same and is this proportional to the kronecker delta $\delta_{i_1 i_2}$.

By considering the last sentence and looking at this for a while you should be able to see that:

$$\mathbb{E}\left[z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_2;\alpha_2}\right] \equiv \delta_{i_1 i_2} G^{(\ell)}_{\alpha_1\alpha_2} \,,$$

and that:

$$G^{(\ell)}_{\alpha_1\alpha_2} = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{E}\left[z^{(\ell)}_{j;\alpha_1} z^{(\ell)}_{j;\alpha_2}\right] \,,$$

So the quantity $G^{(l)}_{\alpha_1\alpha_2}$ can be thought of as the average inner product of activations at a given layer, averaged over neurons. This depends on the sample indices α_1, α_2 only so $G^{(l)}_{\alpha_1\alpha_2} = G^{(l)}(x_{\alpha_1}, x_{\alpha_2})$ can be interpreted as the covariance of the two inputs after passing through layer l.

Now it is easy to see that $G^{(l+1)}_{\alpha_1\alpha_2} = C_W G^{(l)}_{\alpha_1\alpha_2}$ and that $G^{(l+1)}_{\alpha_1\alpha_2} = C^l_W G^{(0)}_{\alpha_1\alpha_2}$

Note: n_l, the width of the network at each layer, in the initialisation of the layers has dropped out: indicating that this is the proper way of scaling the variance.

Criticality: Physics

Now we can see that if $C_W < 1$ the covariance will vanish to 0 and if $C_W > 1$ the covariance will blow up to ∞. The authors refer to any fixed point approached exponentially quickly as a **trivial fixed point**. Such behavior would make it difficult for the NN to approximate the desired function.

If instead $C_W = 1$ we have that the variance is preserved and the covariance (since it is not tending exponentially quickly towards anything) tends towards a **non-trivial fixed point**. A setting of the initialisation hyperparameters that avoids blow up or vanishing is called a **critical initialisation hyperparameters**.

Fluctuations

For zero mean gaussians the covariance completely determines the distribution, so if the distribution $p(z^{(l)}|\mathcal{D})$ were gaussian the critical tuning of $C_W = 1$ ensures all observables are well behaved. If this distribution isn't Gaussian then the behavior of observables that

depend on higher-point connected correlators may not be well behaved with the tuning that makes the covariance well behaved.

Now we consider the recursion for the four-point correlator and for simplicity only consider correlators that depend on one input:

$$G_2^{(\ell)} \equiv G_{\alpha\alpha}^{(\ell)} = G^{(\ell)}(x,x)\,.$$

Recursion for the first layer:

$$\begin{aligned}
&\mathbb{E}\left[z_{i_1}^{(1)} z_{i_2}^{(1)} z_{i_3}^{(1)} z_{i_4}^{(1)}\right] \qquad (3.18)\\
=&\sum_{j_1,j_2,j_3,j_4=1}^{n_0} \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)} W_{i_3 j_3}^{(1)} W_{i_4 j_4}^{(1)}\right] x_{j_1} x_{j_2} x_{j_3} x_{j_4}\\
=&\frac{C_W^2}{n_0^2} \sum_{j_1,j_2,j_3,j_4=1}^{n_0} \left(\delta_{i_1 i_2}\delta_{j_1 j_2}\delta_{i_3 i_4}\delta_{j_3 j_4} + \delta_{i_1 i_3}\delta_{j_1 j_3}\delta_{i_2 i_4}\delta_{j_2 j_4} + \delta_{i_1 i_4}\delta_{j_1 j_4}\delta_{i_2 i_3}\delta_{j_2 j_3}\right) x_{j_1} x_{j_2} x_{j_3} x_{j_4}\\
=&C_W^2 \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)\left(G_2^{(0)}\right)^2 .
\end{aligned}$$

Going from the second line to the first we take the wick contraction which yields 3 pairings of the W_{ij}'s (double factorial $(2*2-1)!! = 3$) and used (3.4) to evaluate the weight variance. Then evaluate the sums over the j's and substitute the the definition of the inner product from above:

$$G_2^{(0)} = \frac{1}{n_0}\sum_{j=1}^{n_0} x_j x_j\,.$$

This is precisely what we'd expect for the fourpoint correlator if the preactivation distribution were Gaussian. For deeper layers this won't be the case:

