Last Chapter

- Computed statistics of preactivations for a deep neural network without activation functions
- Using Wick contractions and the recursive structure of the network we were able to understand the effect of the initialization scheme, depth, width on preactivation correlators
- This highlighted the importance of critical initialization hyperparamters and a sufficiently small depth-to-width ratio in order for the networks outputs to be "well behaved"

In This Chapter

- Overall goal is to discover how a particular neural network learns from a dataset. To that end they start by investigating how *ensembles* of neural networks behave at initialisation, as a function of the data.
- By doing this they hope to isolate the *typical* behavior of a neural network, and how any particular neural network might fluctuate from this typicality.
- In the infinite width limit NNs become GPs (with a fixed kernel) and show reduced capacity for representation learning. To investigate what happens in the large but not infinite width regime, they propose to use a $1/n$ expansion.
- They proceed recursively, going layer by layer.

First Layer

At initialization the biases $b^{(1)}$ and weights $W^{(1)}$ are independently distributed according to mean-zero Gaussian distributions with variances

$$\mathbb{E}\left[b_i^{(1)} b_j^{(1)}\right] = \delta_{ij} C_b^{(1)} , \tag{4.3}$$

$$\mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W^{(1)}}{n_0} . \tag{4.4}$$

The first-layer preactivations $z^{(1)} = z_{i;\alpha}^{(1)}$ form an $(n_1 N_{\mathcal{D}})$-dimensional vector, and we are interested in its distribution at initialization,

$$p\left(z^{(1)}\middle|\mathcal{D}\right) = p\left(z^{(1)}(x_1), \ldots, z^{(1)}(x_{N_{\mathcal{D}}})\right) . \tag{4.5}$$

Next step: Two derivations of the distributions of the first-layer preactivations at initialisation.

Via Wick Contractions

Starting with one-point correlator:

$$\mathbb{E}\left[z_{i;\alpha}^{(1)}\right] = \mathbb{E}\left[b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_{j;\alpha_1}\right] = 0\,, \tag{4.6}$$

Since bias and weights have 0 mean. In fact all odd-point correlators of $p(z^{(1)}|\mathcal{D})$ vanish because there are always an odd number of biases or weights left unpaired under Wick contractions.

For two-point correlator:

$$\begin{aligned}\mathbb{E}\left[z_{i_1;\alpha_1}^{(1)} z_{i_2;\alpha_2}^{(1)}\right] &= \mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1;\alpha_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2;\alpha_2}\right)\right] \\ &= \delta_{i_1 i_2}\left(C_b^{(1)} + C_W^{(1)} \frac{1}{n_0}\sum_{j=1}^{n_0} x_{j;\alpha_1} x_{j;\alpha_2}\right) = \delta_{i_1 i_2} G_{\alpha_1\alpha_2}^{(1)}\,,\end{aligned} \tag{4.7}$$

where to get to the second line we Wick-contracted the biases and weights using (4.3) and (4.4). We also introduced the first-layer **metric**

$$G_{\alpha_1\alpha_2}^{(1)} \equiv C_b^{(1)} + C_W^{(1)} \frac{1}{n_0}\sum_{j=1}^{n_0} x_{j;\alpha_1} x_{j;\alpha_2}\,, \tag{4.8}$$

This is a function of the two samples. $G_{\alpha_1\alpha_2}^{(1)} = G^{(1)}(x_{\alpha_1}, x_{\alpha_2})$ and represents the two-point correlations of preactivations in the first layer between different samples.

Higher-point correlators can be found similarly, here for the four-point correlator:

$$\begin{aligned}&\mathbb{E}\left[z_{i_1;\alpha_1}^{(1)} z_{i_2;\alpha_2}^{(1)} z_{i_3;\alpha_3}^{(1)} z_{i_4;\alpha_4}^{(1)}\right] \\ =&\delta_{i_1 i_2}\delta_{i_3 i_4} G_{\alpha_1\alpha_2}^{(1)} G_{\alpha_3\alpha_4}^{(1)} + \delta_{i_1 i_3}\delta_{i_2 i_4} G_{\alpha_1\alpha_3}^{(1)} G_{\alpha_2\alpha_4}^{(1)} + \delta_{i_1 i_4}\delta_{i_2 i_3} G_{\alpha_1\alpha_4}^{(1)} G_{\alpha_2\alpha_3}^{(1)} \\ =&\mathbb{E}\left[z_{i_1;\alpha_1}^{(1)} z_{i_2;\alpha_2}^{(1)}\right]\mathbb{E}\left[z_{i_3;\alpha_3}^{(1)} z_{i_4;\alpha_4}^{(1)}\right] + \mathbb{E}\left[z_{i_1;\alpha_1}^{(1)} z_{i_3;\alpha_3}^{(1)}\right]\mathbb{E}\left[z_{i_2;\alpha_2}^{(1)} z_{i_4;\alpha_4}^{(1)}\right] \\ &+ \mathbb{E}\left[z_{i_1;\alpha_1}^{(1)} z_{i_4;\alpha_4}^{(1)}\right]\mathbb{E}\left[z_{i_2;\alpha_2}^{(1)} z_{i_3;\alpha_3}^{(1)}\right]\,.\end{aligned} \tag{4.9}$$

The end result is the same as Wick-contracting $z^{(1)}$'s with the variance given by (4.7). recalling chapter 1, we remember that can be summed up by saying the connected four-point correlator vanishes (since the $z^{(1)}$'s are gaussian). Similarly all higher-point

correlators also vanish. So all correlators can be generated from a Gaussian with zero mean and the variance (4.7).

So to write down the first-layer action we need the inverse of this variance:

$$\sum_{j=1}^{n_1} \sum_{\beta \in \mathcal{D}} \left(\delta_{i_1 j} G_{(1)}^{\alpha_1 \beta} \right) \left(\delta_{j i_2} G_{\beta \alpha_2}^{(1)} \right) = \delta_{i_1 i_2} \delta^{\alpha_1}{}_{\alpha_2} \,, \tag{4.11}$$

with the inverse of the first-layer metric $G^{(1)}_{\alpha_1 \alpha_2}$ denoted as $G_{(1)}^{\alpha_1 \alpha_2}$ and defined by

$$\sum_{\beta \in \mathcal{D}} G_{(1)}^{\alpha_1 \beta} G_{\beta \alpha_2}^{(1)} = \delta^{\alpha_1}{}_{\alpha_2} \,. \tag{4.12}$$

Just as in §1, we follow the conventions of *general relativity* and suppress the superscript "-1" for the inverse metric, distinguishing the metric $G^{(1)}_{\alpha_1 \alpha_2}$ and the inverse metric $G_{(1)}^{\alpha_1 \alpha_2}$ by whether sample indices are lowered or raised. With this notation, the Gaussian distribution for the first-layer preactivations is expressed as

$$p\left(z^{(1)} \middle| \mathcal{D}\right) = \frac{1}{Z} e^{-S\left(z^{(1)}\right)} \,, \tag{4.13}$$

with the quadratic action

$$S\left(z^{(1)}\right) = \frac{1}{2} \sum_{i=1}^{n_1} \sum_{\alpha_1, \alpha_2 \in \mathcal{D}} G_{(1)}^{\alpha_1 \alpha_2} z_{i;\alpha_1}^{(1)} z_{i;\alpha_2}^{(1)} \,, \tag{4.14}$$

and the partition function

$$Z = \int \left[\prod_{i,\alpha} dz_{i;\alpha}^{(1)} \right] e^{-S\left(z^{(1)}\right)} = \left| 2\pi G^{(1)} \right|^{\frac{n_1}{2}} \,, \tag{4.15}$$

Via Hubbard-Stratonovich

Instead of computing correlators and backing out the distribution that generates them, instead we can work directly with the distribution. Using the formal expression for the preactivation distribution worked out in the last chapter:

$$p(z|\mathcal{D}) = \int \left[\prod_i db_i \; p(b_i) \right] \left[\prod_{i,j} dW_{ij} \; p(W_{ij}) \right] \prod_{i,\alpha} \delta \left(z_{i;\alpha} - b_i - \sum_j W_{ij} x_{j;\alpha} \right) \,, \tag{4.16}$$

We could try to eliminate some of the integrals over the model parameters with respect to the constraints from the delta-functions, but it can become confusing due to the different numbers of model-parameter integrals and delta-function constraints.

So to simplify things we can use the **Hubbard-Stratonovich transformation**, using the following integral representation of the Dirac delta function:

$$\delta(z-a) = \int \frac{d\Lambda}{2\pi} e^{i\Lambda(z-a)} \tag{4.17}$$

for each constraint and also plugging in explicit expressions for the Gaussian distributions over the parameters, we obtain

$$\begin{aligned} p(z|\mathcal{D}) = \int \left[\prod_i \frac{db_i}{\sqrt{2\pi C_b}}\right] \left[\prod_{i,j} \frac{dW_{ij}}{\sqrt{2\pi C_W/n_0}}\right] \left[\prod_{i,\alpha} \frac{d\Lambda_i{}^\alpha}{2\pi}\right] \\ \times \exp\left[-\sum_i \frac{b_i^2}{2C_b} - n_0 \sum_{i,j} \frac{W_{ij}^2}{2C_W} + i \sum_{i,\alpha} \Lambda_i{}^\alpha \left(z_{i;\alpha} - b_i - \sum_j W_{ij} x_{j;\alpha}\right)\right] . \end{aligned} \tag{4.18}$$

Completing the square for the biases b and weights W gives a quadratic action in model parameters:

$$\begin{aligned} &-\sum_i \frac{b_i^2}{2C_b} - n_0 \sum_{i,j} \frac{W_{ij}^2}{2C_W} + i \sum_{i,\alpha} \Lambda_i{}^\alpha \left(z_{i;\alpha} - b_i - \sum_j W_{ij} x_{j;\alpha}\right) \\ &= -\frac{1}{2C_b} \sum_i \left(b_i + iC_b \sum_\alpha \Lambda_i{}^\alpha\right)^2 - \frac{C_b}{2} \sum_i \left(\sum_\alpha \Lambda_i{}^\alpha\right)^2 \\ &\quad - \frac{n_0}{2C_W} \sum_{i,j} \left(W_{ij} + i\frac{C_W}{n_0} \sum_\alpha \Lambda_i{}^\alpha x_{j;\alpha}\right)^2 - \frac{C_W}{2n_0} \sum_{i,j} \left(\sum_\alpha \Lambda_i{}^\alpha x_{j;\alpha}\right)^2 + i \sum_{i,\alpha} \Lambda_i{}^\alpha z_{i;\alpha} . \end{aligned} \tag{4.19}$$

