8

RG Flow of the Neural Tangent Kernel

People get things backwards and they shouldn't—it has been said, and wisely said, that every successful physical theory swallows its predecessors alive.

Sidney Coleman, more forward and a little bit deeper in that same "Quantum Mechanics in Your Face" Dirac Lecture [56].

In the last chapter, we introduced gradient-based learning as an alternative to Bayesian learning and specifically focused on the gradient descent algorithm. In short, the gradient descent algorithm involved instantiating a network from the prior distribution and then repeatedly updating the model parameters by running training data through the network. This algorithm is straightforward to implement and very efficient to run for any particular network. In practice, it makes things very easy.

In theory, it makes things a little more difficult. For the Bayesian prior, we were able to integrate out the model parameters layer by layer in deriving the output distribution because the initialization distribution of the biases and weights was extremely simple; in addition, the large-width expansion made it possible to derive analytic expressions for the Bayesian posterior for finite-width networks. By contrast, the model parameters and the outputs of any particular network trained by gradient descent are a complicated correlated mess.

To make progress, we first need to shift the perspective back to a statistical one. Rather than focusing on how any particular network learns from the data, we instead ask how a *typical* network behaves when being trained. If we understand the typical behavior (i.e., the mean) under gradient descent and have control of the fluctuations from network instantiation to instantiation (i.e., the variance), then we can describe gradient-based learning as used in practice.

With that statistical perspective in mind, recall from the last chapter that the gradient-descent updates decompose into an error factor times a function-approximation factor. The latter factor was dubbed the *neural tangent kernel* (NTK) and conveniently summarizes the effect of changes in the model parameters on the behavior of the network.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

This means that the statistics of changes in network observables in the initial stage of training are governed by the statistics of the NTKs at initialization. To proceed forward, the core of the current chapter and the next will involve explicitly computing such NTK statistics for deep MLPs; we will postpone the actual analysis of neural network training – enabled by these computations of the NTK statistics – until §10 and §∞.

In §8.0, we will lay the groundwork for the recursive computation of the NTK statistics. Namely, starting from the MLP iteration equation, or the *forward* equation for the preactivations, we'll derive a corresponding forward equation for the NTK. This equation is a layer-to-layer iteration equation that holds for each distinct instantiation of the model parameters. (Here we'll also remark on how the learning-rate tensor should be scaled with network width – an important point that is often neglected in practice.)

By averaging over different instantiations, we can then use the forward equation to recursively compute the joint statistics of the NTK and the preactivations. The approach taken here completely mirrors the *RG-flow* approach taken in §4 for the preactivations. In §8.1, §8.2, and §8.3, we will progressively determine the sequence of joint NTK-preactivation distributions in the first, second, and deeper layers, respectively.

8.0 Forward Equation for the NTK

As we saw in the previous chapter, the evolution of observables $\mathcal{O}(z)$ under gradient descent is governed by the NTK,

$$H_{i_1 i_2;\alpha_1\alpha_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i_1;\alpha_1}}{d\theta_\mu} \frac{dz_{i_2;\alpha_2}}{d\theta_\nu} \,, \tag{8.1}$$

where $\lambda_{\mu\nu}$ is the learning-rate tensor.

Specializing to MLPs, observables can depend not only on the network's output $z_{i;\alpha} = z^{(L)}_{i;\alpha}$ but also on the preactivations $z_{i;\alpha} = z^{(\ell)}_{i;\alpha}$ in any layer. Such ℓ-th-layer observables for $\ell < L$ tell us about the *hidden-layer representations* of the network. For instance, the neural component $\mathcal{O} = z^{(\ell)}_i(x)$ tells us about an ℓ-th-layer feature evaluated on an input x, while $\mathcal{O} = z^{(\ell)}_i(x)\, z^{(\ell)}_j(x)$ with neural indices $i \neq j$ tracks correlations among different features given x.

With similar manipulations as before, we find that an observable $\mathcal{O}$ that depends only on the ℓ-th-layer preactivations

$$\mathcal{O}(\theta) \equiv \mathcal{O}\Big(z^{(\ell)}(x_{\delta_1};\theta), \ldots , z^{(\ell)}(x_{\delta_M};\theta)\Big) \tag{8.2}$$

evolves after a gradient descent update as

$$\mathcal{O}\Big(\theta(t+1)\Big) - \mathcal{O}\Big(\theta(t)\Big) = -\eta \sum_{i_1,i_2=1}^{n_\ell} \sum_{\alpha\in\mathcal{A}} \sum_{\delta\in\mathcal{D}} \left[\frac{d\mathcal{L}_\mathcal{A}}{dz^{(\ell)}_{i_1;\alpha}} \frac{\partial\mathcal{O}}{\partial z^{(\ell)}_{i_2;\delta}} \right] H^{(\ell)}_{i_1 i_2;\alpha\delta} + O(\eta^2), \tag{8.3}$$

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

where $x_{\delta_1}, \ldots, x_{\delta_M} \in \mathcal{D}$ for some dataset $\mathcal{D}$.[1] Here, we have defined the **ℓ-th-layer NTK** as

$$H^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz^{(\ell)}_{i_1;\alpha_1}}{d\theta_\mu} \frac{dz^{(\ell)}_{i_2;\alpha_2}}{d\theta_\nu}, \tag{8.4}$$

which governs the evolution of the ℓ-th-layer observables; in terms of this notation, the output NTK is simply $H^{(\ell=L)}_{i_1 i_2;\alpha_1\alpha_2}$. Note that whenever we write the ℓ-th-layer NTK as above, we will always assume that the learning-rate tensor $\lambda_{\mu\nu}$ does not mix network parameters from different layers, though in general it can still mix the biases and weights within a layer. We will place further restrictions on this in another paragraph.

At initialization, the model parameters are sampled from their initialization distributions, and the ℓ-th-layer NTK is a stochastic object. In order to emphasize this stochasticity, in what follows we'll decorate the NTK *at initialization* with a hat: $\widehat{H}^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2}$. Our goal is to evaluate its statistics.

Before we go any further, it is convenient to make a specialized choice for the learning-rate tensor $\lambda_{\mu\nu}$. In practice, typically $\lambda_{\mu\nu} = \delta_{\mu\nu}$, and there is only the *global* learning rate η for the entire model. Even in a more general setup, a learning rate is often shared among each group of parameters that are sampled from the same distribution. Recalling that the same distribution was shared among the biases in a given layer with the same variance $C^{(\ell)}_b$, (2.19), and similarly for the weights with the *rescaled* weight variance $C^{(\ell)}_W$, (2.20), this suggests an ansatz for our **training hyperparameters**: we should decompose the learning-rate tensor $\lambda_{\mu\nu}$ into a diagonal matrix

$$\lambda_{b^{(\ell)}_{i_1} b^{(\ell)}_{i_2}} = \delta_{i_1 i_2} \lambda^{(\ell)}_b, \qquad \lambda_{W^{(\ell)}_{i_1 j_1} W^{(\ell)}_{i_2 j_2}} = \delta_{i_1 i_2}\delta_{j_1 j_2} \frac{\lambda^{(\ell)}_W}{n_{\ell-1}}, \tag{8.5}$$

giving each group of biases in a layer the same learning rate and each group of weights in a layer the same learning rate, and allowing such learning rates to vary from layer to layer.

Importantly, we have normalized the learning rate for a given weight $W^{(\ell)}_{i_1 j_1}$ by the width of the previous layer $n_{\ell-1}$, just as we did for the variance of the weight's initialization distribution. This normalization is there for much the same reason: the freedom to tune the weight learning rates separately from the bias learning rates will

[1] However, note that this is not quite as simple as the expression for the evolution of the network output that we gave in the last chapter, (7.19). In particular, the derivative of the loss with respect to the ℓ-th-layer preactivations needs to be computed by the chain rule as

$$\frac{d\mathcal{L}_\mathcal{A}}{dz^{(\ell)}_{i_1;\alpha}} = \sum_{j=1}^{n_L} \frac{\partial \mathcal{L}_\mathcal{A}}{\partial z^{(L)}_{j;\alpha}} \frac{dz^{(L)}_{j;\alpha}}{dz^{(\ell)}_{i_1;\alpha}}, \tag{8.6}$$

with the error factor $\partial\mathcal{L}_\mathcal{A}/\partial z^{(L)}_{j;\alpha}$ now multiplied by the chain-rule factor $dz^{(L)}_{j;\alpha}/dz^{(\ell)}_{i_1;\alpha}$. For observables that depend on preactivations from multiple layers, the generalization of (8.3) further involves additional chain-rule factors as well as a sum over NTKs from different layers.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

prove necessary for having a sensible large-width expansion. Going forward, our training hyperparameters will consist of the global learning rate η and the individual ℓ-th-layer learning rates for the biases and weights, $\lambda_b^{(\ell)}$ and $\lambda_W^{(\ell)}$.

Substituting our choice for $\lambda_{\mu\nu}$ back into the definition of the ℓ-th-layer NTK, (8.4), this expression decomposes as

$$\widehat{H}^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2} = \sum_{\ell'=1}^{\ell}\left[\sum_{j=1}^{n_{\ell'}}\left(\lambda_b^{(\ell')}\frac{dz^{(\ell)}_{i_1;\alpha_1}}{db_j^{(\ell')}}\frac{dz^{(\ell)}_{i_2;\alpha_2}}{db_j^{(\ell')}} + \frac{\lambda_W^{(\ell')}}{n_{\ell'-1}}\sum_{k=1}^{n_{\ell'-1}}\frac{dz^{(\ell)}_{i_1;\alpha_1}}{dW_{jk}^{(\ell')}}\frac{dz^{(\ell)}_{i_2;\alpha_2}}{dW_{jk}^{(\ell')}}\right)\right]. \tag{8.7}$$

Here, the part in the square brackets is the per-layer contribution of the model parameters to the ℓ-th-layer NTK, treating the biases and weights separately. We also see that our intuition above in (8.5) was correct: the ℓ'-th-layer weight learning rate $\lambda_W^{(\ell')}$ needs to be accompanied by a factor of $1/n_{\ell'-1}$ in order to compensate for the additional summation over the $(\ell'-1)$-th-layer neural indices in the second term as compared to the first. Even so, the layer sum in (8.7) makes this expression somewhat unwieldy and suggests that we should search for an alternate representation.

Following our analysis of the preactivations, let's try to find a recursive expression. To that end, consider the $(\ell+1)$-th-layer NTK, $\widehat{H}^{(\ell+1)}_{i_1 i_2;\alpha_1\alpha_2}$, and decompose the sum over layers in its definition by separating the $(\ell+1)$-th-layer term from all of the rest, giving

$$\begin{aligned}\widehat{H}^{(\ell+1)}_{i_1 i_2;\alpha_1\alpha_2} = &\sum_{j=1}^{n_{\ell+1}}\left(\lambda_b^{(\ell+1)}\frac{dz^{(\ell+1)}_{i_1;\alpha_1}}{db_j^{(\ell+1)}}\frac{dz^{(\ell+1)}_{i_2;\alpha_2}}{db_j^{(\ell+1)}} + \frac{\lambda_W^{(\ell+1)}}{n_\ell}\sum_{k=1}^{n_\ell}\frac{dz^{(\ell+1)}_{i_1;\alpha_1}}{dW_{jk}^{(\ell+1)}}\frac{dz^{(\ell+1)}_{i_2;\alpha_2}}{dW_{jk}^{(\ell+1)}}\right)\\ &+\sum_{j_1,j_2=1}^{n_\ell}\frac{dz^{(\ell+1)}_{i_1;\alpha_1}}{dz^{(\ell)}_{j_1;\alpha_1}}\frac{dz^{(\ell+1)}_{i_2;\alpha_2}}{dz^{(\ell)}_{j_2;\alpha_2}}\widehat{H}^{(\ell)}_{j_1 j_2;\alpha_1\alpha_2}.\end{aligned} \tag{8.8}$$

Here, the first line is the $(\ell+1)$-th-layer term that we left alone, while the second line gives the terms from all the other layers after applying the chain rule and then recalling the definition (8.7). In this way, the ℓ-th-layer NTK appears naturally. This means that we can find a simple iterative expression for the NTK, similar in spirit to the forward equation for the preactivations that defines the MLP.

To finish our derivation, we need to evaluate the derivatives in (8.8). To do so, recall the preactivation forward iteration equation

$$z^{(\ell+1)}_{i;\alpha} = b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)}\sigma^{(\ell)}_{j;\alpha} \tag{8.9}$$

and remember that the activations are explicit functions of the preactivations $\sigma^{(\ell)}_{i;\alpha} \equiv \sigma\left(z^{(\ell)}_{i;\alpha}\right)$. The factors in the second line of (8.8) coming from the chain rule evaluate to

$$\frac{dz^{(\ell+1)}_{i;\alpha}}{dz^{(\ell)}_{j;\alpha}} = W_{ij}^{(\ell+1)}\sigma'^{(\ell)}_{j;\alpha}, \tag{8.10}$$

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

while the derivatives with respect to the $(\ell+1)$-th-layer parameters evaluate to

$$\frac{dz^{(\ell+1)}_{i;\alpha}}{db^{(\ell+1)}_{j}} = \delta_{ij}\,, \qquad \frac{dz^{(\ell+1)}_{i;\alpha}}{dW^{(\ell+1)}_{jk}} = \delta_{ij}\,\sigma^{(\ell)}_{k;\alpha}. \tag{8.11}$$