$$
\begin{aligned}
&\mathbb{E}\left[z_{i_1}^{(\ell+1)} z_{i_2}^{(\ell+1)} z_{i_3}^{(\ell+1)} z_{i_4}^{(\ell+1)}\right] \qquad (3.20)\\
=&\sum_{j_1,j_2,j_3,j_4=1}^{n_\ell} \mathbb{E}\left[W_{i_1 j_1}^{(\ell+1)} W_{i_2 j_2}^{(\ell+1)} W_{i_3 j_3}^{(\ell+1)} W_{i_4 j_4}^{(\ell+1)}\right] \mathbb{E}\left[z_{j_1}^{(\ell)} z_{j_2}^{(\ell)} z_{j_3}^{(\ell)} z_{j_4}^{(\ell)}\right]\\
=&\frac{C_W^2}{n_\ell^2} \sum_{j_1,j_2,j_3,j_4=1}^{n_\ell} \left(\delta_{i_1 i_2}\delta_{j_1 j_2}\delta_{i_3 i_4}\delta_{j_3 j_4} + \delta_{i_1 i_3}\delta_{j_1 j_3}\delta_{i_2 i_4}\delta_{j_2 j_4} + \delta_{i_1 i_4}\delta_{j_1 j_4}\delta_{i_2 i_3}\delta_{j_2 j_3}\right)\\
&\qquad \times \mathbb{E}\left[z_{j_1}^{(\ell)} z_{j_2}^{(\ell)} z_{j_3}^{(\ell)} z_{j_4}^{(\ell)}\right]\\
=&C_W^2 \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \frac{1}{n_\ell^2} \sum_{j,k=1}^{n_\ell} \mathbb{E}\left[z_j^{(\ell)} z_j^{(\ell)} z_k^{(\ell)} z_k^{(\ell)}\right] ,
\end{aligned}
$$

Where we use the fact that the $(l+1)$th layer weights are independent from the (l)th layer activations.

Similarly in the covariance case we notice that *any* layer is proportional to the factor $(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_4})$ we can write:

$$
\mathbb{E}\left[z_{i_1}^{(\ell)} z_{i_2}^{(\ell)} z_{i_3}^{(\ell)} z_{i_4}^{(\ell)}\right] \equiv \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) G_4^{(\ell)} ,
$$

and put all the layer dependence into $G_4^{(l)}$ so for the first layer we get:

$$
G_4^{(1)} = C_W^2 \left(G_2^{(0)}\right)^2 ,
$$

and the final term in the above summation becomes:

$$
\frac{1}{n_\ell^2} \sum_{j,k=1}^{n_\ell} \mathbb{E}\left[z_j^{(\ell)} z_j^{(\ell)} z_k^{(\ell)} z_k^{(\ell)}\right] = \frac{1}{n_\ell^2} \sum_{j,k=1}^{n_\ell} \left(\delta_{jj}\delta_{kk} + \delta_{jk}\delta_{jk} + \delta_{jk}\delta_{kj}\right) G_4^{(\ell)} = \left(1 + \frac{2}{n_\ell}\right) G_4^{(\ell)} . \qquad (3.23)
$$

(First kronecker delta sum is n_l^2 second is n_l and third is n_l) so we can rewrite the recursion:

$$
G_4^{(\ell+1)} = C_W^2 \left(1 + \frac{2}{n_\ell}\right) G_4^{(\ell)} .
$$

and then the recursion can be solved using the initial condition from above:

$$G_4^{(\ell)} = C_W^{2\ell}\left[\prod_{\ell'=1}^{\ell-1}\left(1+\frac{2}{n_{\ell'}}\right)\right]\left(G_2^{(0)}\right)^2 \tag{3.25}$$
$$= \left[\prod_{\ell'=1}^{\ell-1}\left(1+\frac{2}{n_{\ell'}}\right)\right]\left(G_2^{(\ell)}\right)^2 ,$$

Resulting physics

The point of doing this is to analyse the behavior of the four point correlator. For instance, taking the limit as $n_l \to \infty$ notice that the full four point correlator becomes:

$$\mathbb{E}\left[z_{i_1}^{(\ell)} z_{i_2}^{(\ell)} z_{i_3}^{(\ell)} z_{i_4}^{(\ell)}\right] = \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)\left(G_2^{(\ell)}\right)^2 .$$

I.e in the infinite width limit the preactivation distributions are gaussian and the four point correlator is determined by the two point correlator.

If we instead have equal hidden layer widths for all layers we get for the deviation of the four point correlator from that in the infinite width limit:

$$G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2 = \left[\left(1+\frac{2}{n}\right)^{\ell-1} - 1\right]\left(G_2^{(\ell)}\right)^2$$
$$= \frac{2(\ell-1)}{n}\left(G_2^{(\ell)}\right)^2 + O\left(\frac{1}{n^2}\right) ,$$

Where we expand in $1/n$ and keep the leading correction to the infinite width limit. At criticality where $G_2^{(l)}$ is constant we have that the correction scales proportionally with depth and inversely with width, and is thus proportional to the depth-to-width ratio of the network, something the authors refer to as **emergent scale**.

They give some examples of interpreting this correction, see the text book for more details.

The take away point is that networks show these finite-width effects where behavior depends on the depth-to-width ratio.

Summary: