4

RG Flow of Preactivations

"You can hide a lot in a large-N matrix." – Steve Shenker

John McGreevy [27].

At the end of the last chapter, we computed the statistics of preactivations for deep linear networks at initialization and saw them *run* as a function of the network depth. For that toy model, using a handful of Wick contractions and the recursive structure of the network architecture, we were able to fully understand the effects of the network's hyperparameters – its initialization scheme, width, and depth – on preactivation correlators. This exercise in particular highlighted the importance of *critical initialization hyperparameters* and sufficiently small *depth-to-width ratio* in order for the network outputs to be well-behaved, theoretically and practically. To extend these insights beyond deep linear networks, we need to develop an effective theory of deep learning for networks with any activation function.

While ultimately the goal of our effective theory is to explain how a *particular* neural network learns from a given dataset, our immediate goal in §4 and §5 will be to understand how an *ensemble* of neural networks at initialization behaves as a function of data. In §10, §11, and §∞, we'll find that these goals are closely tied together: through the judicious study of the ensemble, we can systematically evaluate the *typical* behavior of trained networks as well as how any particular network may *fluctuate* away from typicality. Our starting point will thus be a study of the statistics of neural-network preactivations with Gaussian-initialized biases and weights. All in all, the formalism developed in this chapter for analyzing the ensemble of networks at initialization will be the key to a principled understanding of deep learning.

As stressed in the introduction, §0, our focus will always be on describing real finite-width networks, since a lot is lost in idealized infinite-width networks. One salient phenomenon lost in the infinite-width limit is the increasing non-Gaussianity in the preactivation distributions of deeper layers. Such non-Gaussianity makes the behavior of finite-width networks much richer but more complicated to analyze. In order to tame these complications, we'll need to borrow some tools from theoretical physics. In particular, physicists have a long tradition of finding simple descriptions of complicated systems

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

in the limit of a large number of degrees of freedom, while keeping in mind the true goal of modeling real systems. In our context, this hints at tractability and simplification in the regime where networks become very wide, though not infinitely so. To make this precise, in this chapter we introduce the *large-n* expansion or $1/n$ expansion in order to perform perturbative expansions when hidden-layer width n becomes parametrically big. With this tool, we'll be able to systematically study the preactivation distributions of finite neural networks to arbitrary precision.[1]

As we did for deep linear networks, we will proceed recursively, investigating how the distribution of preactivations changes from layer to layer by following the transformation of inputs via the iterative MLP forward-pass equation. We start in §4.1 by computing the distribution of preactivations in the first layer, integrating out the first set of weights and biases. This procedure recovers a well-known result that the distribution of the first-layer preactivations is Gaussian. Since this calculation is so central to the rest of the chapter, we'll present two different derivations: a combinatorial derivation in terms of Wick contractions and an algebraic derivation using the Hubbard–Stratonovich transformation.

Next, in §4.2, we'll consider the distribution of preactivations in the second layer and see the emergence of non-Gaussianity in four-point and higher-point connected correlators. The magnitude of these correlators is suppressed when the network is very wide, vanishing in the strict infinite-width limit. This suppression for wide networks in turn enables us to write down an action describing the preactivation distribution, building on the correspondence explored in §1 between such connected correlators and the couplings in the action. In particular, the large-n expansion lets us start with the quadratic action describing the Gaussian distribution in the infinite-width limit and then perturbatively expand around it in a series of the inverse width, $1/n$, to arbitrary desired precision. Given the importance of this result, we again provide two derivations, one based on Wick contractions and the other based on expanding the stochastic metric.

Finally, in §4.3, we'll analyze the distribution of preactivations at any depth. At this point we can simply repurpose the calculations from the preceding sections to see how the distribution of preactivations recursively transforms from the ℓ-th layer to the $(\ell + 1)$-th layer. In particular, keeping the leading finite-width $1/n$ corrections, we'll obtain recursion equations for the two-point and four-point correlators, encoding how these observables evolve with increasing depth. We'll see that the preactivation distribution of

[1] Back in 1996, Neal introduced the *infinite-width limit* in a seminal work [28], focusing on single-hidden-layer networks. Much later, this program was continued in [29, 30], extending the infinite-width limit to deeper networks, and then was extended further by Yaida in [31] to *finite-width networks.* A large part of this chapter is focused on reproducing the recursions first derived in [31].

However, our perspective here is different than the one taken in this prior work. In particular, our main motivation is in computing the distribution of preactivations at initialization, with an eye toward ultimately understanding gradient-based training (§10, §11, §∞), rather than providing a starting point for Bayesian inference. (We will give our own perspective on Bayesian learning for deep learning in §6.) Additionally, in contrast to [31], our results here are derived by first focusing on the couplings in the *action*, rather than directly on the correlators of the distribution. This method is more intuitive and can be more easily extended.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

the $(\ell+1)$-th layer contains a nearly-Gaussian piece inherited from the ℓ-th layer as well as an additional near-Gaussianity generated in the transition from the ℓ-th to $(\ell+1)$-th layer. In the next chapter, §5, we'll see in detail how the near-Gaussianity accumulates with depth by explicitly solving these recursions and analyzing their solutions, which extends the notion of criticality and emergence of the depth-to-width ratio to networks with general activation functions.

After a short clarifying section on some implications of marginalization (§4.4) and a section on subleading corrections (§4.5), we take a step back in §4.6 in order to draw a parallel between our formalism and the *renormalization group* in theoretical physics. The renormalization group is a powerful recursive method for understanding complicated interacting systems, capturing how the effective interactions between the constituents of a system change when the scale at which they are measured changes from microscopic to macroscopic. Specifically, renormalization marginalizes over the microscopic degrees of freedom in the system to yield an effective *coarse-grained* description at long distances. This is analogous to the way we recursively marginalize over preactivations in previous layers to obtain an effective description of a *representation* at the current layer, in our case capturing how the interactions between neurons change with depth. In both cases the flow of the distributions is created by the marginalization of fine-grained information. Given the complete parallel, we will call our flow *representation group (RG) flow*.

If this sounds like a popular heuristic explanation for what deep neural networks do – transforming fine-grained information at the input level into coarser information at the feature levels and finally into a fully coarse-grained representation at the output level – that's because our formalism makes this heuristic picture of representation coarse-graining concrete.[2] Our formalism will further let us directly probe the effect of the *deep* in deep learning by tracking the change in preactivation distributions as we increase the number of layers. Thus, it is the starting point for an effective theory of deep learning, which we will continue to develop throughout the book.

4.1 First Layer: Good-Old Gaussian

Given a dataset

$$\mathcal{D} = \{x_{i;\alpha}\}_{i=1,\ldots,n_0;\,\alpha=1,\ldots,N_\mathcal{D}} \tag{4.1}$$

containing $N_\mathcal{D}$ inputs of n_0-dimensional vectors, the preactivations in the first layer are given by

$$z^{(1)}_{i;\alpha} \equiv z^{(1)}_i(x_\alpha) = b^{(1)}_i + \sum_{j=1}^{n_0} W^{(1)}_{ij} x_{j;\alpha}\,, \quad \text{for} \quad i = 1,\ldots,n_1. \tag{4.2}$$

[2] There have been many formal and informal comments on the connection between renormalization and deep learning, but the relationship has never before been made precise.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

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,

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

Note how this distribution depends conditionally on the input data, representing the fact that the preactivations are functions of the input.

Now, let us compute the distribution of the first-layer preactivations at initialization. Since this will be so important, we give two derivations, one combinatorial and one algebraic.

Wick This Way: Combinatorial Derivation via Correlators

The first derivation involves direct application of Wick contractions to compute correlators of the first-layer distribution (4.5). Starting with the one-point correlator, simply inserting the definition of the first-layer preactivations (4.2) gives

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

since $\mathbb{E}\left[b_i^{(1)}\right] = \mathbb{E}\left[W_{ij}^{(1)}\right] = 0$. In fact, it's easy to see that all the odd-point correlators of $p\Big(z^{(1)}|\mathcal{D}\Big)$ vanish because there is always an odd number of either biases $b^{(1)}$ or weights $W^{(1)}$ left unpaired under Wick contractions.

Next for the two-point correlator, again inserting the definition (4.2), we see

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

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**

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

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

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

The higher-point correlators can be obtained similarly. For instance, the full four-point correlation can be obtained by inserting the definition (4.2) four times and Wick-contracting the biases and weights, yielding

$$\begin{aligned}
&\mathbb{E}\left[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_2;\alpha_2} z^{(1)}_{i_3;\alpha_3} z^{(1)}_{i_4;\alpha_4}\right] \\
&\quad= \delta_{i_1i_2}\delta_{i_3i_4}G^{(1)}_{\alpha_1\alpha_2}G^{(1)}_{\alpha_3\alpha_4} + \delta_{i_1i_3}\delta_{i_2i_4}G^{(1)}_{\alpha_1\alpha_3}G^{(1)}_{\alpha_2\alpha_4} + \delta_{i_1i_4}\delta_{i_2i_3}G^{(1)}_{\alpha_1\alpha_4}G^{(1)}_{\alpha_2\alpha_3} \\
&\quad= \mathbb{E}\left[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_2;\alpha_2}\right]\mathbb{E}\left[z^{(1)}_{i_3;\alpha_3} z^{(1)}_{i_4;\alpha_4}\right] + \mathbb{E}\left[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_3;\alpha_3}\right]\mathbb{E}\left[z^{(1)}_{i_2;\alpha_2} z^{(1)}_{i_4;\alpha_4}\right] \\
&\qquad+ \mathbb{E}\left[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_4;\alpha_4}\right]\mathbb{E}\left[z^{(1)}_{i_2;\alpha_2} z^{(1)}_{i_3;\alpha_3}\right].
\end{aligned} \tag{4.9}$$

Note that the end result is the same as Wick-contracting $z^{(1)}$'s with the variance given by (4.7). As we recall from §1, this can compactly be summarized by saying that the *connected* four-point correlator vanishes,

$$\mathbb{E}\left[z^{(1)}_{i_1;\alpha_1} z^{(1)}_{i_2;\alpha_2} z^{(1)}_{i_3;\alpha_3} z^{(1)}_{i_4;\alpha_4}\right]\Big|_{\text{connected}} = 0. \tag{4.10}$$

Similar Wick combinatorics show that all the full higher-point correlators can be obtained simply by Wick-contracting $z^{(1)}$'s with the variance given by (4.7), and hence all the connected higher-point correlators vanish. This means that all correlators can be generated from a Gaussian distribution with zero mean and variance (4.7).

Then, in order to write down the first-layer action, all we need is the inverse of this variance, given by a matrix $\delta_{i_1i_2}G^{\alpha_1\alpha_2}_{(1)}$ that satisfies

$$\sum_{j=1}^{n_1}\sum_{\beta\in\mathcal{D}}\left(\delta_{i_1j}G^{\alpha_1\beta}_{(1)}\right)\left(\delta_{ji_2}G^{(1)}_{\beta\alpha_2}\right) = \delta_{i_1i_2}\delta^{\alpha_1}_{\ \alpha_2}, \tag{4.11}$$

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

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

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 $G^{\alpha_1\alpha_2}_{(1)}$ 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

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

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

and the partition function

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

where $\left|2\pi G^{(1)}\right|$ is the determinant of the $N_{\mathcal{D}}$-by-$N_{\mathcal{D}}$ matrix $2\pi G^{(1)}_{\alpha_1\alpha_2}$, and, whenever we write out a determinant involving the metric, it will always be that of the metric and *not* of the inverse metric.[3]

Hubbard–Stratonovich This Way: Algebraic Derivation via Action

Rather than first computing correlators and then backing out the distribution that generates them, we can instead work with the distribution directly. Let's start with the formal expression for the preactivation distribution (2.33) worked out two chapters ago[4]

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

where we have momentarily suppressed the layer superscripts "(1)" because it is distracting. At this point, we could try to eliminate some of the integrals over the model parameters against the constraints imposed by the Dirac delta functions, but it's easy to get confused by the different numbers of model-parameter integrals and delta-function constraints.

To clarify matters, we import a neat trick from theoretical physics called the **Hubbard–Stratonovich transformation**. Specifically, using the following integral representation of the Dirac delta function (2.32)

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

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

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

[3] N.B. compared to the generic quadratic action introduced in (1.66) where the random variable z_μ was a *vector* with a general index μ, here in (4.14) we've subdivided the general index into a pair of indices, $\mu \to (i,\alpha)$, so that the first-layer preactivation $z^{(1)}_{i;\alpha}$ is a *tensor* with a neural index i and a sample index α.

[4] For architectures other than MLPs, the expression inside the Dirac delta function would be different, but we expect much of the following to hold so long as the parameters are sampled from simple distributions.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

Completing the square in the exponential for both the biases b and weights W, we see that the action is quadratic in the model parameters:

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

The biases and weights can then be integrated out, yielding an integral representation for the first-layer distribution $p(z)$ as

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

In essence, we've so far traded the delta-function constraints and the model parameters for the auxiliary Hubbard–Stratonovich variables $\Lambda_i{}^\alpha$, which have quadratic action and a simple linear interaction with the preactivations $z_{i;\alpha}$.

Note that the inverse variance for the Hubbard–Stratonovich variables $\Lambda_i{}^\alpha$ is just the first-layer metric (4.8) we introduced in the Wick-contraction derivation,

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

where by now enough dust has settled that layer superscripts "(1)" have been restored. Once again completing the square, the argument of the exponential becomes

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

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

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

As before, $\left|2\pi G^{(1)}\right|$ represents the determinant of the matrix $2\pi G^{(1)}_{\alpha_1\alpha_2}$. The first-layer distribution is Gaussian with each neuron independent, and correlations between preactivations for different samples are encoded entirely in the metric $G^{(1)}_{\alpha_1\alpha_2}$.[5]

[5] Strictly speaking, the expression (4.23) doesn't make sense for $N_\mathcal{D} > n_0$ as the first-layer metric $G^{(1)}_{\alpha_1\alpha_2}$, given by (4.21), would then be degenerate. Nonetheless, the marginalization rules we'll learn

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

Gaussian Action in Action

Now that we've obtained an action representation for the distribution of the first-layer preactivations in two different ways, let's get a feel for how to compute with it. We'll start by computing the expectation of some quantities that will be needed in §4.2: the expectation of two activations on the same neuron, $\mathbb{E}\left[\sigma\left(z^{(1)}_{i_1;\alpha_1}\right)\sigma\left(z^{(1)}_{i_1;\alpha_2}\right)\right]$, and the expectation of four activations, $\mathbb{E}\left[\sigma\left(z^{(1)}_{i_1;\alpha_1}\right)\sigma\left(z^{(1)}_{i_1;\alpha_2}\right)\sigma\left(z^{(1)}_{i_2;\alpha_3}\right)\sigma\left(z^{(1)}_{i_2;\alpha_4}\right)\right]$, either with all four on the same neuron $i_1 = i_2$ or with each pair on two separate neurons $i_1 \neq i_2$.