Then the weights and biases can be integrated out:

$$\int \left[\prod_{i,\alpha} \frac{d\Lambda_i{}^\alpha}{2\pi}\right] \exp\left[-\frac{1}{2} \sum_{i,\alpha_1,\alpha_2} \Lambda_i{}^{\alpha_1} \Lambda_i{}^{\alpha_2} \left(C_b + C_W \sum_j \frac{x_{j;\alpha_1} x_{j;\alpha_2}}{n_0}\right) + i \sum_{i,\alpha} \Lambda_i{}^\alpha z_{i;\alpha}\right] . \tag{4.20}$$

In effect what we have done is swap the delta-function constraints and model parameters for the Hubbard-Statonovich variables Λ_i^α which have a quadratic action (the first term above) and a linear interaction with the preactivations (second term).

Note the inverse variance here is the first layer metric (4.8) in the wick contraction derivation (restoring layer superscripts):

$$C_b^{(1)} + C_W^{(1)} \sum_j \frac{x_{j;\alpha_1} x_{j;\alpha_2}}{n_0} = G^{(1)}_{\alpha_1 \alpha_2} \,, \tag{4.21}$$

Then we can complete the square again, and integrate out the H-S variables to recover the previous result:

$$-\frac{1}{2} \sum_{i,\alpha_1,\alpha_2} \left[G^{(1)}_{\alpha_1\alpha_2} \left(\Lambda_i{}^{\alpha_1} - i \sum_{\beta_1} G^{\alpha_1\beta_1}_{(1)} z^{(1)}_{i;\beta_1} \right) \left(\Lambda_i{}^{\alpha_2} - i \sum_{\beta_2} G^{\alpha_2\beta_2}_{(1)} z^{(1)}_{i;\beta_2} \right) + G^{\alpha_1\alpha_2}_{(1)} z^{(1)}_{i;\alpha_1} z^{(1)}_{i;\alpha_2} \right] , \tag{4.22}$$

which finally lets us integrate out the Hubbard-Stratonovich variables $\Lambda_i{}^{\alpha}$ and recover our previous result

$$p\left(z^{(1)} \middle| \mathcal{D}\right) = \frac{1}{\left|2\pi G^{(1)}\right|^{\frac{n_1}{2}}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n_1} \sum_{\alpha_1,\alpha_2 \in \mathcal{D}} G^{\alpha_1\alpha_2}_{(1)} z^{(1)}_{i;\alpha_1} z^{(1)}_{i;\alpha_2} \right) . \tag{4.23}$$

Quadratic action in action

Now given this action representation for the distribution of the first layer preactivations we can compute some expectations, in particular we can compute the expectation of two actiavtions on the same neuron, and the expectation of four activations either with all on the same neuron or pairs on two different neurons:

$$\begin{aligned}
&\mathbb{E}\left[\sigma\left(z^{(1)}_{i_1;\alpha_1}\right)\sigma\left(z^{(1)}_{i_1;\alpha_2}\right)\right] \qquad (4.24)\\
&= \int \left[\prod_{i=1}^{n_1} \frac{\prod_{\alpha\in\mathcal{D}} dz_{i;\alpha}}{\sqrt{\left|2\pi G^{(1)}\right|}} \right] \exp\left(-\frac{1}{2} \sum_{j=1}^{n_1} \sum_{\beta_1,\beta_2\in\mathcal{D}} G^{\beta_1\beta_2}_{(1)} z_{j;\beta_1} z_{j;\beta_2} \right) \sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2}) \\
&= \left\{ \prod_{i\neq i_1} \int \left[\frac{\prod_{\alpha\in\mathcal{D}} dz_{i;\alpha}}{\sqrt{\left|2\pi G^{(1)}\right|}} \right] \exp\left(-\frac{1}{2} \sum_{\beta_1,\beta_2\in\mathcal{D}} G^{\beta_1\beta_2}_{(1)} z_{i;\beta_1} z_{i;\beta_2} \right) \right\} \\
&\quad \times \int \left[\frac{\prod_{\alpha\in\mathcal{D}} dz_{i_1;\alpha}}{\sqrt{\left|2\pi G^{(1)}\right|}} \right] \exp\left(-\frac{1}{2} \sum_{\beta_1,\beta_2\in\mathcal{D}} G^{\beta_1\beta_2}_{(1)} z_{i_1;\beta_1} z_{i_1;\beta_2} \right) \sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2}) \\
&= \{1\} \times \left[\int \frac{\prod_{\alpha\in\mathcal{D}} dz_{\alpha}}{\sqrt{\left|2\pi G^{(1)}\right|}} \right] \exp\left(-\frac{1}{2} \sum_{\beta_1,\beta_2\in\mathcal{D}} G^{\beta_1\beta_2}_{(1)} z_{\beta_1} z_{\beta_2} \right) \sigma(z_{\alpha_1})\,\sigma(z_{\alpha_2}) \\
&\equiv \left\langle \sigma(z_{\alpha_1})\,\sigma(z_{\alpha_2}) \right\rangle_{G^{(1)}} .
\end{aligned}$$

The second equality is because the probability distribution factorises for each neuron ($e^{(x+y)} = e^x e^y$). The third equality comes about because the first set of integrals are all trivial, and from renaming the dummy variables $z_{i_1;\alpha_1} \to z_{\alpha_1}$. The final equality uses the notation:

$$\langle F\left(z_{\alpha_1}, \ldots, z_{\alpha_m}\right)\rangle_g \equiv \int \left[\frac{\prod_{\alpha\in\mathcal{D}} dz_\alpha}{\sqrt{|2\pi g|}}\right] \exp\left(-\frac{1}{2}\sum_{\beta_1,\beta_2\in\mathcal{D}} g^{\beta_1\beta_2} z_{\beta_1} z_{\beta_2}\right) F(z_{\alpha_1}, \ldots, z_{\alpha_m}) \tag{4.25}$$

This is a gaussian expectation with variance g and an arbitrary function $F(z_{\alpha_1}, \ldots, z_{\alpha_m})$ over variables with sample indices only. For this chapter we consider computations complete when they can be reduces to gaussian expectations like these.

Using the shorthand $\sigma_\alpha \equiv \sigma(z_\alpha)$, the computation above can be expressed simply as:

$$\mathbb{E}\left[\sigma(z^{(1)}_{z_{i_1};\alpha_1})\sigma(z^{(1)}_{z_{i_1};\alpha_2})\right] = \langle \sigma_{\alpha_1}, \sigma_{\alpha_2}\rangle_{G^{(1)}}$$

This can quite easily be generalised to correlators of more than two activations:

$$\mathbb{E}\left[\sigma\left(z^{(1)}_{i_1;\alpha_1}\right)\sigma\left(z^{(1)}_{i_1;\alpha_2}\right)\sigma\left(z^{(1)}_{i_1;\alpha_3}\right)\sigma\left(z^{(1)}_{i_1;\alpha_4}\right)\right] = \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}} \ , \tag{4.28}$$

and for each pair on two different neurons $i_1 \neq i_2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sigma\left(z^{(1)}_{i_1;\alpha_1}\right)\sigma\left(z^{(1)}_{i_1;\alpha_2}\right)\sigma\left(z^{(1)}_{i_2;\alpha_3}\right)\sigma\left(z^{(1)}_{i_2;\alpha_4}\right)\right] \qquad (4.29)\\
&= \left\{\prod_{i\notin\{i_1,i_2\}} \int \left[\frac{\prod_{\alpha\in\mathcal{D}} dz_{i;\alpha}}{\sqrt{|2\pi G^{(1)}|}}\right] \exp\left(-\frac{1}{2}\sum_{\beta_1,\beta_2\in\mathcal{D}} G^{\beta_1\beta_2}_{(1)} z_{i;\beta_1} z_{i;\beta_2}\right)\right\}\\
&\quad\times \int \left[\frac{\prod_{\alpha\in\mathcal{D}} dz_{i_1;\alpha}}{\sqrt{|2\pi G^{(1)}|}}\right] \exp\left(-\frac{1}{2}\sum_{\beta_1,\beta_2\in\mathcal{D}} G^{\beta_1\beta_2}_{(1)} z_{i_1;\beta_1} z_{i_1;\beta_2}\right) \sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2})\\
&\quad\times \int \left[\frac{\prod_{\alpha\in\mathcal{D}} dz_{i_2;\alpha}}{\sqrt{|2\pi G^{(1)}|}}\right] \exp\left(-\frac{1}{2}\sum_{\beta_1,\beta_2\in\mathcal{D}} G^{\beta_1\beta_2}_{(1)} z_{i_2;\beta_1} z_{i_2;\beta_2}\right) \sigma(z_{i_2;\alpha_3})\,\sigma(z_{i_2;\alpha_4})\\
&= \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}} \langle \sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}} \ ,
\end{aligned}$$

It's clear that each neuron factorises and gives seperate integrals. In deeper layers the preactivation distributions are nearly-Gaussian and things get a bit more complicated.

Second Layer: Genesis of Non-Gaussianity

The joint distribution of the first and second layer preactivations is given:

$$p\left(z^{(2)}, z^{(1)}\middle|\mathcal{D}\right) = p\left(z^{(2)}\middle|z^{(1)}\right) p\left(z^{(1)}\middle|\mathcal{D}\right) . \tag{4.32}$$

We evaluated the last term in the previous section, for the conditional distribution we have:

$$\begin{aligned} & p\left(z^{(2)}\middle|z^{(1)}\right) \\ &= \int \left[\prod_i db_i^{(2)}\; p\left(b_i^{(2)}\right)\right] \left[\prod_{i,j} dW_{ij}^{(2)}\; p\left(W_{ij}^{(2)}\right)\right] \prod_{i,\alpha} \delta\left(z_{i;\alpha}^{(2)} - b_i^{(2)} - \sum_j W_{ij}^{(2)} \sigma_{j;\alpha}^{(1)}\right) , \end{aligned} \tag{4.33}$$

We then marginalise over the first layer preactivations:

$$p\left(z^{(2)}\middle|\mathcal{D}\right) = \int \left[\prod_{i,\alpha} dz_{i;\alpha}^{(1)}\right]\; p\left(z^{(2)}\middle|z^{(1)}\right) p\left(z^{(1)}\middle|\mathcal{D}\right) . \tag{4.34}$$

First we need to figure out how to treat the conditional distribution, then we need to figure out how to integrate out the $z^{(1)}$.

Second layer conditional distribution

If we replace the layer indices $l\; 1 \to 2$ and exchange the network inputs for the first layer preactivations $x_{j;\alpha} \to \sigma_{j;\alpha}^{(1)}$ we can evaluate this in exactly the same way as we evaluated the first-layer distribution:

$$p\left(z^{(2)}\middle|z^{(1)}\right) = \frac{1}{\sqrt{\left|2\pi \widehat{G}^{(2)}\right|^{n_2}}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n_2} \sum_{\alpha_1,\alpha_2 \in \mathcal{D}} \widehat{G}_{(2)}^{\alpha_1\alpha_2} z_{i;\alpha_1}^{(2)} z_{i;\alpha_2}^{(2)}\right) , \tag{4.35}$$

Where the second-layer metric $\hat{G}$ is a random variable that depends on $z^{(1)}$:

$$\widehat{G}_{\alpha_1\alpha_2}^{(2)} \equiv C_b^{(2)} + C_W^{(2)} \frac{1}{n_1} \sum_{j=1}^{n_1} \sigma_{j;\alpha_1}^{(1)} \sigma_{j;\alpha_2}^{(1)} , \tag{4.36}$$

So the second layer conditional distribution is a Gaussian with a variance which is a random variable. This random variable has a mean, and they measure the fluctuation of the second-layer metric by subtracting this mean from the R.V:

$$G^{(2)}_{\alpha_1\alpha_2} \equiv \mathbb{E}\left[\widehat{G}^{(2)}_{\alpha_1\alpha_2}\right] = C_b^{(2)} + C_W^{(2)} \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbb{E}\left[\sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2}\right] \tag{4.37}$$
$$= C_b^{(2)} + C_W^{(2)} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(1)}} ,$$

$$\widehat{\Delta G}^{(2)}_{\alpha_1\alpha_2} \equiv \widehat{G}^{(2)}_{\alpha_1\alpha_2} - G^{(2)}_{\alpha_1\alpha_2} = C_W^{(2)} \frac{1}{n_1}\sum_{j=1}^{n_1} \left(\sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2} - \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(1)}}\right) , \tag{4.38}$$

Which by construction has mean zero when averaged over first layer preactivations.

The variance of the fluctuation is given by its two-point correlator, remembering the expression for the gaussian integrals of the two and four activations on the same neurons from above:

$$\mathbb{E}\left[\widehat{\Delta G}^{(2)}_{\alpha_1\alpha_2}\widehat{\Delta G}^{(2)}_{\alpha_3\alpha_4}\right] \tag{4.40}$$
$$= \left(\frac{C_W^{(2)}}{n_1}\right)^2 \sum_{j,k=1}^{n_1} \mathbb{E}\left[\left(\sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2} - \mathbb{E}\left[\sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2}\right]\right)\left(\sigma^{(1)}_{k;\alpha_3}\sigma^{(1)}_{k;\alpha_4} - \mathbb{E}\left[\sigma^{(1)}_{k;\alpha_3}\sigma^{(1)}_{k;\alpha_4}\right]\right)\right]$$
$$= \left(\frac{C_W^{(2)}}{n_1}\right)^2 \sum_{j=1}^{n_1} \left\{\mathbb{E}\left[\sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2}\sigma^{(1)}_{j;\alpha_3}\sigma^{(1)}_{j;\alpha_4}\right] - \mathbb{E}\left[\sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2}\right]\mathbb{E}\left[\sigma^{(1)}_{j;\alpha_3}\sigma^{(1)}_{j;\alpha_4}\right]\right\}$$
$$= \frac{1}{n_1}\left(C_W^{(2)}\right)^2 \left[\left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\right\rangle_{G^{(1)}} - \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(1)}}\left\langle \sigma_{\alpha_3}\sigma_{\alpha_4}\right\rangle_{G^{(1)}}\right]$$
$$\equiv \frac{1}{n_1} V^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} ,$$

At the end we introduce the four-point vertex $V^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = V(x_{\alpha_1}, x_{\alpha_2}, x_{\alpha_3}, x_{\alpha_4})$ which is depends on four points of input data and is symmetric under exchanges of sample indices: $\alpha_1 \leftrightarrow \alpha_2, \alpha_3 \leftrightarrow \alpha_4, (\alpha_1, \alpha_2) \leftrightarrow (\alpha_3, \alpha_4)$.

Here we also see that as for $n_1 \gg 1$, since V is of order one, the metric fluctuation will become more and more suppressed. We see that the metric fluctuation will become more and more Gaussian due to the central limit theorem. In the limit as n_1 tends to infinity the fluctuation will dissapear.