All together, we can rewrite (8.8) as

$$\begin{aligned}\widehat{H}^{(\ell+1)}_{i_1i_2;\alpha_1\alpha_2} = {} & \delta_{i_1i_2}\left[\lambda^{(\ell+1)}_{b} + \lambda^{(\ell+1)}_{W}\left(\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right)\right] \\ & + \sum_{j_1,j_2=1}^{n_\ell} W^{(\ell+1)}_{i_1j_1}W^{(\ell+1)}_{i_2j_2}\,\sigma'^{(\ell)}_{j_1;\alpha_1}\sigma'^{(\ell)}_{j_2;\alpha_2}\,\widehat{H}^{(\ell)}_{j_1j_2;\alpha_1\alpha_2}.\end{aligned} \tag{8.12}$$

This is the **forward equation for the NTK**, which is an iteration equation that computes the NTK layer by layer for any realization of the biases and weights. This is analogous to the way in which (8.9) computes the network output – as well as all the hidden-layer preactivations – via a layer-to-layer iteration for a given realization of model parameters.

Scaling in the Effective Theory

The forward equation (8.12) further clarifies our decomposition (8.5) in which we made a distinction between the learning rates for the biases and those for the weights, giving each a different scaling with respect to the layer widths n_ℓ of the network.[2]

To see why, first recall from §7 that the change in the training loss after a step of gradient descent is proportional to the product of the global learning rate η and the final-layer NTK $\widehat{H}^{(L)}_{i_1i_2;\alpha_1\alpha_2}$:

$$\Delta\mathcal{L}_{\mathcal{A}} = -\eta\sum_{i_1,i_2=1}^{n_L}\sum_{\alpha_1,\alpha_2\in\mathcal{A}}\epsilon_{i_1;\alpha_1}\epsilon_{i_2;\alpha_2}\,\widehat{H}^{(L)}_{i_1i_2;\alpha_1\alpha_2} + O(\eta^2)\,, \tag{8.13}$$

where here we also recall the definition of the error factor

$$\epsilon_{i;\alpha} \equiv \frac{\partial\mathcal{L}_{\mathcal{A}}}{\partial z^{(L)}_{i;\alpha}}. \tag{8.14}$$

Note that this error factor generally stays of order one in the large-width limit, cf. the explicit expression when using the MSE loss (7.15). Thus, it's essential that the product of the global learning rate and the NTK, $\eta\widehat{H}^{(L)}$, also stays of order one for large-width networks: if it diverged as the width increases, then the higher-order terms in (8.13) would dominate, and the loss would no longer be guaranteed to decrease; if instead it vanished in this limit, then no training would take place. Either way, training would fail.

[2] You'll have to wait until §9 to understand why it is advantageous to give a layer dependence to $\lambda^{(\ell)}_b$ and $\lambda^{(\ell)}_W$ and to learn how they should be scaled with depth.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

With that in mind, we chose the width scaling of our learning-rate tensor so that the NTK naturally stays of order one in the large-width limit and hence a (sufficiently small but not parametrically small) order-one global learning η ensures the success of training. In particular, the $(\ell+1)$-th-layer contribution in the first line of the forward equation (8.12) stays of order one if we take $\lambda_b^{(\ell+1)}, \lambda_W^{(\ell+1)} = O(1)$, with the $1/n_\ell$ normalization of the $(\ell+1)$-th-layer weight learning rate playing an essential role in compensating for the summation over the n_ℓ terms.[3]

If instead we had considered the original version of gradient descent with $\lambda_{\mu\nu} = \delta_{\mu\nu}$ rather than *tensorial gradient descent*, we would have been in trouble. In the language of our effective theory, the original gradient descent corresponds to setting $\lambda_b^{(\ell)} = 1$ and $\lambda_W^{(\ell)} = n_{\ell-1}$, which means that the NTK itself would be $O(n)$. We'd then have to scale the global learning rate as $\eta = O(1/n)$ to compensate for this $O(n)$ scaling of the NTK. However, since in this case $\eta\lambda_b^{(\ell)} = O(1/n)$, the order-one contribution from the weights to the NTK would completely overwhelm the $1/n$-suppressed contribution from the biases. This would lead to a lack of appropriate contribution of the biases to the updates of the weights as well as an extreme under-training of the biases themselves.

Finally, let's make a general point: in any effective theory, it's really essential to make all large or small scales explicit – and rescale hyperparameters accordingly – as we did earlier for the variance of the weight initialization distribution, did here for the weight learning rate, and will do later for the depth scaling of both the bias and weight learning rates. For the effective theorist this ensures that the asymptotic $1/n$ and $1/\ell$ expansions are sound and nontrivial, and for the practical practitioner this enables comparisons of hyperparameter values across architectures with different widths and depths. In particular, we expect very generally that this should help mitigate expensive hyperparameter tuning, remove the need for heuristic fixes, and increase the robustness of optimal hyperparameter settings when scaling a model up.

Getting Things Backwards

N.B. the chain-rule factors (8.10) also appear when evaluating the derivative of the network outputs with respect to model parameters:

[3] With this choice, the recursive term in the second line of the forward equation (8.12) also stays of order one. To see this, let's evaluate its expectation:

$$\mathbb{E}\left[\sum_{j_1,j_2=1}^{n_\ell} W_{i_1j_1}^{(\ell+1)} W_{i_2j_2}^{(\ell+1)} \sigma'^{(\ell)}_{j_1;\alpha_1} \sigma'^{(\ell)}_{j_2;\alpha_2} \widehat{H}^{(\ell)}_{j_1j_2;\alpha_1\alpha_2}\right] = \sum_{j_1,j_2=1}^{n_\ell} \mathbb{E}\left[W_{i_1j_1}^{(\ell+1)} W_{i_2j_2}^{(\ell+1)}\right] \mathbb{E}\left[\sigma'^{(\ell)}_{j_1;\alpha_1} \sigma'^{(\ell)}_{j_2;\alpha_2} \widehat{H}^{(\ell)}_{j_1j_2;\alpha_1\alpha_2}\right]$$
$$= \delta_{i_1i_2}\, C_W^{(\ell+1)} \left(\frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1} \sigma'^{(\ell)}_{j;\alpha_2} \widehat{H}^{(\ell)}_{jj;\alpha_1\alpha_2}\right]\right). \tag{8.15}$$

In particular, we see that the $1/n_\ell$ scaling of the initialization weight variance $C_W^{(\ell+1)}$ is important for ensuring principled behavior of not only the network output, but also the NTK.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

$$\frac{dz_{i;\alpha}^{(L)}}{db_j^{(\ell)}} = \frac{dz_{i;\alpha}^{(L)}}{dz_{j;\alpha}^{(\ell)}}\,, \qquad \frac{dz_{i;\alpha}^{(L)}}{dW_{jk}^{(\ell)}} = \sum_m \frac{dz_{i;\alpha}^{(L)}}{dz_{m;\alpha}^{(\ell)}} \frac{dz_{m;\alpha}^{(\ell)}}{dW_{jk}^{(\ell)}} = \frac{dz_{i;\alpha}^{(L)}}{dz_{j;\alpha}^{(\ell)}} \sigma_{k;\alpha}^{(\ell-1)}. \tag{8.16}$$

Evaluating these derivatives gives another neural-network iteration equation,

$$\frac{dz_{i;\alpha}^{(L)}}{dz_{j;\alpha}^{(\ell)}} = \sum_{k=1}^{n_{\ell+1}} \frac{dz_{i;\alpha}^{(L)}}{dz_{k;\alpha}^{(\ell+1)}} \frac{dz_{k;\alpha}^{(\ell+1)}}{dz_{j;\alpha}^{(\ell)}} = \sum_{k=1}^{n_{\ell+1}} \frac{dz_{i;\alpha}^{(L)}}{dz_{k;\alpha}^{(\ell+1)}} W_{kj}^{(\ell+1)} \sigma'^{(\ell)}_{j;\alpha} \quad \text{for} \quad \ell < L, \tag{8.17}$$

but in this case for the derivative of the output. In particular, (8.17) is a *backward* equation: starting from the *final* condition

$$\frac{dz_{i;\alpha}^{(L)}}{dz_{j;\alpha}^{(L)}} = \delta_{ij}, \tag{8.18}$$

we iterate layer to layer backwards, $\ell = L-1, L-2, \ldots, 1$, by sequential multiplications of the chain-rule factors (8.10).

An algorithm based on this backward equation can be efficiently implemented to compute derivatives with respect to the model parameters and, for that reason, is used by most deep-learning packages to compute the gradient as part of any neural network gradient-based learning algorithm. Such a package typically lets practitioners specify a deep learning model by defining a **forward pass** – for MLPs a practitioner would implement the forward equation (8.9) – and then the package will automatically work out the **backward pass** – i.e., for MLPs it would implement (8.17). The computational algorithm based on (8.17) is termed **backpropagation**, which was discovered and rediscovered numerous times in the history of deep learning. Among these, a particular rediscovery [15] was essential in convincing the machine learning community that multilayer neural networks can be trained efficiently.

All that said, when evaluating the NTK in the effective theory, it's essential that we use the forward equation (8.12) rather than getting things backwards. In the next three-plus-one sections, we'll indeed use the forward equation to recursively compute the *joint* initialization distribution for the ℓ-th-layer preactivations *and* the ℓ-th-layer NTK:

$$p\Big(z^{(\ell)}, \widehat{H}^{(\ell)}\Big|\mathcal{D}\Big) \equiv p\begin{pmatrix} z^{(\ell)}(x_1) & z^{(\ell)}(x_2) & \ldots & z^{(\ell)}(x_{N_\mathcal{D}}) \\ \widehat{H}^{(\ell)}(x_1,x_1) & \widehat{H}^{(\ell)}(x_1,x_2) & \ldots & \widehat{H}^{(\ell)}(x_1,x_{N_\mathcal{D}}) \\ \widehat{H}^{(\ell)}(x_2,x_1) & \widehat{H}^{(\ell)}(x_2,x_2) & \ldots & \widehat{H}^{(\ell)}(x_2,x_{N_\mathcal{D}}) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{H}^{(\ell)}(x_{N_\mathcal{D}},x_1) & \widehat{H}^{(\ell)}(x_{N_\mathcal{D}},x_2) & \ldots & \widehat{H}^{(\ell)}(x_{N_\mathcal{D}},x_{N_\mathcal{D}}) \end{pmatrix}. \tag{8.19}$$

(On the right-hand side, we've suppressed neural indices while explicitly writing out the input dependence. This emphasizes that the preactivations are each functions of a single input and that the NTK components are each functions of a pair of inputs.)

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

8.1 First Layer: Deterministic NTK

Recall from §4.1 that at initialization the first-layer preactivations,

$$z_{i;\alpha}^{(1)} \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_{j;\alpha} \, , \tag{8.20}$$

are distributed according to a zero-mean Gaussian distribution,

$$p\Big(z^{(1)}\Big|\mathcal{D}\Big) = \frac{1}{\left|2\pi G^{(1)}\right|^{\frac{n_1}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n_1}\sum_{\alpha_1,\alpha_2\in\mathcal{D}} G_{(1)}^{\alpha_1\alpha_2} z_{i;\alpha_1}^{(1)} z_{i;\alpha_2}^{(1)}\right) , \tag{8.21}$$

with the first-layer deterministic metric – a function of the inputs – given by

$$G_{\alpha_1\alpha_2}^{(1)} \equiv C_b^{(1)} + C_W^{(1)} \frac{1}{n_0}\sum_{j=1}^{n_0} x_{j;\alpha_1} x_{j;\alpha_2} . \tag{8.22}$$

In particular, the quadratic action in the exponent of (8.21) indicates the absence of interactions between neurons. This enables us to factor expectation values of first-layer observables into separate Gaussian integrals for each neuron.

The first-layer NTK at initialization is even more trivial and can be read off from the original definition of the NTK (8.7) by plugging in the derivatives (8.11) and remembering the identification $\sigma_{i;\alpha}^{(0)} = x_{i;\alpha}$:

$$\widehat{H}_{i_1 i_2;\alpha_1\alpha_2}^{(1)} = \delta_{i_1 i_2}\left[\lambda_b^{(1)} + \lambda_W^{(1)}\left(\frac{1}{n_0}\sum_{j=1}^{n_0} x_{j;\alpha_1} x_{j;\alpha_2}\right)\right] \equiv \delta_{i_1 i_2} H_{\alpha_1\alpha_2}^{(1)} . \tag{8.23}$$

Like the first-layer metric, the first-layer NTK is completely deterministic – hence no hat on the right-hand side of the equation – and is diagonal in its neural indices. Remembering our exposition on the off-diagonal components of the NTK in §7.2, this in particular means that, for single-layer networks, a feature captured by a particular neuron cannot affect the gradient-descent update for another feature on any other neuron.

Finally, recalling our discussion of deterministic distributions in §2.3, the joint distribution of the first-layer preactivations and the first-layer NTK can be written as

$$p\Big(z^{(1)}, \widehat{H}^{(1)}\Big|\mathcal{D}\Big) = p\Big(z^{(1)}\Big|\mathcal{D}\Big) \prod_{(i_1 i_2),(\alpha_1\alpha_2)} \delta\Big(\widehat{H}_{i_1 i_2;\alpha_1\alpha_2}^{(1)} - \delta_{i_1 i_2} H_{\alpha_1\alpha_2}^{(1)}\Big) , \tag{8.24}$$

where the product of the Dirac delta functions runs over all pairs of neural indices and sample indices. Just as the first-layer preactivation distribution was representative of deeper layers in the infinite-width limit, this first-layer joint distribution is also representative of deeper-layer joint distributions in the infinite-width limit: the preactivation distribution is exactly Gaussian, the NTK distribution is completely deterministic, and there is no correlation between the two, i.e., they are statistically independent from each other once the dataset is fixed.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

8.2 Second Layer: Fluctuating NTK

Now, let us see how finite-width corrections can modify this picture in the second layer.

Recall from §4.2 that the second-layer preactivations are given by

$$z^{(2)}_{i;\alpha} = b^{(2)}_i + \sum_{j=1}^{n_1} W^{(2)}_{ij} \sigma^{(1)}_{j;\alpha}. \tag{8.25}$$

After marginalizing over the first-layer preactivations $z^{(1)}$, the correlated fluctuations of the preactivations in the first layer resulted in a nontrivial interaction between different neurons in the second layer. At the leading nontrivial order in $1/n_1$, this led to a nearly-Gaussian distribution with a quartic action (4.60) for the second-layer preactivations, with the leading non-Gaussianity captured by a nonzero connected four-point correlator.

As for the NTK, looking at its forward equation (8.12) and recalling that the first-layer NTK is deterministic, (8.23), we see that the second-layer NTK is given by

$$\widehat{H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2} = \delta_{i_1 i_2}\left[\lambda^{(2)}_b + \lambda^{(2)}_W \left(\frac{1}{n_1}\sum_{j=1}^{n_1} \sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2}\right)\right] + \sum_{j=1}^{n_1} W^{(2)}_{i_1 j} W^{(2)}_{i_2 j}\, \sigma'^{(1)}_{j;\alpha_1}\sigma'^{(1)}_{j;\alpha_2}\, H^{(1)}_{\alpha_1\alpha_2}. \tag{8.26}$$

This second-layer NTK depends on two sets of stochastic variables, the weights $W^{(2)}_{ij}$ and the first-layer preactivations $z^{(1)}_{i;\alpha}$, and hence it fluctuates.

To compute its mean we take an expectation of (8.26), finding

$$\begin{aligned} &\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2}\right] \qquad (8.27)\\ &= \delta_{i_1 i_2}\left[\lambda^{(2)}_b + \lambda^{(2)}_W \left(\frac{1}{n_1}\sum_{j=1}^{n_1} \mathbb{E}\left[\sigma^{(1)}_{j;\alpha_1}\sigma^{(1)}_{j;\alpha_2}\right]\right)\right] + \sum_{j=1}^{n_1} \mathbb{E}\left[W^{(2)}_{i_1 j} W^{(2)}_{i_2 j}\right] \mathbb{E}\left[\sigma'^{(1)}_{j;\alpha_1}\sigma'^{(1)}_{j;\alpha_2}\right] H^{(1)}_{\alpha_1\alpha_2}\\ &= \delta_{i_1 i_2}\left[\lambda^{(2)}_b + \lambda^{(2)}_W \left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(1)}} + C^{(2)}_W \left\langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\right\rangle_{G^{(1)}} H^{(1)}_{\alpha_1\alpha_2}\right]\\ &\equiv \delta_{i_1 i_2} H^{(2)}_{\alpha_1\alpha_2}. \end{aligned}$$

Here, in the second line, the expectation of the recursive term factorized because the second-layer weights $W^{(2)}_{ij}$ are statistically independent from the first-layer preactivations. Additionally, in the third line we recalled (4.27), in which we showed that the two-point correlators can be expressed as separate Gaussian expectations for each neuron, with the variance given by the first-layer metric $G^{(1)}$.[4] Further, inspecting our answer (8.27), we see that the mean of the second-layer NTK is diagonal in its neural indices.

[4] Note that the logic around (4.27) is the same whether or not the Gaussian expectation is of activations or derivatives of the activation. In other words, for the first-layer preactivations we also have $\mathbb{E}\left[\sigma'^{(1)}_{j;\alpha_1}\sigma'^{(1)}_{j;\alpha_2}\right] = \left\langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\right\rangle_{G^{(1)}}$.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

Furthermore, we separated the part that encodes the sample dependence and symbolized it by taking off its hat because it is a mean, not a stochastic variable.

Now, let's compute the variance. First, define the second-layer NTK fluctuation through our usual decomposition,

$$\widehat{H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2} \equiv \delta_{i_1 i_2} H^{(2)}_{\alpha_1\alpha_2} + \widehat{\Delta H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2}, \tag{8.28}$$

so that the expectation of the magnitude of this fluctuation determines the covariance:

$$\mathbb{E}\left[\widehat{\Delta H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2}\widehat{\Delta H}^{(2)}_{i_3 i_4;\alpha_3\alpha_4}\right] = \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2}\widehat{H}^{(2)}_{i_3 i_4;\alpha_3\alpha_4}\right] - \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2}\right]\mathbb{E}\left[\widehat{H}^{(2)}_{i_3 i_4;\alpha_3\alpha_4}\right]. \tag{8.29}$$

Substituting in our expression (8.26) for the second-layer stochastic NTK and using the independence of the second-layer weights from the first-layer preactivations, we find a complicated-looking result,

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\Delta H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2}\widehat{\Delta H}^{(2)}_{i_3 i_4;\alpha_3\alpha_4}\right] \qquad (8.30)\\
&= \frac{1}{n_1}\delta_{i_1 i_2}\delta_{i_3 i_4}\Bigg\{\left(\lambda^{(2)}_W\right)^2\left[\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}} - \langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}}\right]\\
&\qquad + C^{(2)}_W H^{(1)}_{\alpha_1\alpha_2}\lambda^{(2)}_W\left[\langle\sigma'_{\alpha_1}\sigma'_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}} - \langle\sigma'_{\alpha_1}\sigma'_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}}\right]\\
&\qquad + \lambda^{(2)}_W C^{(2)}_W H^{(1)}_{\alpha_3\alpha_4}\left[\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}} - \langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}}\right]\\
&\qquad + \left(C^{(2)}_W\right)^2 H^{(1)}_{\alpha_1\alpha_2}H^{(1)}_{\alpha_3\alpha_4}\left[\langle\sigma'_{\alpha_1}\sigma'_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}} - \langle\sigma'_{\alpha_1}\sigma'_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}}\right]\Bigg\}\\
&\quad + \frac{1}{n_1}\left(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)\left(C^{(2)}_W\right)^2 H^{(1)}_{\alpha_1\alpha_2}H^{(1)}_{\alpha_3\alpha_4}\langle\sigma'_{\alpha_1}\sigma'_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}}.
\end{aligned}$$

To get this expression, we recalled not only (4.27) for the two-point correlators but also both (4.28) and (4.29) for the different pairings of the four-point correlators, with the pairings depending on whether all activations are on the same neuron or are on two different neurons, respectively. As with our computation of the mean above, the computations of these four-point correlators proceed similarly regardless of whether an activation has a derivative or not.

To help make sense of this rather ugly expression (8.30), let's first decompose the second-layer NTK variance into a sum of two different types of tensors,

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\Delta H}^{(2)}_{i_1 i_2;\alpha_1\alpha_2}\widehat{\Delta H}^{(2)}_{i_3 i_4;\alpha_3\alpha_4}\right] \qquad (8.31)\\
&\equiv \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}A^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4}B^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}B^{(2)}_{\alpha_1\alpha_4\alpha_2\alpha_3}\right].
\end{aligned}$$

This decomposition was motivated by the pattern of Kronecker deltas that appear in (8.30). Next, by comparing this to our original expression (8.30), we see that these tensors are given by

$$A^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \left\langle \widehat{\Omega}^{(2)}_{\alpha_1\alpha_2}\widehat{\Omega}^{(2)}_{\alpha_3\alpha_4} \right\rangle_{G^{(1)}} - \left\langle \widehat{\Omega}^{(2)}_{\alpha_1\alpha_2} \right\rangle_{G^{(1)}} \left\langle \widehat{\Omega}^{(2)}_{\alpha_3\alpha_4} \right\rangle_{G^{(1)}}, \tag{8.32}$$

$$B^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4} = \left(C^{(2)}_W\right)^2 H^{(1)}_{\alpha_1\alpha_2} H^{(1)}_{\alpha_3\alpha_4} \left\langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4} \right\rangle_{G^{(1)}}, \tag{8.33}$$

where on the first line we've introduced an auxiliary stochastic variable,

$$\widehat{\Omega}^{(2)}_{\alpha_1\alpha_2} \equiv \lambda^{(2)}_W \, \sigma^{(1)}_{\alpha_1}\sigma^{(1)}_{\alpha_2} + C^{(2)}_W H^{(1)}_{\alpha_1\alpha_2}\, \sigma'^{(1)}_{\alpha_1}\sigma'^{(1)}_{\alpha_2}, \tag{8.34}$$

in order to remedy the ugliness of what would have otherwise been a very long expression.[5] From (8.32) and (8.33) we see clearly that these tensors are of order one. Given (8.31), this in turn means that the second-layer NTK variance is suppressed by $1/n_1$ in the large-width limit. In other words, the second-layer NTK is deterministic in the strict infinite-width limit, but in backing off that limit it fluctuates according to (8.31), (8.32), and (8.33).

Moreover, at finite width the second-layer NTK not only fluctuates but also has nontrivial cross correlation with the second-layer preactivations. This can ultimately be traced to the fact that the second-layer preactivations (8.25) and the second-layer NTK (8.26) are both functions of the same stochastic variables: the second-layer weights $W^{(2)}_{ij}$ and the first-layer preactivations $z^{(1)}_{i;\alpha}$.

This cross correlation can be computed analogously to the way we computed the NTK mean and variance. Substituting in the definition of the second-layer preactivations (8.25) and the second-layer NTK (8.26), and again using the statistical independence of the second-layer weights $W^{(2)}_{ij}$ from the first-layer preactivations $z^{(1)}_{i;\alpha}$, we find

$$\mathbb{E}\left[z^{(2)}_{i_1;\alpha_1}\widehat{\Delta H}^{(2)}_{i_2i_3;\alpha_2\alpha_3}\right] = 0, \tag{8.35}$$

$$\begin{aligned}\mathbb{E}\left[z^{(2)}_{i_1;\alpha_1} z^{(2)}_{i_2;\alpha_2}\widehat{\Delta H}^{(2)}_{i_3i_4;\alpha_3\alpha_4}\right] &= \mathbb{E}\left[z^{(2)}_{i_1;\alpha_1} z^{(2)}_{i_2;\alpha_2}\widehat{H}^{(2)}_{i_3i_4;\alpha_3\alpha_4}\right] - \mathbb{E}\left[z^{(2)}_{i_1;\alpha_1} z^{(2)}_{i_2;\alpha_2}\right]\mathbb{E}\left[\widehat{H}^{(2)}_{i_3i_4;\alpha_3\alpha_4}\right] \\ &= \frac{1}{n_1}\delta_{i_1i_2}\delta_{i_3i_4}\Big\{\lambda^{(2)}_W C^{(2)}_W \Big[\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}} - \langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma_{\alpha_3}\sigma_{\alpha_4}\rangle_{G^{(1)}}\Big]\end{aligned}$$

[5] Note that $A^{(2)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ (8.32) has the same symmetries as the four-point vertex $V^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$. In particular, it's symmetric under exchanges of sample indices $\alpha_1 \leftrightarrow \alpha_2$, $\alpha_3 \leftrightarrow \alpha_4$, and $(\alpha_1\alpha_2) \leftrightarrow (\alpha_3\alpha_4)$, and this symmetry persists to deeper layers, cf. (8.97).

Now, you might raise your hand and say that $B^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4}$ (8.33) also respects the same symmetry. That's correct here, but this symmetry will be broken in deeper layers. In general – cf. (8.89) – $B^{(\ell)}_{\alpha_1\alpha_3\alpha_2\alpha_4}$ will be symmetric under $(\alpha_1\alpha_2) \leftrightarrow (\alpha_3\alpha_4)$ and $(\alpha_1\alpha_3) \leftrightarrow (\alpha_2\alpha_4)$ but *not* under $\alpha_1 \leftrightarrow \alpha_2$ or $\alpha_3 \leftrightarrow \alpha_4$ individually.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

$$+\left(C_W^{(2)}\right)^2 H^{(1)}_{\alpha_3\alpha_4}\Big[\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}}-\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\langle\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}}\Big]\Bigg\}$$
$$+\frac{1}{n_1}\left(\delta_{i_1i_3}\delta_{i_2i_4}+\delta_{i_1i_4}\delta_{i_2i_3}\right)\left(C_W^{(2)}\right)^2 H^{(1)}_{\alpha_3\alpha_4}\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}}\,. \tag{8.36}$$

Here, as for the variance (8.30), we recalled the suitably generalized versions of (4.27), (4.28), and (4.29) for the two- and four-point correlators. Thus, we see that the first measure of cross correlation between the second-layer preactivations and the second-layer NTK (8.35) vanishes, but the second one (8.36) is nonzero at finite width.

To aid us in our deep-layer analysis, it will be convenient to decompose this cross correlation (8.36) into two tensors with sample indices only, just as we did for the variance in (8.31):

$$\mathbb{E}\left[z^{(2)}_{i_1;\alpha_1}z^{(2)}_{i_2;\alpha_2}\widehat{\Delta H}^{(2)}_{i_3i_4;\alpha_3\alpha_4}\right] \tag{8.37}$$
$$=\frac{1}{n_1}\left[\delta_{i_1i_2}\delta_{i_3i_4}D^{(2)}_{\alpha_1\alpha_2\alpha_3\alpha_4}+\delta_{i_1i_3}\delta_{i_2i_4}F^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4}+\delta_{i_1i_4}\delta_{i_2i_3}F^{(2)}_{\alpha_1\alpha_4\alpha_2\alpha_3}\right].$$

Comparing this decomposition with our explicit formula for the correlator (8.36), we can identify expressions for these tensors:

$$D^{(2)}_{\alpha_1\alpha_2\alpha_3\alpha_4}=C_W^{(2)}\left[\left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\widehat{\Omega}^{(2)}_{\alpha_3\alpha_4}\right\rangle_{G^{(1)}}-\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(1)}}\left\langle\widehat{\Omega}^{(2)}_{\alpha_3\alpha_4}\right\rangle_{G^{(1)}}\right], \tag{8.38}$$
$$F^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4}=\left(C_W^{(2)}\right)^2 H^{(1)}_{\alpha_3\alpha_4}\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(1)}}\,, \tag{8.39}$$

where we've also recalled the stochastic tensor $\widehat{\Omega}^{(2)}_{\alpha_1\alpha_2}$ defined in (8.34).[6] Just as for $A^{(2)}$ and $B^{(2)}$ above, both $D^{(2)}$ and $F^{(2)}$ are manifestly of order one. Similar to the second-layer NTK variance, this means that the cross correlator (8.37) is suppressed by $1/n_1$ in the large-width limit, vanishing in the strict infinite-width limit.

In summary, the joint distribution of the second-layer preactivations and second-layer NTK,

$$p\Big(z^{(2)},\widehat{H}^{(2)}\Big|\mathcal{D}\Big)\,, \tag{8.40}$$

at leading nontrivial order in the $1/n$ expansion, is a *nearly-Gaussian distribution* with *(i)* a quartic interaction among preactivations on different neurons, *(ii)* a fluctuating NTK, and *(iii)* cross correlation between the preactivations and NTK. All of these finite-width effects become more complicated for deeper layers.

[6] The cross-correlation tensor $D^{(2)}_{\alpha_1\alpha_2\alpha_3\alpha_4}$ (8.38) – and more generally $D^{(\ell)}_{\alpha_1\alpha_2\alpha_3\alpha_4}$ in deeper layers, cf. (8.77) – is symmetric under exchanges of sample indices $\alpha_1\leftrightarrow\alpha_2$ and $\alpha_3\leftrightarrow\alpha_4$, but of course *not* under $(\alpha_1\alpha_2)\leftrightarrow(\alpha_3\alpha_4)$. The other tensor $F^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4}$ (8.39) respects this same symmetry in the second layer but has no symmetry at all in deeper layers, cf. (8.79).

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

8.3 Deeper Layers: Accumulation of NTK Fluctuations

As before with §4.1 ‖ §8.1 and §4.2 ‖ §8.2, this section parallels §4.3. In §4.3, we investigated the nearly-Gaussian distribution of preactivations $p\Big(z^{(\ell+1)}\big|\mathcal{D}\Big)$ at finite width by considering an interlayer joint distribution $p\Big(z^{(\ell+1)}, z^{(\ell)}\big|\mathcal{D}\Big)$ and then integrating out the ℓ-th-layer preactivations. In particular, due to correlated dependence on the preactivations in previous layers, the non-Gaussianity in the preactivation distribution accumulated as depth increased, manifesting itself in the running four-point vertex $V^{(\ell)}$.

This same mechanism makes the NTK fluctuations accumulate, amplifying the NTK variance as well as the cross correlation between the NTK and preactivations. In this section, we will derive recursions for the NTK mean, the NTK-preactivation cross correlation, and the NTK variance that together determine the ℓ-th-layer joint distribution at leading nontrivial order in $1/n$. What follows is a *goode olde calculation*, so please sharpen your quills, unfurl your parchment, and inform your majordomo that you require a cleared schedule for the rest of the day.

8.3.0 *Inter*lude: *Inter*layer Correlations

If you have an eidetic memory, then perhaps you recall that the main complication with our derivation of the general $(\ell+1)$-th-layer preactivation statistics – as compared to the second-layer statistics – was that the ℓ-th-layer preactivation distribution $p\Big(z^{(\ell)}\big|\mathcal{D}\Big)$ was also non-Gaussian, unlike the Gaussian preactivation distribution in the first layer. For such a nearly-Gaussian distribution $p\Big(z^{(\ell)}\big|\mathcal{D}\Big)$, *interactions* imply a nontrivial **intralayer correlation** between observables of the preactivations across different neurons $i_1 \neq i_2$. Specifically, the covariance of two arbitrary single-neuron functions $\mathcal{F}\Big(z^{(\ell)}_{i_1;\mathcal{A}_1}\Big)$ and $\mathcal{G}\Big(z^{(\ell)}_{i_2;\mathcal{A}_2}\Big)$ depending on data subsamples $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{D}$, respectively, is given by (4.64) and reprinted here:

$$\begin{aligned}
&\mathrm{Cov}\Big[\mathcal{F}\Big(z^{(\ell)}_{i_1;\mathcal{A}_1}\Big), \mathcal{G}\Big(z^{(\ell)}_{i_2;\mathcal{A}_2}\Big)\Big] \equiv \mathbb{E}\Big[\mathcal{F}\Big(z^{(\ell)}_{i_1;\mathcal{A}_1}\Big)\, \mathcal{G}\Big(z^{(\ell)}_{i_2;\mathcal{A}_2}\Big)\Big] - \mathbb{E}\Big[\mathcal{F}\Big(z^{(\ell)}_{i_1;\mathcal{A}_1}\Big)\Big]\, \mathbb{E}\Big[\mathcal{G}\Big(z^{(\ell)}_{i_2;\mathcal{A}_2}\Big)\Big] \\
&= \sum_{\beta_1,\ldots,\beta_4\in\mathcal{D}} \frac{1}{4n_{\ell-1}} V_{(\ell)}^{(\beta_1\beta_2)(\beta_3\beta_4)} \Big\langle \Big(z_{\beta_1}z_{\beta_2} - G^{(\ell)}_{\beta_1\beta_2}\Big)\, \mathcal{F}(z_{\mathcal{A}_1})\Big\rangle_{G^{(\ell)}} \Big\langle \Big(z_{\beta_3}z_{\beta_4} - G^{(\ell)}_{\beta_3\beta_4}\Big)\, \mathcal{G}(z_{\mathcal{A}_2})\Big\rangle_{G^{(\ell)}} \\
&\quad + O\!\left(\frac{1}{n^2}\right). \qquad (8.41)
\end{aligned}$$

In this reprinting, we substituted in our leading large-width expressions (4.81) and (4.82) for the quadratic coupling $g_{(\ell)}$ and quartic coupling $v_{(\ell)}$, respectively. We have also recalled our long-forgotten shorthand notation for the covariance of random variables (1.53), which we will use judiciously throughout this section. This *intralayer* formula will soon prove itself useful.