Let's start with the two-point correlator of activations. Using the definition of the expectation and inserting the action representation of the distribution (4.23), we get

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

The second equality states that the probability distribution factorizes for each neuron due to the relation $e^{x+y} = e^x e^y$. To go from the second equality to the third, we compute the integrals for the neurons with $i \neq i_1$, which are all trivial, and we also rename the dummy integral variable $z_{i_1;\alpha}$ to z_α. The final equality reintroduces the notation (1.68)

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

to describe a Gaussian expectation with variance g and an arbitrary function $F(z_{\alpha_1},\ldots,z_{\alpha_m})$ over variables with sample indices *only*. In other parts of this book we'll explicitly evaluate this type of Gaussian expectation in various setups for concrete choices

in §4.4 make it clear that this degeneracy is harmless so long as the number of inputs involved in an observable of interest is less than the rank of the metric. In particular, the derivations of the recursion relations in the following sections – which require us to consider only two or four inputs at a time – stay intact if $n_0 \geq 2, 4$, respectively.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

of activation functions, but for the purpose of this chapter we will view computations as complete when they are reduced to such Gaussian expectations without any neural indices. Introducing further the simplifying notation

$$\sigma_\alpha \equiv \sigma(z_\alpha)\,, \tag{4.26}$$

the result of the computation above can be succinctly summarized as

$$\mathbb{E}\left[\sigma\left(z^{(1)}_{i_1;\alpha_1}\right)\sigma\left(z^{(1)}_{i_1;\alpha_2}\right)\right] = \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\,. \tag{4.27}$$

It's easy to generalize this to correlators of more than two activations. For instance, for four activations on the same neuron $i_1 = i_2$, we have by the exact same manipulations

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

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

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

where it's clear each neuron factorizes and gives separate Gaussian integrals. This illustrates the fact that neurons are independent, and thus there is no interaction among different neurons in the first layer. In deeper layers, the preactivation distributions are nearly-Gaussian, and things will be a bit more complicated.

4.2 Second Layer: Genesis of Non-Gaussianity

In this section, we'll move onto evaluating the distribution of preactivations in the second layer of an MLP. The second-layer preactivations are defined via

$$z^{(2)}_{i;\alpha} \equiv z^{(2)}_i(x_\alpha) = b^{(2)}_i + \sum_{j=1}^{n_1} W^{(2)}_{ij}\sigma^{(1)}_{j;\alpha}\,, \quad \text{for} \quad i = 1,\ldots,n_2, \tag{4.30}$$

with the first-layer activations denoted as

$$\sigma^{(1)}_{i;\alpha} \equiv \sigma\left(z^{(1)}_{i;\alpha}\right), \tag{4.31}$$

and the biases $b^{(2)}$ and weights $W^{(2)}$ sampled from Gaussian distributions.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

The joint distribution of preactivations in the first and second layers can be factorized as

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

Here the first-layer marginal distribution $p\left(z^{(1)}\middle|\mathcal{D}\right)$ was evaluated in the last section, §4.1, to be a Gaussian distribution (4.23) with the variance given in terms of the first-layer metric $G^{(1)}_{\alpha_1\alpha_2}$. As for the conditional distribution, we know that it can be expressed as[6]

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

from the formal expression (2.34) for the preactivation distribution conditioned on the activations in the previous layer. The marginal distribution of the second-layer preactivations can then be obtained by **marginalizing over** or **integrating out** the first-layer preactivations as

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

To evaluate this expression for the marginal distribution $p\left(z^{(2)}\middle|\mathcal{D}\right)$, first we'll discuss how to treat the conditional distribution $p\left(z^{(2)}\middle|z^{(1)}\right)$, and then we'll explain how to integrate over the first-layer preactivations $z^{(1)}$ governed by the Gaussian distribution (4.23).

Second-Layer Conditional Distribution

The conditional distribution (4.33) can be evaluated in exactly the same way as we evaluated the first-layer distribution (4.16) conditioned on the inputs, with the simple replacement of the layer indices ℓ as $1 \to 2$ and exchanging the network input for the first-layer activation as $x_{j;\alpha} \to \sigma^{(1)}_{j;\alpha}$. Giving you a moment to flip back to (4.16) to make these substitutions and then remind yourself of the answer (4.23), it's easy to see that this evaluation yields

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

[6] Again, the expression in the Dirac delta function is specific to multilayer perceptron architectures, but this formalism can easily be adapted for other architectures.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

where we have defined the *stochastic* second-layer metric

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

with a hat to emphasize that it is a random variable that depends on the stochastic variable $z^{(1)}$ through $\sigma^{(1)} \equiv \sigma\left(z^{(1)}\right)$. Thus, we see that the second-layer conditional distribution (4.35) is a Gaussian whose variance itself is a random variable. In particular, the stochastic second-layer metric fluctuates around the *mean* second-layer metric

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

where in the last step we recalled the result (4.27) for evaluating the two-point correlator of the first-layer activations on the same neuron.

Around this mean, we define the fluctuation of the second-layer metric as

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

which by construction has mean zero when averaged over the first-layer preactivations,

$$\mathbb{E}\left[\widehat{\Delta G}^{(2)}_{\alpha_1\alpha_2}\right] = 0. \tag{4.39}$$

The typical size of the fluctuations is given by its two-point correlator. Recalling the expressions we derived for Gaussian integrals (4.27) and (4.28) of two and four activations on the same neuron and their factorization property on separate neurons (4.29), we obtain

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

where at the end we introduced the second-layer **four-point vertex** $V^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = V\left(x_{\alpha_1}, x_{\alpha_2}; x_{\alpha_3}, x_{\alpha_4}\right)$, which depends on four input data points and is symmetric under

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

the exchanges of sample indices $\alpha_1 \leftrightarrow \alpha_2$, $\alpha_3 \leftrightarrow \alpha_4$, and $(\alpha_1, \alpha_2) \leftrightarrow (\alpha_3, \alpha_4)$. We will understand the significance of this quantity soon in a future equation, (4.43).

Here, we also see our first hint of simplification in the wide regime $n_1 \gg 1$: since the four-point vertex here is manifestly of order one, we see that the metric fluctuation will be suppressed in that regime. Essentially, as the number of neurons in the first layer grows, the metric fluctuation becomes more and more Gaussian due to the central limit theorem. In the strict limit of infinite n_1, the metric would *self-average*, meaning that the fluctuation would vanish.

Now that we have a feel for the distribution of metric fluctuations, we are only too ready to actually integrate out the first-layer preactivations $z^{(1)}$ and obtain the marginal distribution of the second-layer preactivations $p\Big(z^{(2)}\Big|\mathcal{D}\Big)$. We again provide two derivations, one brute-force and the other clever.

Wick Wick Wick: Combinatorial Derivation

The correlators of the second-layer preactivations can be written nicely in terms of the expectations of the stochastic metric that we just computed. In order to compute the correlators, first we use the fact that the conditional distribution $p\Big(z^{(2)}\Big|z^{(1)}\Big)$ is Gaussian (4.35) to Wick-contract the second-layer preactivations $z^{(2)}$, resulting in expressions involving expectations of the stochastic metric $\widehat{G}^{(2)}_{\alpha_1\alpha_2}$; we then insert expressions for the expectations of the stochastic metric obtained above.

With this in mind, the two-point correlator of the second-layer preactivations is given by

$$\mathbb{E}\left[z^{(2)}_{i_1;\alpha_1} z^{(2)}_{i_2;\alpha_2}\right] = \delta_{i_1i_2}\mathbb{E}\left[\widehat{G}^{(2)}_{\alpha_1\alpha_2}\right] = \delta_{i_1i_2} G^{(2)}_{\alpha_1\alpha_2} = \delta_{i_1i_2}\left(C_b^{(2)} + C_W^{(2)}\left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(1)}}\right), \quad (4.41)$$

where to be clear we first used (4.35) to do the single Wick contraction and then inserted the expression (4.37) for the mean of the stochastic metric.

Similarly, the full four-point function can be evaluated as

$$\begin{aligned}
&\mathbb{E}\left[z^{(2)}_{i_1;\alpha_1} z^{(2)}_{i_2;\alpha_2} z^{(2)}_{i_3;\alpha_3} z^{(2)}_{i_4;\alpha_4}\right] \qquad (4.42)\\
&= \delta_{i_1i_2}\delta_{i_3i_4}\mathbb{E}\left[\widehat{G}^{(2)}_{\alpha_1\alpha_2}\widehat{G}^{(2)}_{\alpha_3\alpha_4}\right] + \delta_{i_1i_3}\delta_{i_2i_4}\mathbb{E}\left[\widehat{G}^{(2)}_{\alpha_1\alpha_3}\widehat{G}^{(2)}_{\alpha_2\alpha_4}\right] + \delta_{i_1i_4}\delta_{i_2i_3}\mathbb{E}\left[\widehat{G}^{(2)}_{\alpha_1\alpha_4}\widehat{G}^{(2)}_{\alpha_2\alpha_3}\right],\\
&= \delta_{i_1i_2}\delta_{i_3i_4} G^{(2)}_{\alpha_1\alpha_2} G^{(2)}_{\alpha_3\alpha_4} + \delta_{i_1i_3}\delta_{i_2i_4} G^{(2)}_{\alpha_1\alpha_3} G^{(2)}_{\alpha_2\alpha_4} + \delta_{i_1i_4}\delta_{i_2i_3} G^{(2)}_{\alpha_1\alpha_4} G^{(2)}_{\alpha_2\alpha_3}\\
&\quad + \frac{1}{n_1}\left[\delta_{i_1i_2}\delta_{i_3i_4} V^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1i_3}\delta_{i_2i_4} V^{(2)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)} + \delta_{i_1i_4}\delta_{i_2i_3} V^{(2)}_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)}\right],
\end{aligned}$$

where in the first line we made three Wick contractions of the four second-layer preactivations $z^{(2)}$'s using the Gaussian distribution (4.35), and then in the second line we recalled (4.39) and (4.40) for the expectations of the stochastic metric $\widehat{G}^{(2)}_{\alpha_1\alpha_2} = G^{(2)}_{\alpha_1\alpha_2} + \widehat{\Delta G}^{(2)}_{\alpha_1\alpha_2}$ over the first-layer preactivations $z^{(1)}$. This means that the *connected* four-point correlator – recall (1.54) – after subtracting the contributions from the two-point correlators of the second-layer preactivations, is given by

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

$$\mathbb{E}\left[z^{(2)}_{i_1;\alpha_1}z^{(2)}_{i_2;\alpha_2}z^{(2)}_{i_3;\alpha_3}z^{(2)}_{i_4;\alpha_4}\right]\Big|_{\text{connected}} \tag{4.43}$$
$$= \frac{1}{n_1}\left[\delta_{i_1i_2}\delta_{i_3i_4}V^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1i_3}\delta_{i_2i_4}V^{(2)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)} + \delta_{i_1i_4}\delta_{i_2i_3}V^{(2)}_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)}\right].$$

Here we see the true importance of the four-point vertex we introduced in (4.40); it gives the connected second-layer four-point correlator and controls the near-Gaussianity of the second-layer preactivation distribution. Thus, we see that this connected correlator is suppressed in the wide regime of $n_1 \gg 1$, suggesting that the preactivation distribution will become more and more Gaussian as the network gets wider and wider. Given this, we see that the second-layer preactivation distribution $p\left(z^{(2)}\middle|\mathcal{D}\right)$ is in general *non-Gaussian* but also simplifies significantly in the large-n_1 regime, becoming Gaussian in the strict $n_1 = \infty$ limit and with the four-point vertex $V^{(2)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)}$ measuring the leading deviation from Gaussianity.

To complete our combinatorial derivation, we need to find an action that generates correlations (4.41) and (4.43). As we know, a quadratic action cannot generate non-Gaussian distributions with nontrivial connected four-point correlators, so we need a different action that's appropriate for a nearly-Gaussian distribution. Intuition from single-variable non-Gaussian integrals in §1.2 suggests that we could perhaps generate the requisite correlations by including a quartic term in the action.

With that in mind, let's start with a quartic action for an $(nN_{\mathcal{D}})$-dimensional random variable z

$$S[z] = \frac{1}{2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}} g^{\alpha_1\alpha_2}\sum_{i=1}^{n} z_{i;\alpha_1}z_{i;\alpha_2}$$
$$-\frac{1}{8}\sum_{\alpha_1,\ldots,\alpha_4\in\mathcal{D}} v^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}\sum_{i_1,i_2=1}^{n} z_{i_1;\alpha_1}z_{i_1;\alpha_2}\,z_{i_2;\alpha_3}z_{i_2;\alpha_4}, \tag{4.44}$$

with undetermined couplings g and v. We will treat the quartic coupling v perturbatively, an assumption that we will justify later by relating the quartic coupling v to the $1/n_1$-suppressed connected four-point correlator. Note that by construction the quartic coupling $v^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ has the same symmetric structure as the four-point vertex $V^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ with respect to the sample indices.[7] Using this action, we can compute the two-point and four-point correlators to the first order in v. Then, by matching with the expressions (4.41) and (4.43) for these quantities, we'll learn how to adjust the couplings g and v to reproduce the right statistics of second-layer preactivations in the wide regime.

Before proceeding further, it is convenient to introduce some notation. In (4.25), we defined $\langle F(z_{\alpha_1},\ldots,z_{\alpha_m})\rangle_g$ for the average of an arbitrary function F over a Gaussian distribution with variance g, where preactivation variables z_α have sample indices *only*. In addition, we here define

[7] The conventional factor of $1/8$ in (4.44) is to account for this symmetry.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

$$
\begin{aligned}
&\langle\langle F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})\rangle\rangle_g \qquad (4.45)\\
&\equiv \int\left[\prod_{i=1}^{n}\frac{\prod_{\alpha\in\mathcal{D}}dz_{i;\alpha}}{\sqrt{|2\pi g|}}\right]\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\sum_{\beta_1,\beta_2\in\mathcal{D}}g^{\beta_1\beta_2}z_{j;\beta_1}z_{j;\beta_2}\right)F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})\,,
\end{aligned}
$$

which now includes neural indices. As we saw while working through (4.27) and (4.29), this type of average factorizes into integrals of the form (4.25) for each neuron.

With this notation in hand, the expectation of an arbitrary function $F(z_{i_1;\alpha_1},\ldots,z_{i_m;\alpha_m})$ against a distribution with the quartic action (4.44) can be rewritten in terms of Gaussian expectations, enabling the perturbative expansion in the coupling v as

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

where in the first line we used the definition of the expectation, in the second line we rewrote the numerator and denominator using the notation (4.45) that we just introduced, and in the third line we expanded the exponential in the coupling v, both in the denominator and numerator. In short, this tells us how to perturbatively express an expectation against the full distribution with the quartic action (4.44) in terms of the leading Gaussian expectation and perturbative corrections; these perturbative contributions nonetheless involve only Gaussian expectations and hence are easy to evaluate.

With this in mind, let's consider some particular choices for F. Starting with the two-point correlator, we get

$$
\begin{aligned}
&\mathbb{E}\left[z_{i_1;\alpha_1}z_{i_2;\alpha_2}\right] \qquad (4.47)\\
&=\delta_{i_1i_2}\left[g_{\alpha_1\alpha_2}+\frac{1}{2}\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}v^{(\beta_1\beta_2)(\beta_3\beta_4)}\left(ng_{\alpha_1\beta_1}g_{\alpha_2\beta_2}g_{\beta_3\beta_4}+2g_{\alpha_1\beta_1}g_{\alpha_2\beta_3}g_{\beta_2\beta_4}\right)\right]+O\!\left(v^2\right).
\end{aligned}
$$

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

Here the variance $g_{\alpha_1\alpha_2}$ is the inverse of the quadratic coupling, with $\sum_\beta g_{\alpha_1\beta} g^{\beta\alpha_2} = \delta_{\alpha_1}^{\ \alpha_2}$. Similarly, we find that the connected four-point correlator evaluates to

$$\begin{aligned} &\mathbb{E}\left[z_{i_1;\alpha_1} z_{i_2;\alpha_2} z_{i_3;\alpha_3} z_{i_4;\alpha_4}\right]\Big|_{\text{connected}} \\ &\equiv \mathbb{E}\left[z_{i_1;\alpha_1} z_{i_2;\alpha_2} z_{i_3;\alpha_3} z_{i_4;\alpha_4}\right] - \mathbb{E}\left[z_{i_1;\alpha_1} z_{i_2;\alpha_2}\right] \mathbb{E}\left[z_{i_3;\alpha_3} z_{i_4;\alpha_4}\right] \\ &\quad - \mathbb{E}\left[z_{i_1;\alpha_1} z_{i_3;\alpha_3}\right] \mathbb{E}\left[z_{i_2;\alpha_2} z_{i_4;\alpha_4}\right] - \mathbb{E}\left[z_{i_1;\alpha_1} z_{i_4;\alpha_4}\right] \mathbb{E}\left[z_{i_2;\alpha_2} z_{i_3;\alpha_3}\right] \\ &= \delta_{i_1 i_2}\delta_{i_3 i_4} \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} v^{(\beta_1\beta_2)(\beta_3\beta_4)} g_{\alpha_1\beta_1} g_{\alpha_2\beta_2} g_{\alpha_3\beta_3} g_{\alpha_4\beta_4} \\ &\quad + \delta_{i_1 i_3}\delta_{i_2 i_4} \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} v^{(\beta_1\beta_3)(\beta_2\beta_4)} g_{\alpha_1\beta_1} g_{\alpha_3\beta_3} g_{\alpha_2\beta_2} g_{\alpha_4\beta_4} \\ &\quad + \delta_{i_1 i_4}\delta_{i_2 i_3} \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} v^{(\beta_1\beta_4)(\beta_2\beta_3)} g_{\alpha_1\beta_1} g_{\alpha_4\beta_4} g_{\alpha_2\beta_2} g_{\alpha_3\beta_3} + O\!\left(v^2\right). \end{aligned} \tag{4.48}$$

Comparing these expressions, (4.47) and (4.48), with correlators in the second layer, (4.41) and (4.43), it's easy to see that setting the couplings as

$$g^{\alpha_1\alpha_2} = G_{(2)}^{\alpha_1\alpha_2} + O\!\left(\frac{1}{n_1}\right), \tag{4.49}$$

$$v^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \frac{1}{n_1} V_{(2)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + O\!\left(\frac{1}{n_1^2}\right) \tag{4.50}$$

reproduces the second-layer preactivation correlators to the leading order in $1/n_1$, with the marginal distribution

$$p\!\left(z^{(2)}\middle|\mathcal{D}\right) = \frac{1}{Z} e^{-S\left(z^{(2)}\right)} \tag{4.51}$$

and quartic action (4.44). Here for convenience we have defined a version of the four-point vertex with indices *raised* by the inverse of the second-layer mean metric

$$V_{(2)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \equiv \sum_{\beta_1,\ldots,\beta_4} G_{(2)}^{\alpha_1\beta_1} G_{(2)}^{\alpha_2\beta_2} G_{(2)}^{\alpha_3\beta_3} G_{(2)}^{\alpha_4\beta_4} V^{(2)}_{(\beta_1\beta_2)(\beta_3\beta_4)}. \tag{4.52}$$

Note that the quartic coupling v is $O(1/n_1)$, justifying our earlier perturbative treatment of the coupling for wide networks. Note also that these couplings – the inverse metric $G_{(2)}^{\alpha_1\alpha_2}$ and the quartic coupling $V_{(2)}^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ – are input-dependent. In particular, the effective strength of interaction between neurons is set by the particular set of inputs to the network.

This completes our first combinatorial derivation of the second-layer preactivation distribution.

Schwinger–Dyson This Way: Algebraic Derivation

Here is a neat way to derive the action for the second-layer preactivation distribution. Plugging the conditional distribution (4.35) into the marginalization equation (4.34), the second-layer marginal distribution becomes

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

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

We saw that the stochastic metric has a natural decomposition into mean and fluctuating parts as

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

Inverting this matrix to the second order in the fluctuation around the mean, we get the inverse stochastic metric[8]

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

Putting this into the exponential that appears in the integrand of the marginal distribution (4.53) and Taylor-expanding in the fluctuation $\widehat{\Delta G}^{(2)}_{\alpha_1\alpha_2}$, we find

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

[8] This together with the defining equation for the metric fluctuation (4.38) are sometimes called the Schwinger–Dyson equations [32, 33], from which this subsubsection takes its title.

Using this expression, the determinant in the denominator becomes

$$\sqrt{\left|2\pi\widehat{G}^{(2)}\right|^{n_2}} = \int\left[\prod_{i,\alpha} dz^{(2)}_{i;\alpha}\right]\exp\left(-\frac{1}{2}\sum_{j=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\widehat{G}^{\alpha_1\alpha_2}_{(2)} z^{(2)}_{j;\alpha_1} z^{(2)}_{j;\alpha_2}\right) \tag{4.57}$$
$$= \sqrt{\left|2\pi G^{(2)}\right|^{n_2}}\Bigg[1+\frac{n_2}{2}\sum_{\beta_1,\beta_2\in\mathcal{D}}\widehat{\Delta G}^{(2)}_{\beta_1\beta_2}G^{\beta_1\beta_2}_{(2)}$$
$$+\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}\widehat{\Delta G}^{(2)}_{\beta_1\beta_2}\widehat{\Delta G}^{(2)}_{\beta_3\beta_4}\left(\frac{n_2^2}{8}G^{\beta_1\beta_2}_{(2)}G^{\beta_3\beta_4}_{(2)}-\frac{n_2}{4}G^{\beta_1\beta_3}_{(2)}G^{\beta_2\beta_4}_{(2)}\right)+O\!\left(\Delta^3\right)\Bigg],$$

where on the first line we re-expressed the determinant as a Gaussian integral, and on the subsequent line we plugged in (4.56) and integrated over the second-layer preactivations $z^{(2)}$.

Next, plugging these two expressions (4.56) and (4.57) back into our expression for the second-layer distribution (4.53), we can now integrate out the first-layer preactivations, giving

$$p\!\left(z^{(2)}\middle|\mathcal{D}\right) = \frac{1}{\sqrt{\left|2\pi G^{(2)}\right|^{n_2}}}\exp\left(-\frac{1}{2}\sum_{j=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}G^{\alpha_1\alpha_2}_{(2)} z^{(2)}_{j;\alpha_1} z^{(2)}_{j;\alpha_2}\right) \tag{4.58}$$
$$\times\Bigg\{\left[1+O\!\left(\frac{1}{n_1}\right)\right]+\sum_{i=1}^{n_2}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\left[O\!\left(\frac{1}{n_1}\right)\right]z^{(2)}_{i_1;\alpha_1}z^{(2)}_{i_1;\alpha_2}$$
$$+\frac{1}{8n_1}\sum_{i_1,i_2=1}^{n_2}\sum_{\alpha_1,\ldots,\alpha_4\in\mathcal{D}}V^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(2)} z^{(2)}_{i_1;\alpha_1}z^{(2)}_{i_1;\alpha_2}z^{(2)}_{i_2;\alpha_3}z^{(2)}_{i_2;\alpha_4}\Bigg\}+O\!\left(\frac{1}{n_1^2}\right),$$

where we have used the fact that expectations of the metric fluctuation are given by $\mathbb{E}\left[\widehat{\Delta G}^{(2)}_{\beta_1\beta_2}\right]=0$ and $\mathbb{E}\left[\widehat{\Delta G}^{(2)}_{\beta_1\beta_2}\widehat{\Delta G}^{(2)}_{\beta_3\beta_4}\right]=\frac{1}{n_1}V^{(2)}_{(\beta_1\beta_2)(\beta_3\beta_4)}$.[9] Taking the logarithm to isolate the action and absorbing the irrelevant constant terms into the partition function,

[9] We tacitly assumed that the expectations of $\widehat{\Delta G}^{m\geq 3}$ are of order $O\!\left(1/n_1^2\right)$ or greater. For instance, you can follow exactly the same steps as in (4.40) and compute

$$\mathbb{E}\left[\widehat{\Delta G}^{(2)}_{\beta_1\beta_2}\widehat{\Delta G}^{(2)}_{\beta_3\beta_4}\widehat{\Delta G}^{(2)}_{\beta_5\beta_6}\right] \tag{4.59}$$
$$=\frac{1}{n_1^2}\left(C_W^{(2)}\right)^3\Big[\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\sigma_{\alpha_5}\sigma_{\alpha_6}\rangle_{G^{(1)}}-\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\sigma_{\alpha_5}\sigma_{\alpha_6}\rangle_{G^{(1)}}$$
$$-\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}}\langle\sigma_{\alpha_5}\sigma_{\alpha_6}\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}-\langle\sigma_{\alpha_5}\sigma_{\alpha_6}\rangle_{G^{(1)}}\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}}$$
$$+2\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}}\langle\sigma_{\alpha_5}\sigma_{\alpha_6}\rangle_{G^{(1)}}\Big].$$

Just as in the middle step of (4.40), here again you've likely noticed that nonzero contributions arise only when all the neural indices coincide. You can further use that same insight to show that $\mathbb{E}\left[\left(\widehat{\Delta G}^{(2)}\right)^m\right]=O\!\left(1/n_1^{m-1}\right)$.

we arrive at the correct expression for the second-layer quartic action to leading order in the first layer width:

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

Here, a prudent reader might wonder about our dropping of the $1/n_1$ correction to the quadratic coupling, while keeping the quartic coupling despite it being of the same order. The main reason for this is 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. Indeed, we shall encounter various observables whose leading contributions stem solely from the nontrivial neuron–neuron interaction induced by the quartic coupling. By contrast, the correction to the quadratic coupling at finite width is just a small quantitative effect. Nevertheless, we will compute this subleading correction in §4.5 for completeness.[10]

Nearly-Gaussian Action in Action

Having completed the two derivations, before moving on to the next section, let's use this opportunity to get a bit more of a feel for how to compute with a nearly-Gaussian distribution. Paralleling what we did with the Gaussian action in the last section, let's evaluate the expectation of two activations on the same neuron and four activations, with all four on the same neuron or pairs on separate neurons. The resulting expressions will enable us to obtain the distributions of the preactivations in deeper layers.

In the following, we are just applying the formula (4.46) for the expectation of a general function. These expressions will be valid for any layer $\ell > 1$. First, for two activations on the same neuron, we find

$$
\begin{aligned}
& \mathbb{E}\left[\sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2})\right] \\
& = \langle \sigma_{\alpha_1}\sigma_{\alpha_2} \rangle_g + \frac{1}{8} \sum_{\beta_1,\ldots,\beta_4 \in \mathcal{D}} v^{(\beta_1\beta_2)(\beta_3\beta_4)} \\
& \quad \times \Big[\left\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \left(z_{\beta_1}z_{\beta_2} - g_{\beta_1\beta_2}\right)\left(z_{\beta_3}z_{\beta_4} - g_{\beta_3\beta_4}\right) \right\rangle_g \\
& \qquad + 2n \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \left(z_{\beta_1}z_{\beta_2} - g_{\beta_1\beta_2}\right) \right\rangle_g g_{\beta_3\beta_4} - 2 \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \right\rangle_g g_{\beta_1\beta_3} g_{\beta_2\beta_4} \Big] + O\!\left(v^2\right),
\end{aligned}
\tag{4.61}
$$

where we assume the reader is by now familiar enough with Gaussian integrals and factorization into separate neurons that we can omit the middle steps. This result

[10] It will also turn out (§5.4) that by fine-tuning the initialization hyperparameters, such subleading corrections are suppressed with depth in comparison to nearly-Gaussian corrections. So, in a sense, this subleading correction to the quadratic coupling can be doubly ignored.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

highlights that the addition of the quartic coupling v has a nontrivial effect even on the two-point correlator of same-neuron activations. We can similarly compute the expectation of four activations on the same neuron, but we'll need only the leading Gaussian contribution, namely

$$\begin{aligned}&\mathbb{E}\left[\sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2})\,\sigma(z_{i_1;\alpha_3})\,\sigma(z_{i_1;\alpha_4})\right]-\mathbb{E}\left[\sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2})\right]\mathbb{E}\left[\sigma(z_{i_1;\alpha_3})\,\sigma(z_{i_1;\alpha_4})\right] \qquad (4.62)\\ &\quad=\left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\right\rangle_g-\left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_g\left\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\right\rangle_g+O(v)\ ,\end{aligned}$$

where we subtracted off the contribution from the two-point correlators as that's what will appear in the next section. Finally, the similar expectation of four activations on two different pairs of neurons $i_1 \neq i_2$ can be evaluated by the application of the formula (4.46) and neuron factorizations in Gaussian expectations, yielding

$$\begin{aligned}&\mathbb{E}\left[\sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2})\,\sigma(z_{i_2;\alpha_3})\,\sigma(z_{i_2;\alpha_4})\right]-\mathbb{E}\left[\sigma(z_{i_1;\alpha_1})\,\sigma(z_{i_1;\alpha_2})\right]\mathbb{E}\left[\sigma(z_{i_2;\alpha_3})\,\sigma(z_{i_2;\alpha_4})\right]\\ &=\frac{1}{8}\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}v^{(\beta_1\beta_2)(\beta_3\beta_4)}\sum_{j_1,j_2=1}^{n} \qquad (4.63)\\ &\quad\times\Big[\left\langle\!\left\langle z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}\sigma_{i_1;\alpha_1}\sigma_{i_1;\alpha_2}\sigma_{i_2;\alpha_3}\sigma_{i_2;\alpha_4}\right\rangle\!\right\rangle_g\\ &\qquad-\left\langle\!\left\langle z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}\sigma_{i_1;\alpha_1}\sigma_{i_1;\alpha_2}\right\rangle\!\right\rangle_g\left\langle\!\left\langle\sigma_{i_2;\alpha_3}\sigma_{i_2;\alpha_4}\right\rangle\!\right\rangle_g\\ &\qquad-\left\langle\!\left\langle z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}\sigma_{i_2;\alpha_3}\sigma_{i_2;\alpha_4}\right\rangle\!\right\rangle_g\left\langle\!\left\langle\sigma_{i_1;\alpha_1}\sigma_{i_1;\alpha_2}\right\rangle\!\right\rangle_g\\ &\qquad+\left\langle\!\left\langle z_{j_1;\beta_1}z_{j_1;\beta_2}\,z_{j_2;\beta_3}z_{j_2;\beta_4}\right\rangle\!\right\rangle_g\left\langle\!\left\langle\sigma_{i_1;\alpha_1}\sigma_{i_1;\alpha_2}\right\rangle\!\right\rangle_g\left\langle\!\left\langle\sigma_{i_2;\alpha_3}\sigma_{i_2;\alpha_4}\right\rangle\!\right\rangle_g\Big]\\ &=\frac{1}{4}\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}}v^{(\beta_1\beta_2)(\beta_3\beta_4)}\left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2}-g_{\beta_1\beta_2}\right)\right\rangle_g\left\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\left(z_{\beta_3}z_{\beta_4}-g_{\beta_3\beta_4}\right)\right\rangle_g+O\!\left(v^2\right),\end{aligned}$$

where we get nonzero contributions only when $j_1 = i_1$ and $j_2 = i_2$ or when $j_1 = i_2$ and $j_2 = i_1$. This shows that pairs of activations can only correlate with the addition of the quartic coupling to the action, hinting at the role of finite width for feature learning. More generally, consider functions $\mathcal{F}(z_{i_1;\mathcal{A}_1})$ and $\mathcal{G}(z_{i_2;\mathcal{A}_2})$ of preactivations that depend on subsamples $\mathcal{A}_1$ and $\mathcal{A}_2 \subset \mathcal{D}$, respectively, where with a slight abuse of notation we put the set dependences into the subscripts. For distinct neurons $i_1 \neq i_2$, the calculation identical to the one just above shows that their covariance is given by

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

This formula will be very useful in the future.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

4.3 Deeper Layers: Accumulation of Non-Gaussianity

The preactivations in the deeper layers are recursively given by

$$z_{i;\alpha}^{(\ell+1)} = b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma_{j;\alpha}^{(\ell)}\,, \quad \text{for} \quad i = 1, \ldots, n_{\ell+1}, \tag{4.65}$$

with the activations in the previous layer abbreviated as

$$\sigma_{i;\alpha}^{(\ell)} \equiv \sigma\left(z_{i;\alpha}^{(\ell)}\right). \tag{4.66}$$

We can obtain the marginal distributions of the preactivations in these deeper layers – including the output distribution $p\left(z^{(L)}\middle|\mathcal{D}\right)$ – by following the procedure that we implemented for the second-layer distribution. The only complication is that the preactivation distribution in the previous layer is no longer Gaussian, as it was for the first layer.

The three key concepts of the derivation are recursion, action, and $1/n$-expansion. Let's walk through them one by one.

Recursion

The idea of recursion is to start with information contained in the marginal distribution $p\left(z^{(\ell)}\middle|\mathcal{D}\right)$ in the ℓ-th layer and obtain the marginal distribution for the $(\ell+1)$-th layer. The change of the marginal preactivation distribution from layer to layer can be captured by first writing out the joint probability distribution of preactivations in adjacent layers ℓ and $\ell+1$,

$$p\left(z^{(\ell+1)}, z^{(\ell)}\middle|\mathcal{D}\right) = p\left(z^{(\ell+1)}\middle|z^{(\ell)}\right) p\left(z^{(\ell)}\middle|\mathcal{D}\right), \tag{4.67}$$

then calculating the conditional probability distribution $p\left(z^{(\ell+1)}\middle|z^{(\ell)}\right)$, and finally marginalizing over the preactivations at the ℓ-th layer as

$$p\left(z^{(\ell+1)}\middle|\mathcal{D}\right) = \int \left[\prod_{i,\alpha} dz_{i;\alpha}^{(\ell)}\right] p\left(z^{(\ell+1)}\middle|z^{(\ell)}\right) p\left(z^{(\ell)}\middle|\mathcal{D}\right). \tag{4.68}$$

In particular, the conditional probability distribution $p\left(z^{(\ell+1)}\middle|z^{(\ell)}\right)$ serves as a **transition matrix**, bridging preactivation distributions in adjacent layers.

The calculation of this conditional distribution proceeds identically to the one we performed for the first layer (4.16) and then repurposed for computing the second-layer conditional distribution (4.33). If you'd like, you can again follow along with §4.1, replacing $z^{(1)}$ by $z^{(\ell+1)}$ and $x_{j;\alpha}$ by $\sigma_{j;\alpha}^{(\ell)}$, and obtain

$$p\left(z^{(\ell+1)}\middle|z^{(\ell)}\right) = \frac{1}{\sqrt{\left|2\pi\widehat{G}^{(\ell+1)}\right|^{n_{\ell+1}}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n_{\ell+1}} \sum_{\alpha_1,\alpha_2\in\mathcal{D}} \widehat{G}_{(\ell+1)}^{\alpha_1\alpha_2} z_{i;\alpha_1}^{(\ell+1)} z_{i;\alpha_2}^{(\ell+1)}\right), \tag{4.69}$$

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

with the $(\ell+1)$-th-layer stochastic metric

$$\widehat{G}^{(\ell+1)}_{\alpha_1\alpha_2} \equiv C_b^{(\ell+1)} + C_W^{(\ell+1)} \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \sigma^{(\ell)}_{j;\alpha_1} \sigma^{(\ell)}_{j;\alpha_2}, \tag{4.70}$$

depending on the random variables $z^{(\ell)}$ in the previous layer ℓ through the activations $\sigma^{(\ell)}$. Note that all the correlators with odd numbers of the $(\ell+1)$-th-layer preactivations vanish while even-point correlators are obtained through Wick contractions, yielding

$$\mathbb{E}\left[z^{(\ell+1)}_{i_1;\alpha_1} \cdots z^{(\ell+1)}_{i_{2m};\alpha_{2m}}\right] = \sum_{\text{all pairings}} \delta_{i_{k_1} i_{k_2}} \cdots \delta_{i_{k_{2m-1}} i_{k_{2m}}} \mathbb{E}\left[\widehat{G}^{(\ell+1)}_{\alpha_{k_1}\alpha_{k_2}} \cdots \widehat{G}^{(\ell+1)}_{\alpha_{k_{2m-1}}\alpha_{k_{2m}}}\right], \tag{4.71}$$

where the sum runs over all the $(2m-1)!!$ parings of auxiliary indices $(k_1, \ldots, k_{2m})$. On the left-hand side, the expectation value characterizes the $(\ell+1)$-th-layer preactivation distribution; on the right-hand side, the expectation value becomes a correlator of ℓ-th-layer activations upon plugging in the stochastic metric (4.70), which can be evaluated with the ℓ-th-layer distribution.

The mean of the stochastic metric is given by

$$G^{(\ell+1)}_{\alpha_1\alpha_2} \equiv \mathbb{E}\left[\widehat{G}^{(\ell+1)}_{\alpha_1\alpha_2}\right] = C_b^{(\ell+1)} + C_W^{(\ell+1)} \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1} \sigma^{(\ell)}_{j;\alpha_2}\right], \tag{4.72}$$

and this mean metric governs the two-point correlator in the $(\ell+1)$-th layer through

$$\mathbb{E}\left[z^{(\ell+1)}_{i_1;\alpha_1} z^{(\ell+1)}_{i_2;\alpha_2}\right] = \delta_{i_1 i_2} \mathbb{E}\left[\widehat{G}^{(\ell+1)}_{\alpha_1\alpha_2}\right] = \delta_{i_1 i_2} G^{(\ell+1)}_{\alpha_1\alpha_2}, \tag{4.73}$$

as we saw for the second layer (4.41) as a special case of (4.71). Meanwhile, the fluctuation around the mean,

$$\widehat{\Delta G}^{(\ell+1)}_{\alpha_1\alpha_2} \equiv \widehat{G}^{(\ell+1)}_{\alpha_1\alpha_2} - G^{(\ell+1)}_{\alpha_1\alpha_2} = C_W^{(\ell+1)} \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \left(\sigma^{(\ell)}_{j;\alpha_1} \sigma^{(\ell)}_{j;\alpha_2} - \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1} \sigma^{(\ell)}_{j;\alpha_2}\right]\right), \tag{4.74}$$

obviously has zero mean,

$$\mathbb{E}\left[\widehat{\Delta G}^{(\ell+1)}_{\alpha_1\alpha_2}\right] = 0, \tag{4.75}$$

and has a magnitude

$$\frac{1}{n_\ell} V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \equiv \mathbb{E}\left[\widehat{\Delta G}^{(\ell+1)}_{\alpha_1\alpha_2} \widehat{\Delta G}^{(\ell+1)}_{\alpha_3\alpha_4}\right] = \mathbb{E}\left[\widehat{G}^{(\ell+1)}_{\alpha_1\alpha_2} \widehat{G}^{(\ell+1)}_{\alpha_3\alpha_4}\right] - G^{(\ell+1)}_{\alpha_1\alpha_2} G^{(\ell+1)}_{\alpha_3\alpha_4}. \tag{4.76}$$

Here we have introduced the $(\ell+1)$-th-layer four-point vertex $V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$, generalizing the second-layer four-point vertex (4.40), which governs the connected four-point correlator in the $(\ell+1)$-th layer. Specifically, following along with the manipulations for the

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

second layer – cf. (4.42) and (4.43) – or simply applying the general expression (4.71), we see

$$\mathbb{E}\left[z^{(\ell+1)}_{i_1;\alpha_1} z^{(\ell+1)}_{i_2;\alpha_2} z^{(\ell+1)}_{i_3;\alpha_3} z^{(\ell+1)}_{i_4;\alpha_4}\right]\Big|_{\text{connected}} \tag{4.77}$$
$$= \frac{1}{n_\ell}\left[\delta_{i_1 i_2}\delta_{i_3 i_4} V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4} V^{(\ell+1)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)} + \delta_{i_1 i_4}\delta_{i_2 i_3} V^{(\ell+1)}_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)}\right].$$

In summary, what we have so far are the expressions for the two-point correlator (4.73) and the connected four-point correlator (4.77) of the $(\ell+1)$-th-layer preactivations in terms of the correlators of the ℓ-th-layer activations, and related expressions for higher-point correlators (4.71) if the need arises. The strategy of our recursive approach is to first evaluate these ℓ-th-layer activation correlators given the ℓ-th-layer distribution $p\left(z^{(\ell)}\middle|\mathcal{D}\right)$ and from them obtain the $(\ell+1)$-th-layer preactivation correlators. Using these correlators, we can then reconstruct the $(\ell+1)$-th-layer marginal distribution $p\left(z^{(\ell+1)}\middle|\mathcal{D}\right)$. Both the evaluation of the ℓ-th-layer activation correlators and the reconstruction of the distribution at the $(\ell+1)$-th layer can be efficiently implemented through the use of the action.

Action

The preactivation distribution $p\left(z^{(\ell)}\middle|\mathcal{D}\right)$ can be written in terms of an action as

$$p\left(z^{(\ell)}\middle|\mathcal{D}\right) = \frac{e^{-S\left(z^{(\ell)}\right)}}{Z(\ell)}, \tag{4.78}$$

with the ℓ-th-layer partition function given by

$$Z(\ell) \equiv \int \left[\prod_{i,\alpha} dz^{(\ell)}_{i;\alpha}\right] e^{-S\left(z^{(\ell)}\right)} \tag{4.79}$$

and our ansatz for the action given by the following expansion:

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

This ansatz encompasses both the actions we had in §4.1 for the first-layer preactivations – with $g^{\alpha_1\alpha_2}_{(1)} = G^{\alpha_1\alpha_2}_{(1)}$ and $v_{(1)} = 0$ – and for the second-layer preactivations in §4.2 – with the couplings $g_{(2)}$ and $v_{(2)}$ given by (4.49) and (4.50), respectively. In fact, this represents the most general expansion around the Gaussian action, given the symmetries of preactivation correlators (4.71). In particular, only even powers of preactivations show up in the action since we know that correlators with odd numbers of preactivations vanish.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

Here, the coefficients $g^{\alpha_1\alpha_2}_{(\ell)}$, $v^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)}$, and the implied additional terms in the expansion, are **data-dependent couplings** that together govern the interactions of the neural preactivations and are simply related to the correlators of preactivations $z^{(\ell)}$. In particular, in §4.2 we gave two derivations for the relations between quadratic and quartic couplings on the one hand and two-point and four-point correlators on the other hand. The same argument applies for an arbitrary layer ℓ, and so we have

$$g^{\alpha_1\alpha_2}_{(\ell)} = G^{\alpha_1\alpha_2}_{(\ell)} + O(v, \dots) \,, \tag{4.81}$$

$$v^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)} = \frac{1}{n_{\ell-1}} V^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)} + O\Big(v^2, \dots\Big) \,, \tag{4.82}$$

with the understanding that the raised indices of the four-point vertex are shorthand for contraction with the ℓ-th-layer inverse metric:

$$V^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)} \equiv \sum_{\beta_1,\dots,\beta_4\in\mathcal{D}} G^{\alpha_1\beta_1}_{(\ell)} G^{\alpha_2\beta_2}_{(\ell)} G^{\alpha_3\beta_3}_{(\ell)} G^{\alpha_4\beta_4}_{(\ell)} V^{(\ell)}_{(\beta_1\beta_2)(\beta_3\beta_4)} . \tag{4.83}$$

Note that the higher-order terms $O(\dots)$ in (4.81) and (4.82) can be neglected self-consistently if and only if the quartic coupling v and higher-order couplings are perturbatively small. This is indeed the case when networks are sufficiently wide, as we will show next.

Large-Width Expansion

Now we have our work cut out for us. First, note that these mappings, (4.81) and (4.82), between the correlators and couplings already accomplish one task mentioned in our recursive strategy. Namely, when applied to the $(\ell+1)$-th layer, they reconstruct the $(\ell+1)$-th-layer distribution out of the $(\ell+1)$-th-layer preactivation correlators. The only remaining task then is to use the ℓ-th-layer action (4.80) to compute the expectations of the ℓ-th-layer activations $\sigma^{(\ell)}$ that appear in the expressions for the two-point correlator (4.73) and four-point correlator (4.77) of the $(\ell+1)$-th-layer preactivations $z^{(\ell+1)}$.

These calculations simplify in the wide regime with a large number of neurons per layer

$$n_1, n_2, \dots, n_{L-1} \sim n \gg 1. \tag{4.84}$$

As has been advertised, this large-but-finite-width regime is where networks become both practically usable and theoretically tractable. Specifically, the relations (4.81) and (4.82) between correlators and couplings simplify in this regime, and higher-order non-Gaussian corrections can be self-consistently truncated in a series in $1/n$.[11] To be precise, we inductively assume that the mean metric $G^{(\ell)} = O(1)$ and the four-point vertex $V^{(\ell)} = O(1)$ are both of order one at the ℓ-th layer – as was the case for the first and second layers – and show that the same holds true at the $(\ell+1)$-th layer. This inductive

[11] In the language of §4.6, such a truncation is preserved under the RG flow.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

assumption in particular implies through (4.81) and (4.82) that the quartic coupling $v_{(\ell)} = O(1/n)$ is perturbatively small at the ℓ-th layer and that the quadratic coupling is given by $g_{(\ell)} = G_{(\ell)} + O(1/n)$. In carrying out this inductive proof, we obtain the recursion relations that govern the change in the preactivation distributions from the ℓ-th layer to the $(\ell+1)$-th layer.

To begin, we see that the two-point correlator in the $(\ell+1)$-th layer (4.73) is given simply in terms of the metric

$$G^{(\ell+1)}_{\alpha_1\alpha_2} = C^{(\ell+1)}_b + C^{(\ell+1)}_W \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right] . \tag{4.85}$$

With foresight, we already evaluated this particular two-point correlator of activations (4.61) in the last section. Inserting this result, along with the quadratic coupling $g_{(\ell)} = G_{(\ell)} + O(1/n)$ and quartic coupling $v_{(\ell)} = O(1/n)$, we find

$$G^{(\ell+1)}_{\alpha_1\alpha_2} = C^{(\ell+1)}_b + C^{(\ell+1)}_W \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}} + O\!\left(\frac{1}{n}\right) , \tag{4.86}$$

which is the leading recursion for the two-point correlator of preactivations.[12] We see that this is self-consistent; any metric $G^{(\ell)}$ that is of order one will give an order-one metric $G^{(\ell+1)}$ in the next layer as well. The correction is suppressed by $O(1/n)$, which affects only the subleading term in the quadratic coupling $g_{(\ell+1)} = G_{(\ell+1)} + O(1/n)$. Note that, neglecting the subleading $1/n$ correction and replacing G by g, the recursion (4.86) for the two-point correlator can also be thought of as the leading recursion for the quadratic coupling.

Next, let's evaluate the four-point correlator (4.77), which involves computing the magnitude of the metric fluctuation (4.76). Substituting in our general expression for the $(\ell+1)$-th-layer metric fluctuation (4.74), we get

$$\begin{aligned} &\frac{1}{n_\ell} V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \\ &= \left(\frac{C^{(\ell+1)}_W}{n_\ell}\right)^2 \sum_{j,k=1}^{n_\ell} \left\{\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right] - \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right]\mathbb{E}\left[\sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right]\right\} . \end{aligned} \tag{4.87}$$

Here, there are two types of contributions: from coincident neurons and from separate pairs of neurons. Again, with foresight, we have already evaluated both types of four-point activation correlators in the last section. When all four are coincident, $j = k$, substituting in (4.62) we find

$$\begin{aligned} &\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\sigma^{(\ell)}_{j;\alpha_3}\sigma^{(\ell)}_{j;\alpha_4}\right] - \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right]\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_3}\sigma^{(\ell)}_{j;\alpha_4}\right] \\ &\quad = \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\right\rangle_{G^{(\ell)}} - \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}}\left\langle \sigma_{\alpha_3}\sigma_{\alpha_4}\right\rangle_{G^{(\ell)}} + O\!\left(\frac{1}{n}\right) , \end{aligned} \tag{4.88}$$

[12] Note that the difference from the second-layer calculation in §4.2 is just that the expectation in (4.85) is not exactly Gaussian but has a $1/n$ correction. This highlights the main difference with that section, which is that the distribution in the prior layer is nearly-Gaussian.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

where we have truncated to leading order in $1/n$ as a consequence of the inductive assumption at the ℓ-th layer. Meanwhile, when $j \neq k$ and the correlation is between two neurons, we substitute in our expression (4.63), finding

$$\begin{aligned}
&\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right] - \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right]\mathbb{E}\left[\sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right] \qquad (4.89)\\
&= \frac{1}{4n_{\ell-1}} \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} V^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2} - g_{\beta_1\beta_2}\right)\right\rangle_{G^{(\ell)}} \left\langle \sigma_{\alpha_3}\sigma_{\alpha_4}\left(z_{\beta_3}z_{\beta_4} - g_{\beta_3\beta_4}\right)\right\rangle_{G^{(\ell)}}\\
&\quad + O\left(\frac{1}{n^2}\right),
\end{aligned}$$

where again we have truncated to leading order in the large-width expansion using the inductive assumption.[13] Inserting both of these expressions back into (4.87) and performing the sums, we get a recursion for the four-point vertex

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

Importantly, we see that

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

and $V^{(\ell+1)} = O(1)$, thus completing our inductive proof and concluding our derivations of the recursion relations (4.86) and (4.90) for the two-point and four-point correlators. As was the case for the quadratic coupling, if we neglect the subleading $1/n^2$ correction and replace G by g and V by v, the recursion (4.90) for the connected four-point correlator can also be thought of as the recursion for the quartic coupling.

Note that in the strict $n \to \infty$ limit, the quartic coupling vanishes, and the marginal distribution of preactivations $p\left(z^{(\ell)}\middle|\mathcal{D}\right)$ is Gaussian for all layers ℓ. The first nontrivial correction to this infinite-width limit is captured by studying the quartic action with couplings $v_{(\ell)}$. In what follows, we will mostly focus on the effective theory with this quartic action, as we expect significant qualitative differences in the behavior of networks described by the quadratic action vs. the quartic action. 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.

[13] Again, the difference from the second-layer calculation is that in §4.2 these expectations are over the exactly Gaussian first-layer distribution. In that case, there was a contribution of the form (4.88) from the case with all neurons coincident but *not* of the form (4.89) from the two neurons – cf. (4.40).

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

4.4 Marginalization Rules

In the past sections, at each step in the recursions we marginalized over all the preactivations in a given layer. This section collects two remarks on other sorts of *partial* marginalizations we can perform, rather than integrating out an *entire* layer. In particular, we'll discuss marginalization over a subset of the $N_{\mathcal{D}}$ samples in the dataset $\mathcal{D}$ and marginalization over a subset of neurons in a layer.

Loosely speaking, these marginalizations let us focus on specific input data and neurons of interest. Tightly speaking, let's consider evaluating the expectation of a function $F(z_{I;\mathcal{A}}) = F\Big(\{z_{i;\alpha}\}_{i\in I;\alpha\in\mathcal{A}}\Big)$ that depends on a subsample $\mathcal{A}\subset\mathcal{D}$ and a subset of neurons $I\subset\{1,\ldots,n_\ell\}\equiv\mathcal{N}$ in a layer ℓ, where with a slight abuse of notation we put the set dependences into the subscripts. We then have

$$\begin{aligned}&\mathbb{E}\left[F(z_{I;\mathcal{A}})\right] \qquad (4.92)\\ &=\int\left[\prod_{i\in\mathcal{N}}\prod_{\alpha\in\mathcal{D}}dz_{i;\alpha}\right]F(z_{I;\mathcal{A}})\;p\Big(z_{\mathcal{N};\mathcal{D}}\Big|\mathcal{D}\Big)\\ &=\int\left[\prod_{i\in I}\prod_{\alpha\in\mathcal{A}}dz_{i;\alpha}\right]F(z_{I;\mathcal{A}})\left\{\int\left[\prod_{(j;\beta)\in[\mathcal{N}\times\mathcal{D}-I\times\mathcal{A}]}dz_{j;\beta}\right]p\Big(z_{\mathcal{N};\mathcal{D}}\Big|\mathcal{D}\Big)\right\}\\ &=\int\left[\prod_{i\in I}\prod_{\alpha\in\mathcal{A}}dz_{i;\alpha}\right]F(z_{I;\mathcal{A}})\;p\Big(z_{I;\mathcal{A}}\Big|\mathcal{A}\Big)\,,\end{aligned}$$

where the last equality is just the marginalization over the spectator variables that do not enter into the observable of interest and, in a sense, defines the subsampled and subneuroned distribution as

$$p\Big(z_{I;\mathcal{A}}\Big|\mathcal{A}\Big)\equiv\int\left[\prod_{(j;\beta)\in[\mathcal{N}\times\mathcal{D}-I\times\mathcal{A}]}dz_{j;\beta}\right]p\Big(z_{\mathcal{N};\mathcal{D}}\Big|\mathcal{D}\Big)\,. \qquad (4.93)$$

In words, in evaluating the expectation of the function $F(z_{I;\mathcal{A}})$, the full distribution $p(z_{\mathcal{N};\mathcal{D}}|\mathcal{D})$ can simply be restricted to that of the subsample $\mathcal{A}$ and subneurons I, i.e., $p(z_{I;\mathcal{A}}|\mathcal{A})$. We call this property a **marginalization rule**. Yes, this is somewhat trivial – we're just restating the consistency of probability distributions with respect to marginalization – but it has two rather useful consequences for us.

Marginalization over Samples

The first corollary of the marginalization rule is that we can use it to reduce a gigantic integral over all the samples in the dataset to a compact integral over only a handful of samples. For example, in recursively obtaining the two-point correlator through

$$\mathbb{E}\left[z^{(\ell+1)}_{i_1;\alpha_1}z^{(\ell+1)}_{i_2;\alpha_2}\right]=\delta_{i_1i_2}\left[C^{(\ell+1)}_b+C^{(\ell+1)}_W\left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}}+O\!\left(\frac{1}{n}\right)\right], \qquad (4.94)$$

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

we can reduce the $N_\mathcal{D}$-dimensional Gaussian integrals $\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(\ell)}}$ with the $N_\mathcal{D}$-by-$N_\mathcal{D}$ variance matrix $G^{(\ell)}$ to a manageable two-dimensional integral with a two-by-two submatrix spanned by α_1 and α_2 (or a one-dimensional integral if $\alpha_1 = \alpha_2$). Similarly, a Gaussian integral for four activations $\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(\ell)}}$ that appears in the recursion for four-point vertex involves integrals over four variables *at most*. Generally, in using the action (4.80) to evaluate a specific expectation, the summation over the whole dataset $\mathcal{D}$ in the action can be restricted to the subset of input data that actually appears in the expectation. By the same token, in recursively evaluating the four-point vertex $V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ via the recursion (4.90), the summation on the right-hand side over the dataset $\mathcal{D}$ can be restricted to the set of samples being correlated, $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$. However, please keep in mind that the inverse metrics used to construct $V^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)}$ in (4.83) must then be taken to be the inverse of the metric submatrix on this restricted subspace.[14]

Marginalization over Neurons

The second corollary involves integrating out a subset of neurons in a layer. Prudent readers might have worried that the quartic term in the ℓ-th-layer action,

$$-\frac{1}{8}\sum_{i_1,i_2=1}^{n_\ell}\sum_{\alpha_1,\ldots,\alpha_4\in\mathcal{D}} v^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)}\, z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_1;\alpha_2}\, z^{(\ell)}_{i_2;\alpha_3} z^{(\ell)}_{i_2;\alpha_4}, \tag{4.95}$$

seems naively to scale like $\sim n_\ell^2/n_{\ell-1} = O(n)$, since there are two sums over n_ℓ, and we know from (4.82) that the coupling $v_{(\ell)}$ scales like $\sim 1/n_{\ell-1}$. Similarly, the quadratic term,

$$\frac{1}{2}\sum_{i=1}^{n_\ell}\sum_{\alpha_1,\alpha_2\in\mathcal{D}} g^{\alpha_1\alpha_2}_{(\ell)}\, z^{(\ell)}_{i;\alpha_1} z^{(\ell)}_{i;\alpha_2}, \tag{4.96}$$

has a single sum over n_ℓ and so seems naively $O(n)$ as well. This would imply that the quartic term isn't perturbatively suppressed in comparison to the quadratic term, naively calling our perturbative approach into question.