There are now two ways of integrating out $z^{(1)}$ and obtaining $p(z^{(2)}|\mathcal{D})$, one brute-force involving lots of wick contractions, and one clever:

Wick Derivation Results:

- Expectation of a function under a nearly Gaussian distribution

$$
\begin{aligned}
&\mathbb{E}\left[F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})\right] \qquad (4.46)\\
=&\frac{\int\left[\prod_{i,\alpha}dz_{i;\alpha}\right]e^{-S(z)}F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})}{\int\left[\prod_{i,\alpha}dz_{i;\alpha}\right]e^{-S(z)}}\\
=&\frac{\left\langle\!\left\langle \exp\left\{\frac{1}{8}\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}v^{(\beta_1\beta_2)(\beta_3\beta_4)}\sum_{j_1,j_2=1}^{n}z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}\right\}F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})\right\rangle\!\right\rangle_g}{\left\langle\!\left\langle \exp\left\{\frac{1}{8}\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}v^{(\beta_1\beta_2)(\beta_3\beta_4)}\sum_{j_1,j_2=1}^{n}z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}\right\}\right\rangle\!\right\rangle_g}\\
=&\left\langle\!\left\langle F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})\right\rangle\!\right\rangle_g\\
&+\frac{1}{8}\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}v^{(\beta_1\beta_2)(\beta_3\beta_4)}\sum_{j_1,j_2=1}^{n}\Big[\left\langle\!\left\langle z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})\right\rangle\!\right\rangle_g\\
&\qquad-\left\langle\!\left\langle z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}\right\rangle\!\right\rangle_g\left\langle\!\left\langle F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})\right\rangle\!\right\rangle_g\Big]\\
&+O\left(v^2\right),
\end{aligned}
$$

Clever derivation:

Plug the conditional distribution (4.35) into the marginalization equation (4.34):

$$
p\left(z^{(2)}\middle|\mathcal{D}\right)=\int\left[\prod_{i,\alpha}dz^{(1)}_{i;\alpha}\right]p\left(z^{(1)}\middle|\mathcal{D}\right)\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\widehat{G}^{\alpha_1\alpha_2}_{(2)}z^{(2)}_{j;\alpha_1}z^{(2)}_{j;\alpha_2}\right)}{\sqrt{\left|2\pi\widehat{G}^{(2)}\right|^{n_2}}}\,. \qquad (4.53)
$$

Use the previously discussed decomposition of the stochastic metric into its mean and fluctuating parts:

$$
\widehat{G}^{(2)}_{\alpha_1\alpha_2}=G^{(2)}_{\alpha_1\alpha_2}+\widehat{\Delta G}^{(2)}_{\alpha_1\alpha_2}\,. \qquad (4.54)
$$

To write a Neumann series for the inverse of this:

$$
\begin{aligned}
\widehat{G}^{\alpha_1\alpha_2}_{(2)}=&G^{\alpha_1\alpha_2}_{(2)}-\sum_{\beta_1,\beta_2\in\mathcal{D}}G^{\alpha_1\beta_1}_{(2)}\widehat{\Delta G}^{(2)}_{\beta_1\beta_2}G^{\beta_2\alpha_2}_{(2)} \qquad (4.55)\\
&+\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}G^{\alpha_1\beta_1}_{(2)}\widehat{\Delta G}^{(2)}_{\beta_1\beta_2}G^{\beta_2\beta_3}_{(2)}\widehat{\Delta G}^{(2)}_{\beta_3\beta_4}G^{\beta_4\alpha_2}_{(2)}+O\left(\Delta^3\right).
\end{aligned}
$$

Then put this into the exponential of the marginal distribution in (4.53):

$$
\begin{aligned}
&\exp\left(-\frac{1}{2}\sum_{j=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\widehat{G}_{(2)}^{\alpha_1\alpha_2}z_{j;\alpha_1}^{(2)}z_{j;\alpha_2}^{(2)}\right) \qquad (4.56)\\
=&\exp\left(-\frac{1}{2}\sum_{j=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}G_{(2)}^{\alpha_1\alpha_2}z_{j;\alpha_1}^{(2)}z_{j;\alpha_2}^{(2)}\right)\\
&\times\left\{1+\frac{1}{2}\sum_{i=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\left(\sum_{\beta_1,\beta_2\in\mathcal{D}}G_{(2)}^{\alpha_1\beta_1}\widehat{\Delta G}_{\beta_1\beta_2}^{(2)}G_{(2)}^{\beta_2\alpha_2}\right)z_{i;\alpha_1}^{(2)}z_{i;\alpha_2}^{(2)}\right.\\
&-\frac{1}{2}\sum_{i=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\left(\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}G_{(2)}^{\alpha_1\beta_1}\widehat{\Delta G}_{\beta_1\beta_2}^{(2)}G_{(2)}^{\beta_2\beta_3}\widehat{\Delta G}_{\beta_3\beta_4}^{(2)}G_{(2)}^{\beta_4\alpha_2}\right)z_{i;\alpha_1}^{(2)}z_{i;\alpha_2}^{(2)}\\
&\left.+\frac{1}{2!}\left(\frac{1}{2}\right)^2\sum_{i_1,i_2=1}^{n_2}\sum_{\alpha_1,\ldots,\beta_4\in\mathcal{D}}G_{(2)}^{\alpha_1\beta_1}\cdots G_{(2)}^{\alpha_4\beta_4}\widehat{\Delta G}_{\beta_1\beta_2}^{(2)}\widehat{\Delta G}_{\beta_3\beta_4}^{(2)}z_{i_1;\alpha_1}^{(2)}z_{i_1;\alpha_2}^{(2)}z_{i_2;\alpha_3}^{(2)}z_{i_2;\alpha_4}^{(2)}+O\left(\Delta^3\right)\right\}.
\end{aligned}
$$

Then the denominator becomes:

$$
\begin{aligned}
&\sqrt{\left|2\pi\widehat{G}^{(2)}\right|^{n_2}}=\int\left[\prod_{i,\alpha}dz_{i;\alpha}^{(2)}\right]\exp\left(-\frac{1}{2}\sum_{j=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\widehat{G}_{(2)}^{\alpha_1\alpha_2}z_{j;\alpha_1}^{(2)}z_{j;\alpha_2}^{(2)}\right) \qquad (4.57)\\
=&\sqrt{\left|2\pi G^{(2)}\right|^{n_2}}\left[1+\frac{n_2}{2}\sum_{\beta_1,\beta_2\in\mathcal{D}}\widehat{\Delta G}_{\beta_1\beta_2}^{(2)}G_{(2)}^{\beta_1\beta_2}\right.\\
&\left.+\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}\widehat{\Delta G}_{\beta_1\beta_2}^{(2)}\widehat{\Delta G}_{\beta_3\beta_4}^{(2)}\left(\frac{n_2^2}{8}G_{(2)}^{\beta_1\beta_2}G_{(2)}^{\beta_3\beta_4}-\frac{n_2}{4}G_{(2)}^{\beta_1\beta_3}G_{(2)}^{\beta_2\beta_4}\right)+O\left(\Delta^3\right)\right],
\end{aligned}
$$

The first line is expressing the determinant as a Gaussian integral and the second is substituting in (4.56).

Now we plug these two expressions back into (4.53) to get:

$$
\begin{aligned}
p\left(z^{(2)}\middle|\mathcal{D}\right)=&\frac{1}{\sqrt{\left|2\pi G^{(2)}\right|^{n_2}}}\exp\left(-\frac{1}{2}\sum_{j=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}G_{(2)}^{\alpha_1\alpha_2}z_{j;\alpha_1}^{(2)}z_{j;\alpha_2}^{(2)}\right) \qquad (4.58)\\
&\times\left\{\left[1+O\left(\frac{1}{n_1}\right)\right]+\sum_{i=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\left[O\left(\frac{1}{n_1}\right)\right]z_{i_1;\alpha_1}^{(2)}z_{i_1;\alpha_2}^{(2)}\right.\\
&\left.+\frac{1}{8n_1}\sum_{i_1,i_2=1}^{n_2}\sum_{\alpha_1,\ldots,\alpha_4\in\mathcal{D}}V_{(2)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}z_{i_1;\alpha_1}^{(2)}z_{i_1;\alpha_2}^{(2)}z_{i_2;\alpha_3}^{(2)}z_{i_2;\alpha_4}^{(2)}\right\}+O\left(\frac{1}{n_1^2}\right),
\end{aligned}
$$

(Not sure about this) Where they have used that $\mathbb{E}\left[\hat{\Delta G}_{\beta_1\beta_2}^{(2)}\right]=0$ and $\mathbb{E}\left[\hat{\Delta G}_{\beta_1\beta_2}^{(2)}\hat{\Delta G}_{\beta_3\beta_4}^{(2)}\right]=\frac{1}{n_1}V_{(\beta_1\beta_2)(\beta_3\beta_4)}^{(2)}$. Then taking log:

$$S(z) = \frac{1}{2} \sum_{\alpha_1,\alpha_2 \in \mathcal{D}} \left[G_{(2)}^{\alpha_1 \alpha_2} + O\left(\frac{1}{n_1}\right) \right] \sum_{i=1}^{n_2} z_{i;\alpha_1} z_{i;\alpha_2} \tag{4.60}$$
$$- \frac{1}{8} \sum_{\alpha_1,\ldots,\alpha_4 \in \mathcal{D}} \frac{1}{n_1} V_{(2)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \sum_{i_1,i_2=1}^{n_2} z_{i_1;\alpha_1} z_{i_1;\alpha_2} z_{i_2;\alpha_3} z_{i_2;\alpha_4} + O\left(\frac{1}{n_1^2}\right) .$$

They mention that one might wonder why they drop the $1/n_1$ correction to the quadratic coupling, but keep the quartic coupling despite it being of the same order. They mention that such a correction is a subleading contribution to the two-point correlator, while the quartic coupling gives the leading contribution to the connected four-point correlator. In short, the first is a minor quantitative effect, while for the latter there will be observables whose leading contributions come soley from the quartic coupling.

Nearly-Gaussian action in action

- Using a previously derived equation for the expectation of a function under a nearly-gaussian distribution they show how to obtain the expectations of two activations on the same neuron, and four activations: two pairs on seperate neurons and all four on one neuron.
  - There is a non-trivial effect from the quartic coupling v, and pairs of neurons can only correlate by adding a quartic action, highlighting the importance of finite width for feature learning.
- They give a formula for the covariance between two functions that depend on subsamples of the data:

$$\begin{aligned} &\mathrm{Cov}\Big[\mathcal{F}(z_{i_1;\mathcal{A}_1}),\, \mathcal{G}(z_{i_2;\mathcal{A}_2})\Big] \qquad (4.64)\\ \equiv &\mathbb{E}\Big[\mathcal{F}(z_{i_1;\mathcal{A}_1})\,\mathcal{G}(z_{i_2;\mathcal{A}_2})\Big] - \mathbb{E}\Big[\mathcal{F}(z_{i_1;\mathcal{A}_1})\Big]\,\mathbb{E}\Big[\mathcal{G}(z_{i_2;\mathcal{A}_2})\Big] \\ = &\frac{1}{4} \sum_{\beta_1,\ldots,\beta_4 \in \mathcal{D}} v^{(\beta_1\beta_2)(\beta_3\beta_4)} \Big\langle (z_{\beta_1} z_{\beta_2} - g_{\beta_1\beta_2})\,\mathcal{F}(z_{\mathcal{A}_1}) \Big\rangle_g \Big\langle (z_{\beta_3} z_{\beta_4} - g_{\beta_3\beta_4})\,\mathcal{G}(z_{\mathcal{A}_2}) \Big\rangle_g + O\left(v^2\right) . \end{aligned}$$