We first observe that this problem never arises for the final layer $\ell = L$, since the output dimension n_L is never parametrically large: the quadratic term scales as $\sim n_L = O(1)$ while the quartic term scales as $\sim n_L^2/n_{L-1} = O(1/n)$, which is perturbatively suppressed.

This observation, combined with the marginalization rule, points at a resolution to the naive scale-counting problem above for the hidden layers. Indeed, all the expectations we have evaluated so far – both preactivation and activation correlators – each individually involves only a few neurons m_ℓ in any given layer ℓ, with $m_\ell \ll n_\ell$. This will always be true; we can't actually correlate an infinite number of neurons at once! Thus, when

[14] A similar restriction of the summation can be applied to any of our other recursions and will prove especially useful when you try to evaluate them numerically or analytically.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

using the action representation (4.80) of the probability distribution to compute these correlators at the ℓ-th layer, we can first use the marginalization rule (4.92) to integrate out the $(n_\ell - m_\ell)$ spectator neurons that do not participate in the computation, letting us focus on those m_ℓ relevant neurons that actually appear in the expectation. This in turn lets us replace the summations over n_ℓ neurons by ones over the m_ℓ neurons.[15]

All the while, the numbers of neurons in the previous layers, $n_1, \dots, n_{\ell-1}$, having been integrated out to get the action representation at the ℓ-th layer, *are* parametrically large. This means that the quadratic term in the ℓ-th-layer action, reduced to the m_ℓ relevant neurons, scales as $\sim m_\ell = O(1)$, while the quartic term scales as $\sim m_\ell^2/n_{\ell-1} = O(1/n)$. Thus, this ensures a perturbative treatment of the non-Gaussianity.

Running Couplings with Partial Marginalizations

In focusing our attention on only a subset of samples or neurons, the data-dependent couplings of the action need to be adjusted. Since this running of the couplings is instructive and will be necessary for later computations, let us illustrate here how the quadratic coupling $g^{\alpha_1\alpha_2}_{(\ell),m_\ell}$ depends on the number of neurons m_ℓ in the action.

For simplicity in our illustration, let us specialize to a single input x and drop all the sample indices. Then, denote the distribution over m_ℓ neurons as

$$\begin{aligned} p\Big(z^{(\ell)}_1, \dots, z^{(\ell)}_{m_\ell}\Big) &\propto e^{-S\left(z^{(\ell)}_1, \dots, z^{(\ell)}_{m_\ell}\right)} \\ &= \exp\left[-\frac{g_{(\ell),m_\ell}}{2}\sum_{j=1}^{m_\ell} z^{(\ell)}_j z^{(\ell)}_j + \frac{v_{(\ell)}}{8}\sum_{j_1,j_2=1}^{m_\ell} z^{(\ell)}_{j_1} z^{(\ell)}_{j_1} z^{(\ell)}_{j_2} z^{(\ell)}_{j_2}\right], \end{aligned} \tag{4.97}$$

which is expressed by the same action we've already been using (4.80), though now the dependence of the quadratic coupling on m_ℓ is made explicit.[16] We'll now see in two ways how the quadratic coupling $g_{(\ell),m_\ell}$ *runs* with m_ℓ.

The first way is to begin with the action for n_ℓ neurons and formally integrate out $(n_\ell - m_\ell)$ neurons. Without loss of generality, let's integrate out the *last* $(n_\ell - m_\ell)$ neurons, leaving the *first* m_ℓ neurons labeled as $1, \dots, m_\ell$. Using the marginalization rule (4.93), we see that

$$\begin{aligned} e^{-S\left(z^{(\ell)}_1, \dots, z^{(\ell)}_{m_\ell}\right)} &\propto p\Big(z^{(\ell)}_1, \dots, z^{(\ell)}_{m_\ell}\Big) = \int dz^{(\ell)}_{m_\ell+1} \cdots dz^{(\ell)}_{n_\ell}\; p\Big(z^{(\ell)}_1, \dots, z^{(\ell)}_{n_\ell}\Big) \\ &\propto \int dz^{(\ell)}_{m_\ell+1} \cdots dz^{(\ell)}_{n_\ell} \exp\left[-\frac{g_{(\ell),n_\ell}}{2}\sum_{i=1}^{n_\ell} z^{(\ell)}_i z^{(\ell)}_i + \frac{v_{(\ell)}}{8}\sum_{i_1,i_2=1}^{n_\ell} z^{(\ell)}_{i_1} z^{(\ell)}_{i_1} z^{(\ell)}_{i_2} z^{(\ell)}_{i_2}\right], \end{aligned} \tag{4.98}$$

[15] In evaluating generic expectation values such as (4.46), one can always check that the contributions from the $(n_\ell - m_\ell)$ spectator neurons consistently cancel out at each order in $1/n_{\ell-1}$ expansion. If you go back to your personal note that fills in the small gaps between lines in our computations, you will surely notice this cancellation due to Gaussian factorization.

[16] Note that in principle the quartic coupling should also depend on m_ℓ: $v_{(\ell)} \to v_{(\ell),m_\ell}$. However, since such a dependence only shows up at higher order in v, we will suppress it.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

throughout which we neglected normalization factors that are irrelevant if we're just interested in the running of the coupling. Next, we can separate out the dependence on the m_ℓ neurons, perturbatively expand the integrand in quartic coupling, and finally integrate out the last $(n_\ell - m_\ell)$ neurons by computing a few simple Gaussian integrals:

$$
\begin{aligned}
&p\Big(z_1^{(\ell)}, \dots, z_{m_\ell}^{(\ell)}\Big) \qquad (4.99)\\
&\propto \exp\left[-\frac{g_{(\ell),n_\ell}}{2}\sum_{j=1}^{m_\ell} z_j^{(\ell)} z_j^{(\ell)} + \frac{v_{(\ell)}}{8}\sum_{j_1,j_2=1}^{m_\ell} z_{j_1}^{(\ell)} z_{j_1}^{(\ell)} z_{j_2}^{(\ell)} z_{j_2}^{(\ell)}\right]\\
&\quad \times \int dz_{m_\ell+1}^{(\ell)} \cdots dz_{n_\ell}^{(\ell)} \exp\left[-\frac{g_{(\ell),n_\ell}}{2}\sum_{k=m_\ell+1}^{n_\ell} z_k^{(\ell)} z_k^{(\ell)}\right]\\
&\quad \times \left[1 + \frac{2v_{(\ell)}}{8}\sum_{j=1}^{m_\ell}\sum_{k=m_\ell+1}^{n_\ell} z_j^{(\ell)} z_j^{(\ell)} z_k^{(\ell)} z_k^{(\ell)} + \frac{v_{(\ell)}}{8}\sum_{k_1,k_2=m_\ell+1}^{n_\ell} z_{k_1}^{(\ell)} z_{k_1}^{(\ell)} z_{k_2}^{(\ell)} z_{k_2}^{(\ell)} + O\Big(v^2\Big)\right]\\
&= \exp\left[-\frac{g_{(\ell),n_\ell}}{2}\sum_{j=1}^{m_\ell} z_j^{(\ell)} z_j^{(\ell)} + \frac{v_{(\ell)}}{8}\sum_{j_1,j_2=1}^{m_\ell} z_{j_1}^{(\ell)} z_{j_1}^{(\ell)} z_{j_2}^{(\ell)} z_{j_2}^{(\ell)}\right]\\
&\quad \times \left\{1 + \frac{(n_\ell - m_\ell)}{4}\frac{v_{(\ell)}}{g_{(\ell),n_\ell}}\left(\sum_{i=1}^{m_\ell} z_i^{(\ell)} z_i^{(\ell)}\right) + \frac{v_{(\ell)}}{8g_{(\ell),n_\ell}^2}\Big[(n_\ell - m_\ell)^2 + 2(n_\ell - m_\ell)\Big] + O\Big(v^2\Big)\right\}.
\end{aligned}
$$

Finally, resumming the correction arising from the quartic coupling proportional to $\sum_{i=1}^{m_\ell} z_i^{(\ell)} z_i^{(\ell)}$ back into the exponential, ignoring the proportionality factor, and comparing with the action for m_ℓ neurons (4.97), we find

$$
g_{(\ell),m_\ell} = g_{(\ell),n_\ell} - \frac{(n_\ell - m_\ell)}{2}\frac{v_{(\ell)}}{g_{(\ell),n_\ell}} + O\Big(v^2\Big) \qquad (4.100)
$$

as the running equation for the quadratic coupling.

The second way to see the coupling run – and find a solution to the running equation (4.100) – is to compute the single-input metric $G^{(\ell)} \equiv \mathbb{E}\left[z_i^{(\ell)} z_i^{(\ell)}\right]$ and compute it directly using the m_ℓ-neuron action (4.97). We've already computed this in (4.47) using the quartic action for multiple inputs. Specializing to a single input, considering an action of m_ℓ neurons, and being explicit about the dependence of the quadratic coupling on the number of neurons, we get

$$
G^{(\ell)} = \left[\frac{1}{g_{(\ell),m_\ell}} + \frac{(m_\ell + 2)}{2}\frac{v^{(\ell)}}{g_{(\ell),m_\ell}^3}\right] + O\Big(v^2\Big)\,. \qquad (4.101)
$$

Solving this equation for $g_{(\ell),m_\ell}$ by perturbatively expanding in $v^{(\ell)}$, we find

$$
\frac{1}{g_{(\ell),m_\ell}} = G^{(\ell)} - \frac{(m_\ell + 2)}{2}\frac{V^{(\ell)}}{n_{\ell-1}G^{(\ell)}} + O\left(\frac{1}{n^2}\right), \qquad (4.102)
$$

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

where we have also plugged in

$$v_{(\ell)} = \frac{V^{(\ell)}}{n_{\ell-1}\left(G^{(\ell)}\right)^4} + O\left(\frac{1}{n^2}\right), \tag{4.103}$$

using (4.82) and (4.83) to relate the quartic coupling to the four-point vertex and again specializing to a single input. Now, it's easy to check that this expression (4.102) solves the running equation (4.100).[17]

The key step in this alternative derivation is realizing that observables without any neural indices, such as $G^{(\ell)}$, should *not* depend on which version of the m_ℓ action we use in computing them. Interpreted another way, what this running of the coupling means is that for different numbers of neurons in a layer ℓ – e.g., m_ℓ and n_ℓ – we need different quadratic couplings – in this case $g_{(\ell),m_\ell}$ and $g_{(\ell),n_\ell}$ – in order to give the correct value for an ℓ-th-layer observable such as $G^{(\ell)}$. If you're ever in doubt, it's always safest to express an observable of interest in terms of the metric $G^{(\ell)}$ and the four-point vertex $V^{(\ell)}$ rather than the couplings.

4.5 Subleading Corrections

At finite width, all of the correlators receive an infinite series of subleading corrections. Concretely, the metric governing the two-point correlator and the four-point vertex governing the connected four-point correlator have $1/n$ series expansions of the form

$$G^{(\ell)}_{\alpha_1\alpha_2} = G^{\{0\}(\ell)}_{\alpha_1\alpha_2} + \frac{1}{n_{\ell-1}} G^{\{1\}(\ell)}_{\alpha_1\alpha_2} + \frac{1}{n_{\ell-1}^2} G^{\{2\}(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n^3}\right), \tag{4.104}$$

$$V^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = V^{\{0\}(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \frac{1}{n_{\ell-1}} V^{\{1\}(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + O\left(\frac{1}{n^2}\right). \tag{4.105}$$

While so far we have focused on the leading contributions $G^{\{0\}(\ell)}_{\alpha_1\alpha_2}$ and $V^{\{0\}(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$, the subleading corrections can be systematically calculated as well. Let us illustrate the procedure by deriving the recursion for the next-to-leading-order (NLO) correction to the metric, $G^{\{1\}(\ell)}_{\alpha_1\alpha_2}$.

Before proceeding, let us remark that the leading contribution of the mean metric fully describes the infinite-width limit of the preactivation distributions and so is given a symbol,

$$K^{(\ell)}_{\alpha_1\alpha_2} \equiv G^{\{0\}(\ell)}_{\alpha_1\alpha_2}, \tag{4.106}$$

and a name, the **kernel**. Since the kernel captures the leading-order correlation between any pair of samples, it will be a central object of study for us in the following chapters. In a similar vein, we will call $G^{\{1\}(\ell)}_{\alpha_1\alpha_2}$ the **NLO metric**.