Deeper Layers: Accumulation of Non-Gaussianity

By following the same procedure as the one for the second-layer distribution we can proceed to find the marginal distribution of an arbitrary layer. Care must be taken however as the previous layers will no longer be gaussian.

- Recursive strategy: Reconstruct the $(l+1)$-th layer marginal distribution out of the $(l+1)$-th layer preactivation correlators, then use the l-th layer action to evaluate the expectations of the l-th layer activations that occur in the expressions for the $(l+1)$-th layer preactivation correlators.

Recursion

- Using previous work they are able to derive expressions for the two-point correlator and the connected four-point correlator of the $(l+1)$-th layer in terms of the correlators for the l-th layer. These correlators can be used to obtain the $(l+1)$-th layer marginal.
- This can be done efficiently by finding the action of the preactivation distribution.

Action

- The preactivation function can be written in terms of an action $S(z^{(l)})$
- The ansatz for the action:

$$S\left(z^{(\ell)}\right) \equiv \frac{1}{2}\sum_{i=1}^{n_\ell}\sum_{\alpha_1,\alpha_2\in\mathcal{D}} g_{(\ell)}^{\alpha_1\alpha_2} z_{i;\alpha_1}^{(\ell)} z_{i;\alpha_2}^{(\ell)} \tag{4.80}$$
$$- \frac{1}{8}\sum_{i_1,i_2=1}^{n_\ell}\sum_{\alpha_1,\ldots,\alpha_4\in\mathcal{D}} v_{(\ell)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} z_{i_1;\alpha_1}^{(\ell)} z_{i_1;\alpha_2}^{(\ell)} z_{i_2;\alpha_3}^{(\ell)} z_{i_2;\alpha_4}^{(\ell)} + \ldots .$$

- The coefficients are data-dependent couplings, and the results in (4.2) can be generalised to the l-th layer

$$g_{(\ell)}^{\alpha_1\alpha_2} = G_{(\ell)}^{\alpha_1\alpha_2} + O(v,\ldots) , \tag{4.81}$$
$$v_{(\ell)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \frac{1}{n_{\ell-1}} V_{(\ell)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + O\left(v^2,\ldots\right) , \tag{4.82}$$

- The higher order terms $\mathcal{O}(\ldots)$ can only be ignored iff the quartic coupling v and higher order couplings are perturbatively small, which they show next.

Large-width expansion

- The calculations needed for the recursive strategy simplify in the wide regime, with a larger number of neurons per layer. (i.e the regime in which the networks are "practically usable and theoretically tractable").
- To be brief, when in this regime the order of the mean metric G and four-point vertex V is order one at layer l and remains order one at layer $l+1$
- They derive recursion relations for this mean metric and four-point vertex:

$$G_{\alpha_1\alpha_2}^{(\ell+1)} = C_b^{(\ell+1)} + C_W^{(\ell+1)} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right) , \tag{4.86}$$

$$\frac{1}{n_\ell} V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \tag{4.90}$$
$$= \frac{1}{n_\ell}\left(C_W^{(\ell+1)}\right)^2 \left[\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(\ell)}} - \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(\ell)}} \langle \sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(\ell)}}\right]$$
$$+ \frac{1}{n_{\ell-1}} \frac{\left(C_W^{(\ell+1)}\right)^2}{4} \sum_{\beta_1,\ldots,\beta_4 \in \mathcal{D}} V_{(\ell)}^{(\beta_1\beta_2)(\beta_3\beta_4)} \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2} - g_{\beta_1\beta_2}\right)\rangle_{G^{(\ell)}}$$
$$\times \langle \sigma_{\alpha_3}\sigma_{\alpha_4}\left(z_{\beta_3}z_{\beta_4} - g_{\beta_3\beta_4}\right)\rangle_{G^{(\ell)}} + O\left(\frac{1}{n^2}\right) .$$

Importantly, we see that

$$\frac{1}{n_\ell} V^{(\ell+1)} = O\left(\frac{1}{n}\right) , \tag{4.91}$$

- "The additional finite-width corrections given by the higher-order terms in the action can change quantitative results but should not really exhibit qualitative differences." - Is this true?

Marginalisation rules

- In this section they show how to perform marginalisations over a subset of the data or a subset of neurons.
  - Over Data:
    - Allows us to simplify the recursion for the two-point correlator by considering integrals only over the two samples of interest rather than using $\mathcal{N}_\mathcal{D}$ integrals.
  - Over neurons:
    - Can be used to overcome some perturbation scaling issues by integrating over a reduced set of neurons. (See the book)
- The couplings depend on the number of neurons in the action, and they show how how to account for this. The key takeaway is that observables of the l-th layer depend on the number of neurons in that layer.

Subleading Corrections