[17] Note that the coupling $g_{(\ell),m_\ell}$ depends on m_ℓ – and also on the other hidden-layer widths $n_1, n_2, \ldots, n_{\ell-1}$ – but does *not* depend on the overall width of the current layer n_ℓ. This implies that the quadratic coupling $g_{(\ell),m_\ell}$ is the same coupling we would have used if instead there were actually only m_ℓ neurons in the ℓ-th layer.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

Our first step will be to express the layer-ℓ quadratic coupling $g^{\beta_1\beta_2}_{(\ell)}$ to order $1/n$ in terms of the $1/n$ correlator data in (4.104) and (4.105). Let's begin by recalling the expression (4.47) for the two-point correlator that we derived from the quartic action, reprinted here for layer ℓ

$$\begin{aligned}\mathbb{E}\left[z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_2;\alpha_2}\right] &= \delta_{i_1 i_2} G^{(\ell)}_{\alpha_1\alpha_2} \qquad (4.107)\\ &= \delta_{i_1 i_2}\left[g^{(\ell)}_{\alpha_1\alpha_2} + \frac{1}{2}\sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} v^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)}\left(n_\ell\, g^{(\ell)}_{\alpha_1\beta_1} g^{(\ell)}_{\alpha_2\beta_2} g^{(\ell)}_{\beta_3\beta_4} + 2 g^{(\ell)}_{\alpha_1\beta_1} g^{(\ell)}_{\alpha_2\beta_3} g^{(\ell)}_{\beta_2\beta_4}\right)\right]\\ &\quad + O\left(v^2\right).\end{aligned}$$

As a reminder, $g^{(\ell)}_{\alpha_1\alpha_2}$ is the matrix inverse of the quadratic coupling $g^{\alpha_1\alpha_2}_{(\ell)}$. Substituting the expansion (4.104) into (4.107), substituting for the quartic coupling $v_{(\ell)} = V_{(\ell)}/n_{\ell-1}$ (4.82), and rearranging to solve for $g^{(\ell)}_{\alpha_1\alpha_2}$ to the subleading order, we get

$$\begin{aligned}g^{(\ell)}_{\alpha_1\alpha_2} &= K^{(\ell)}_{\alpha_1\alpha_2} + \frac{1}{n_{\ell-1}}\left[G^{\{1\}(\ell)}_{\alpha_1\alpha_2} - \sum_{\beta_1,\beta_2\in\mathcal{D}} K^{\beta_1\beta_2}_{(\ell)}\left(\frac{n_\ell}{2} V^{(\ell)}_{(\alpha_1\alpha_2)(\beta_1\beta_2)} + V^{(\ell)}_{(\alpha_1\beta_1)(\alpha_2\beta_2)}\right)\right]\\ &\quad + O\left(\frac{1}{n^2}\right). \qquad (4.108)\end{aligned}$$

Note that in obtaining the above, we have self-consistently replaced $g^{(\ell)}$ by $K^{(\ell)}$ in the subleading term, which in turn let us lower the indices of the four-point vertices. Inverting this expression (4.108) yields the subleading correction to the quadratic coupling in terms of the correlators

$$\begin{aligned}&g^{\beta_1\beta_2}_{(\ell)} - K^{\beta_1\beta_2}_{(\ell)} \qquad (4.109)\\ &= \frac{1}{n_{\ell-1}}\sum_{\beta_3,\beta_4\in\mathcal{D}}\left[-K^{\beta_1\beta_3}_{(\ell)} K^{\beta_2\beta_4}_{(\ell)} G^{\{1\}(\ell)}_{\beta_3\beta_4} + K^{(\ell)}_{\beta_3\beta_4}\left(\frac{n_\ell}{2} V^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)} + V^{(\beta_1\beta_3)(\beta_2\beta_4)}_{(\ell)}\right)\right]\\ &\quad + O\left(\frac{1}{n^2}\right).\end{aligned}$$

Note that one term in this correction scales as $n_\ell/n_{\ell-1}$. As discussed in the previous section, the marginalization rule for the ℓ-th-layer action guarantees that we can treat this quantity as small, $n_\ell/n_{\ell-1} \ll 1$, ensuring that $g^{(\ell)}_{\alpha_1\alpha_2} - K^{(\ell)}_{\alpha_1\alpha_2}$ is a subleading-in-$1/n$ correction to the quadratic coupling. In line with this statement, we'll soon see the cancellation for the factor of n_ℓ when computing the recursion for this subleading correction to the metric $G^{\{1\}(\ell)}$.

Having finished working out the $1/n$-corrected ℓ-th-layer action, we turn to computing the $(\ell+1)$-th-layer two-point correlator.[18] This will let us express the $(\ell+1)$-th-layer

[18] We already knew the $1/n$ contribution to the quartic coupling, namely the relation $v_{(\ell)} = V_{(\ell)}/n_{\ell-1}$.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

two-point correlator in terms of the ℓ-th-layer statistics, ultimately yielding a recursion for $G^{\{1\}(\ell)}_{\alpha_1\alpha_2}$. Starting with the expansion (4.104) in the $(\ell+1)$-th layer and substituting in the expression (4.85) for the two-point correlator, we obtain

$$K^{(\ell+1)}_{\alpha_1\alpha_2} + \frac{1}{n_\ell} G^{\{1\}(\ell+1)}_{\alpha_1\alpha_2} + O\!\left(\frac{1}{n^2}\right) = G^{(\ell+1)}_{\alpha_1\alpha_2} = C_b^{(\ell+1)} + C_W^{(\ell+1)} \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1} \sigma^{(\ell)}_{j;\alpha_2}\right]. \tag{4.110}$$

Thus, we need the expectation of two activations in the ℓ-th layer up to the order $O(1/n)$, which we evaluated before in expression (4.61) in terms of the ℓ-th-layer couplings.

Looking at (4.61), there are two types of contributions at the subleading order, one arising from the $1/n$ correction to the quadratic coupling $g_{(\ell)}$ in (4.108) and the other from the near-Gaussianity of the distribution due to the quartic coupling $v_{(\ell)}$. The latter contribution is easy to handle: since the quartic coupling is already suppressed by $1/n$, we can just make the replacement $g^{(\ell)} \to K^{(\ell)}$ in the second term in (4.61), yielding

$$\begin{aligned}
&\frac{1}{8n_{\ell-1}} \sum_{\beta_1,\ldots,\beta_4 \in \mathcal{D}} V^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)} \\
&\quad\times \Big[\Big\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \left(z_{\beta_1} z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right) \left(z_{\beta_3} z_{\beta_4} - K^{(\ell)}_{\beta_3\beta_4}\right) \Big\rangle_{K^{(\ell)}} \\
&\quad + 2n_\ell \Big\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \left(z_{\beta_1} z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right) \Big\rangle_{K^{(\ell)}} K^{(\ell)}_{\beta_3\beta_4} - 2 \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \right\rangle_{K^{(\ell)}} K^{(\ell)}_{\beta_1\beta_3} K^{(\ell)}_{\beta_2\beta_4} \Big] + O\!\left(\frac{1}{n^2}\right).
\end{aligned} \tag{4.111}$$

However, for the former contribution, the Gaussian term $\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{g^{(\ell)}}$ needs be carefully separated into the leading and subleading pieces. To that end, we can trade the Gaussian expectation with $g^{(\ell)}$ for one in terms of the leading kernel $K^{(\ell)}$:

$$\begin{aligned}
&\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{g^{(\ell)}} \\
&\quad= \frac{\Big\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \exp\!\Big[-\frac{1}{2}\sum_{\beta_1,\beta_2} \left(g^{\beta_1\beta_2}_{(\ell)} - K^{\beta_1\beta_2}_{(\ell)}\right) z_{\beta_1} z_{\beta_2}\Big] \Big\rangle_{K^{(\ell)}}}{\Big\langle \exp\!\Big[-\frac{1}{2}\sum_{\beta_1,\beta_2} \left(g^{\beta_1\beta_2}_{(\ell)} - K^{\beta_1\beta_2}_{(\ell)}\right) z_{\beta_1} z_{\beta_2}\Big] \Big\rangle_{K^{(\ell)}}} \\
&\quad= \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{K^{(\ell)}} - \frac{1}{2} \sum_{\beta_1,\beta_2} \left(g^{\beta_1\beta_2}_{(\ell)} - K^{\beta_1\beta_2}_{(\ell)}\right) \Big\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \left(z_{\beta_1} z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right) \Big\rangle_{K^{(\ell)}} + O\!\left(\frac{1}{n^2}\right).
\end{aligned} \tag{4.112}$$

Plugging (4.109) into (4.112), we obtain the subleading contribution due to the change in the quadratic coupling, giving

$$\begin{aligned}
&\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{g^{(\ell)}} \\
&= \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{K^{(\ell)}} + \frac{1}{2n_{\ell-1}} K^{\beta_1\beta_3}_{(\ell)} K^{\beta_2\beta_4}_{(\ell)} G^{\{1\}(\ell)}_{\beta_3\beta_4} \Big\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \left(z_{\beta_1} z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right) \Big\rangle_{K^{(\ell)}} \\
&\quad - \frac{1}{n_{\ell-1}} \sum_{\beta_1,\ldots,\beta_4} \left(\frac{n_\ell}{4} V^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)} + \frac{1}{2} V^{(\beta_1\beta_3)(\beta_2\beta_4)}_{(\ell)} \right) K^{(\ell)}_{\beta_3\beta_4} \Big\langle \sigma_{\alpha_1}\sigma_{\alpha_2} \left(z_{\beta_1} z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right) \Big\rangle_{K^{(\ell)}} \\
&\quad + O\!\left(\frac{1}{n^2}\right).
\end{aligned} \tag{4.113}$$

Now that we've computed everything, we can add the two contributions to $\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right]$, (4.111) and (4.113), and plug them into the expression for the preactivation correlator (4.110). Collecting terms, we recover the leading contribution, the recursion for the kernel,

$$K^{(\ell+1)}_{\alpha_1\alpha_2} = C^{(\ell+1)}_b + C^{(\ell+1)}_W \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{K^{(\ell)}}, \tag{4.114}$$

and also find a recursion for the NLO metric as promised:

$$\begin{aligned} &\frac{1}{n_\ell} G^{\{1\}(\ell+1)}_{\alpha_1\alpha_2} \qquad (4.115)\\ &= C^{(\ell+1)}_W \frac{1}{n_{\ell-1}} \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} \Bigg[\frac{1}{2} K^{\beta_1\beta_3}_{(\ell)} K^{\beta_2\beta_4}_{(\ell)} G^{\{1\}(\ell)}_{\beta_3\beta_4} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{K^{(\ell)}} \\ &\qquad + \frac{1}{8} V^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right)\left(z_{\beta_3}z_{\beta_4} - K^{(\ell)}_{\beta_3\beta_4}\right)\right\rangle_{K^{(\ell)}} \\ &\qquad + \frac{1}{4} V^{(\beta_1\beta_3)(\beta_2\beta_4)}_{(\ell)} K^{(\ell)}_{\beta_3\beta_4} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(-2z_{\beta_1}z_{\beta_2} + K^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{K^{(\ell)}}\Bigg]. \end{aligned}$$

In going through this calculation in your personal notes or on the margins of this book, you can explicitly see the cancellation of contributions from $n_\ell - 1$ spectator neurons that do not participate in the expectation $\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right]$, as required by the marginalization rule for the ℓ-th-layer action. Indeed, every term in the square bracket on the right-hand side of (4.115) is manifestly of order one.

This process can be systematically pushed to higher orders. Just as the computation of the NLO metric $G^{\{1\}(\ell)}_{\alpha_1\alpha_2}$ involved the leading quartic coupling, the computation of the subleading correction to the four-point vertex, $V^{\{1\}(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$, and the computation of the order $1/n^2$ correction to the two-point correlator, $G^{\{2\}(\ell)}_{\alpha_1\alpha_2}$, would involve the leading sextic coupling. Such a sextic coupling appears at order $1/n^2$ in the action and contributes to the connected six-point function, which also vanishes as $O(1/n^2)$.[19]

4.6 RG Flow and RG Flow

Since the past five sections have been a whirlwind of equations, algebra, and integration, let's take a moment to recap and assemble the main results.

The goal of this chapter was to find the marginal distribution of preactivations $p\left(z^{(\ell)}\middle|\mathcal{D}\right)$ in a given layer ℓ in terms of an **effective action** with data-dependent couplings. These couplings change – or **run** – from layer to layer, and the running

[19] For those familiar with field theory, the leading part of the couplings in the action are *tree-level* contributions to correlators. They are to be contrasted with subleading corrections to the two-point correlator discussed in this section, which included both *loop-level* contributions from quartic interaction and tree-level contributions from the NLO correction to the bare quadratic coupling.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

is determined via recursions, which in turn determine how the distribution of preactivations changes with depth. Equivalently, these recursions tell us how correlators of preactivations evolve with layer. In this language, starting with independent neurons in the first layer (§4.1), we saw how interactions among neurons are induced in the second layer (§4.2) and then amplified in deeper layers (§4.3).

Concretely, let's summarize the behavior of finite-width networks to leading order in the wide-network expansion. Expressing the two-point correlator of preactivations in terms of the **kernel** $K^{(\ell)}_{\alpha_1\alpha_2}$ as

$$\mathbb{E}\left[z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_2;\alpha_2}\right] = \delta_{i_1 i_2} G^{(\ell)}_{\alpha_1\alpha_2} = \delta_{i_1 i_2}\left[K^{(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n}\right)\right], \tag{4.116}$$

and expressing the four-point connected correlator in terms of the **four-point vertex** $V^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ as

$$\begin{aligned}&\mathbb{E}\left[z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_2;\alpha_2} z^{(\ell)}_{i_3;\alpha_3} z^{(\ell)}_{i_4;\alpha_4}\right]\Big|_{\text{connected}} \qquad (4.117)\\ &= \frac{1}{n_{\ell-1}}\left[\delta_{i_1 i_2}\delta_{i_3 i_4} V^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4} V^{(\ell)}_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)} + \delta_{i_1 i_4}\delta_{i_2 i_3} V^{(\ell)}_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)}\right],\end{aligned}$$

the running of these correlators is given by the recursions

$$K^{(\ell+1)}_{\alpha_1\alpha_2} = C^{(\ell+1)}_b + C^{(\ell+1)}_W \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{K^{(\ell)}}, \tag{4.118}$$

$$\begin{aligned}V^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} &= \left(C^{(\ell+1)}_W\right)^2 \left[\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{K^{(\ell)}} - \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{K^{(\ell)}}\langle \sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{K^{(\ell)}}\right] \qquad (4.119)\\ &+ \frac{1}{4}\left(C^{(\ell+1)}_W\right)^2 \frac{n_\ell}{n_{\ell-1}} \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} V^{(\beta_1\beta_2)(\beta_3\beta_4)}_{(\ell)} \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2} - K^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{K^{(\ell)}}\\ &\times \left\langle \sigma_{\alpha_3}\sigma_{\alpha_4}\left(z_{\beta_3}z_{\beta_4} - K^{(\ell)}_{\beta_3\beta_4}\right)\right\rangle_{K^{(\ell)}} + O\left(\frac{1}{n}\right),\end{aligned}$$

where the indices on the four-point vertex are raised by the inverse metric $G_{(\ell)}$:

$$\begin{aligned}V^{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}_{(\ell)} &\equiv \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} G^{\alpha_1\beta_1}_{(\ell)} G^{\alpha_2\beta_2}_{(\ell)} G^{\alpha_3\beta_3}_{(\ell)} G^{\alpha_4\beta_4}_{(\ell)} V^{(\ell)}_{(\beta_1\beta_2)(\beta_3\beta_4)} \qquad (4.120)\\ &= \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} K^{\alpha_1\beta_1}_{(\ell)} K^{\alpha_2\beta_2}_{(\ell)} K^{\alpha_3\beta_3}_{(\ell)} K^{\alpha_4\beta_4}_{(\ell)} V^{(\ell)}_{(\beta_1\beta_2)(\beta_3\beta_4)} + O\left(\frac{1}{n}\right).\end{aligned}$$

These recursions dictate how the statistics of preactivations flow with depth.

This flow is very reminiscent of the following heuristic picture, which is offered as an explanation for how neural networks are supposed to work: given an input, such as the image of a cat, the first few layers identify low-level features from the pixels – such as the `edges` between areas of low and high intensity – and then the middle layers assemble these low-level features into mid-level features – such as the texture and pattern of `fur` – which are further aggregated in deeper layers into higher-level representations – such as `tails` and `ears` – which the last layer combines into an estimate of the probability that the

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

original pixels represent a `cat`. Indeed, some studies support this hierarchically-ordered arrangement of feature representation in trained networks [34].[20] The desirability of such an arrangement emphasizes both the role and importance of depth in deep learning.

Some of the terms we used in discussing this heuristic picture can actually be given more precise definitions. For instance, each neuron in the network – including not only those in the output layer but also those in the hidden layers – is a scalar function of the input and called a **feature**. The neurons of a given layer can be organized into a vector-valued function of the input, which we'll refer to as a **representation**.[21] In terms of these concepts, our formalism tracks the transformation of representations from one layer to the next. It is this flow of representations that we term **representation group flow** or **RG flow** for short.[22] RG flow is induced via the repeated marginalization of fine-grained features in the shallow layers to give a coarse-grained representation in the output layer. Our notion of RG flow makes the heuristic picture given above concrete.

This pattern of coarse-graining has a parallel in theoretical physics, known as **renormalization group flow** or **RG flow** for short. In this case, the RG flow is generated by the repeated marginalization of the microscopic fine-grained degrees of freedom in the system in order to obtain an *effective theory* of the system in terms of macroscopic coarse-grained variables. Analogously, the physical couplings controlling the interactions of these effective degrees of freedom *run* with the length scale at which they are probed – e.g., the effective charge of the electron will change when interrogated at different scales. Similar to the recursion equations describing the running couplings of the network representations, one can derive differential equations – historically called beta functions – that govern the running of the physical couplings with scale.[23]

[20] It has been suggested that even untrained networks have features that can act as types of filters, effectively allowing for primitive edge detection in untrained networks. For a related set of ideas, see [35].

[21] While the main focus of our study of supervised learning (§7) will be understanding how the representation $z^{(L)}$ in the output layer is learned via gradient-based training, it is also important to understand how representations are learned in hidden layers (§11). In addition to being necessary components of determining the coarse-grained representation at the output, in some applications of deep learning, learned representations in the hidden layers can be used as inputs themselves for other learning tasks. This occurs quite often in unsupervised learning, for example with the word embeddings of natural language processing tasks. In these scenarios, the embeddings – representations for an input word in the larger context of a full sentence – typically are taken not just from the final layer but from the concatenation of the final few layers. See, e.g., [36].

[22] Two apologies are in order for the name *representation group flow*: *(i)* it is confusingly close to the notion of *group representation theory* in mathematics; and *(ii)* the flow is technically a *semigroup*, rather than a group. (Both group theory and semigroup theory are the studies of transformations, but a group requires inverses, while a semigroup does not; and the flow has no inverse.) This is just to repeat a historic mistake in physics as we'll explain further in a footnote below.

[23] *A brief history of renormalization in physics.*

Renormalization was originally developed in the 1930s and 1940s to deal with divergences – infinities – that plagued the calculations of experimental observables in quantum field theory. At first, these infinities were simply subtracted off – swept under the rug, if you will – yielding answers that, despite these shenanigans, matched extremely well with experiments. This whole state of affairs was considered embarrassing, leading to near abandonment of the theory.

These divergences arose essentially due to a failure to properly take into account that couplings can be scale-dependent. The idea of running couplings was first put forth by Gell-Mann and Low [37] in

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

To make the connection between this RG flow and that RG flow abundantly clear, let's peek into how it is implemented in field theory in physics. In this scenario, the degrees of freedom are represented by a field $\phi(x)$ that may take different values as a function of spacetime coordinate x. First, one divides $\phi(x)$ into fine-grained variables ϕ^+ consisting of high-frequency modes and coarse-grained variables ϕ^- consisting of low-frequency modes, such that the field decomposes as $\phi(x) = \phi^+(x) + \phi^-(x)$. The full distribution is governed by the full action

$$S_{\text{full}}(\phi) = S(\phi^+) + S(\phi^-) + S_{\text{I}}(\phi^+, \phi^-), \tag{4.121}$$

where in particular the last term describes the interactions between these two sets of modes.

Now, if all we care about are observables that depend only on the coarse-grained modes ϕ^- at macroscopic scales – and such long-range scales are usually the relevant ones for experiments – then this full description is too cumbersome to usefully describe the outcome of such experiments. In order to obtain an effective description in terms of only these coarse-grained variable ϕ^-, we can integrate out (i.e., marginalize over) the fine-grained variables ϕ^+ as

$$e^{-S_{\text{eff}}(\phi^-)} = \int d\phi^+ \; e^{-S_{\text{full}}(\phi)} \tag{4.122}$$

and obtain an **effective action** $S_{\text{eff}}(\phi^-)$, providing an effective theory for the observables of experimental interest. In practice, this marginalization is carried out scale by scale, dividing up the field as $\phi = \phi^{(1)} + \cdots + \phi^{(L)}$ from microscopic modes $\phi^{(1)}$ all the way to macroscopic modes $\phi^{(L)} = \phi^-$, and then integrating out the variables $\phi^{(1)}$, $\ldots$, $\phi^{(L-1)}$ in sequence. Tracking the flow of couplings in the effective action through this marginalization results in the aforementioned beta functions, and in solving these differential equations up to the scale of interest, we get an effective description of observables at that scale.

1954; however, a full conceptualization of renormalization wasn't available until Wilson developed the modern notion of RG flow [38, 39] in 1971, offering a theoretical explanation for critical phenomena in statistical physics as well as giving a sound grounding for the understanding of divergences in quantum field theory.

At this point, all historical accounts of RG are contractually obligated to mention the following: the renormalization group is not a group; it's a semigroup. (The mistake was made in an early paper by Stueckelberg and Petermann, referring to the flow as a "group of normalization" [40].) Mathematically, this is because there are no inverse elements; the marginalization of variables out of a joint distribution deletes information and cannot be undone. In particular, two different joint distributions can sometimes flow to the same distribution after marginalization. Intuitively, this is because these flows go from fine-grained descriptions to coarse-grained descriptions. (Such convergent flows lead to the notion of *universality*, which we will explain in §5 in the context of neural networks with different activations that flow to the same marginal distributions under RG.)

Clearly, RG flow in physics is a very rich subject. If you're interested in learning more, we recommend both [41, 42].

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

This is precisely what we have been doing in this chapter for neural networks. The full field ϕ is analogous to a collection of all the preactivations $\left\{z^{(1)},\dots,z^{(L)}\right\}$. Their distribution is governed by the full joint distribution of preactivations

$$p\Big(z^{(1)},\dots,z^{(L)}\Big|\mathcal{D}\Big)=p\Big(z^{(L)}\Big|z^{(L-1)}\Big)\cdots p\Big(z^{(2)}\Big|z^{(1)}\Big)\,p\Big(z^{(1)}\Big|\mathcal{D}\Big)\,, \tag{4.123}$$

with the full action

$$S_{\text{full}}\Big(z^{(1)},\dots,z^{(L)}\Big)\equiv\sum_{\ell=1}^{L}S_{\text{M}}\Big(z^{(\ell)}\Big)+\sum_{\ell=1}^{L-1}S_{\text{I}}\Big(z^{(\ell+1)}\Big|z^{(\ell)}\Big)\,. \tag{4.124}$$

Here, the full action is decomposed into the mean quadratic action for variables $z^{(\ell)}$,

$$S_{\text{M}}\Big(z^{(\ell)}\Big)=\frac{1}{2}\sum_{i=1}^{n_\ell}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}G_{(\ell)}^{\alpha_1\alpha_2}z_{i;\alpha_1}^{(\ell)}z_{i;\alpha_2}^{(\ell)}, \tag{4.125}$$

in terms of the mean metric $G^{(\ell)}$, (4.72), and the interaction between neighboring layers

$$S_{\text{I}}\Big(z^{(\ell+1)}\Big|z^{(\ell)}\Big)=\frac{1}{2}\sum_{i=1}^{n_{\ell+1}}\sum_{\alpha_1,\alpha_2\in\mathcal{D}}\left[\widehat{G}_{(\ell+1)}^{\alpha_1\alpha_2}\Big(z^{(\ell)}\Big)-G_{(\ell+1)}^{\alpha_1\alpha_2}\right]z_{i;\alpha_1}^{(\ell+1)}z_{i;\alpha_2}^{(\ell+1)}. \tag{4.126}$$

Here we emphasized that the stochastic metric $\widehat{G}_{\alpha_1\alpha_2}^{(\ell+1)}$ is a function of $z^{(\ell)}$, and the induced coupling of $z^{(\ell)}$ with $z^{(\ell+1)}$ is what leads to the interlayer interactions.

Now, if all we care about are observables that depend only on the outputs of the network – which includes a very important observable ... the output! – then this full description is too cumbersome. In order to obtain an effective (i.e., useful) description of the distribution of outputs $z^{(L)}$, we can marginalize over all the features $\left\{z^{(1)},\dots,z^{(L-1)}\right\}$ as

$$e^{-S_{\text{eff}}\left(z^{(L)}\right)}=\int\left[\prod_{\ell=1}^{L-1}dz^{(\ell)}\right]e^{-S_{\text{full}}\left(z^{(1)},\dots,z^{(L)}\right)}, \tag{4.127}$$

just as we integrated out the fine-grained modes ϕ^+ in (4.122) to get the effective description in terms of coarse-grained modes ϕ^-. And, just as in the field theory example, rather than carrying out this marginalization all at once, we proceeded sequentially, integrating out the preactivations layer by layer. This resulted in the recursion relations (4.118) and (4.119), and in solving these recursion relations up to the depth of interest, we get an effective description of neural network output at that depth.[24]

Now, this last sentence suggests a subtle but interesting shift of perspective, so let us elaborate. So far in this chapter, we have implicitly assumed a fixed network depth L and

[24] Note to physicists: the flow in networks from input to output is a flow from the ultraviolet to the infrared.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press

described how the preactivation distribution changes as an input x propagates through the intermediate layers, yielding recursion relations for correlators and couplings for the evolution from layer ℓ to layer $\ell+1$, for $\ell = 0, \ldots, L-1$. However, it is also valid to view the resulting recursion equations as governing the change in output distributions as the overall network depth changes from L to $L+1$.[25] In other words, these recursion relations describe the effect of adding an additional layer to the neural network by comparing distributions $p\Big(z^{(L)}\Big|\mathcal{D}\Big)$ and $p\Big(z^{(L+1)}\Big|\mathcal{D}\Big)$.

Given this perspective, our RG flow can address head-on the effect of the *deep* in deep learning. For instance, as a network gets deeper, do the interactions between neurons – encoded in the finite-width corrections such as the four-point vertex $V^{(\ell)}$ – get amplified or attenuated? In the language of RG flow, couplings that grow with the flow are called **relevant**, and those that shrink are called **irrelevant**.[26] These names are evocative of whether the interaction matters or not for the effective theory, and so we'll employ the same terminology. Thus, to explore the effect of depth on the neuron–neuron interactions, we are simply asking whether the four-point vertex $V^{(\ell)}$ is relevant or irrelevant.

This question has important implications for deep learning. If all the finite-width couplings were irrelevant, then finite-width networks would asymptote to infinite-width architectures under RG flow. This would then mean that these networks behave more like infinite-width models as they get deeper, and so deep learning would really be the study of these much simpler Gaussian models. Fortunately we'll soon find that the couplings *are* relevant, making our life richer, albeit more complicated. In the next chapter, we'll show that finite networks deviate more and more from their infinite-width counterparts as they get deeper. This has important practical consequences in controlling the instantiation-to-instantiation fluctuations in supervised training and also in allowing networks to learn nontrivial representations of their input (§11).

The next chapter explores these relevant questions by explicitly solving recursion equations such as (4.118) and (4.119).

[25] To be precise, the output dimension n_{out} is fixed. So, as the depth changes from L to $L+1$, we imagine holding fixed the widths for $\ell < L$, inserting a new layer L with $n_L \sim n \gg 1$, and then setting the final layer $L+1$ to have width n_{out}.

[26] Couplings that neither grow nor shrink are called **marginal**.

https://doi.org/10.1017/9781009023405.006 Published online by Cambridge University Press