Enlarging our view to the preactivation–NTK joint distribution (8.19), we'll encounter another complication due to ***inter*layer correlation** of the form

$$\mathbb{E}\Big[\mathcal{O}\Big(z^{(\ell+1)}\Big)\, \mathcal{P}\Big(W^{(\ell+1)}\Big)\, \mathcal{Q}\Big(z^{(\ell)}, \widehat{H}^{(\ell)}\Big)\Big], \qquad (8.42)$$

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

where $\mathcal{O}$ is some function of $(\ell+1)$-th-layer preactivations, $\mathcal{P}$ is a polynomial of $(\ell+1)$-th-layer weights, and $\mathcal{Q}$ is a function of ℓ-th-layer preactivations and the ℓ-th-layer NTK. For instance, taking the NTK–preactivation cross correlation

$$\mathbb{E}\left[z^{(\ell+1)}_{i_1;\alpha_1} z^{(\ell+1)}_{i_2;\alpha_2} \widehat{H}^{(\ell+1)}_{i_3 i_4;\alpha_3\alpha_4}\right] \tag{8.43}$$

and unraveling the NTK through its forward equation (8.12), we get an interlayer correlation of the form (8.42), with $\mathcal{P}=1$ from the additive term in the square brackets, and an interlayer correlation of the same form with $\mathcal{P}=W^{(\ell+1)}_{i_3 j_3} W^{(\ell+1)}_{i_4 j_4}$ from the recursive term. While it was simple enough to evaluate such an expectation for the second layer, it's somewhat subtle for a general layer.

That said, there's actually a pretty neat trick that lets us reduce such interlayer correlations (8.42) to expectations of solely ℓ-th-layer variables. Such expectations can subsequently be evaluated with the *intra*layer formula (8.41) above. Let us now teach you this magic trick before diving deep into learning the deeper-layer analysis.[7]

First, using the definition of the expectation and the conditional structure of the distribution, the interlayer correlation (8.42) can be expressed as (suppressing all indices)

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{O}\Big(z^{(\ell+1)}\Big)\, \mathcal{P}\Big(W^{(\ell+1)}\Big)\, \mathcal{Q}\Big(z^{(\ell)}, \widehat{H}^{(\ell)}\Big)\right] \\
&= \int dz^{(\ell)} d\widehat{H}^{(\ell)} p\Big(z^{(\ell)}, \widehat{H}^{(\ell)} \Big| \mathcal{D}\Big)\, \mathcal{Q}\Big(z^{(\ell)}, \widehat{H}^{(\ell)}\Big) \\
&\quad \times \left[\int db^{(\ell+1)} dW^{(\ell+1)} p\Big(b^{(\ell+1)}\Big)\, p\Big(W^{(\ell+1)}\Big)\, \mathcal{P}\Big(W^{(\ell+1)}\Big)\right. \\
&\quad\quad \left. \times \int dz^{(\ell+1)} p\Big(z^{(\ell+1)} \Big| b^{(\ell+1)}, W^{(\ell+1)}, z^{(\ell)}\Big)\, \mathcal{O}\Big(z^{(\ell+1)}\Big)\right].
\end{aligned} \tag{8.44}$$

Our strategy will be to *integrate out* or marginalize over the $(\ell+1)$-th-layer parameters in order to express the object inside the square bracket as a function of the ℓ-th-layer variables *only*. In this way, the entire object will become an ℓ-th-layer expectation that we already know how to handle.

Second, rather than working with an abstract polynomial $\mathcal{P}$, let's construct a *generating function* for these interlayer correlations through the use of a *source term*:

$$\mathcal{P}\Big(W^{(\ell+1)}\Big) = e^{\sum_{i,j} \mathcal{J}_{ij} W^{(\ell+1)}_{ij}}. \tag{8.45}$$

Recall from your pretraining days (§1.1) that a generating function such as (8.45) could be used to evaluate expectations such as (8.42) with any polynomial insertions of weights $W^{(\ell+1)}_{ij}$. To do this, we differentiate the evaluated generating function some number of times with respect to the source $\mathcal{J}_{ij}$ and then set the source to zero.

[7] As similar interlayer correlations appear in §11, we'll keep our exposition completely general rather than specializing to the NTK–preactivation cross correlation (8.43).

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

Now with our choice (8.45) in mind for $\mathcal{P}$, we can explicitly evaluate the expression in the square brackets in (8.44) as follows: *(i)* recall the initialization distributions for the biases (2.21) and weights (2.22), *(ii)* recall from (2.34) that the conditional distribution $p\left(z^{(\ell+1)}\middle|b^{(\ell+1)}, W^{(\ell+1)}, z^{(\ell)}\right)$ encodes the MLP forward equation (8.9) as a Dirac delta function, and finally *(iii)* recall the integral representation of the Dirac delta function (2.32). All together, this gives the following set of integrals:

$$\begin{aligned}\int &\left[\prod_i \frac{db_i}{\sqrt{2\pi C_b^{(\ell+1)}}}\right]\left[\prod_{i,j}\frac{dW_{ij}}{\sqrt{2\pi C_W^{(\ell+1)}/n_\ell}}\right]\left[\prod_{i,\alpha}\frac{d\Lambda_i{}^{\alpha}\, dz_{i;\alpha}^{(\ell+1)}}{2\pi}\right]\mathcal{O}\left(z^{(\ell+1)}\right) \\ &\times\exp\left[-\sum_i \frac{b_i^2}{2C_b^{(\ell+1)}} - \sum_{i,j}\frac{n_\ell W_{ij}^2}{2C_W^{(\ell+1)}} + i\sum_{i,\alpha}\Lambda_i{}^{\alpha}\left(z_{i;\alpha}^{(\ell+1)} - b_i - \sum_j W_{ij}\,\sigma_{j;\alpha}^{(\ell)}\right)\right. \\ &\qquad\left. + \sum_{i,j}\mathcal{J}_{ij}W_{ij}\right],\end{aligned} \tag{8.46}$$

which we recognize as the good-old *Hubbard–Stratonovich transformation* that we first encountered in §4.1.

Next, as we did in §4.1 and in high school, we can *complete the squares* with respect to the biases and weights and integrate them out. The only substantial deviation here from the presentation in §4.1 is that the source term shifts the linear coupling of the weights as

$$-iW_{ij}\sum_\alpha \Lambda_i{}^{\alpha}\sigma_{j;\alpha}^{(\ell)} \to -iW_{ij}\left(\sum_\alpha \Lambda_i{}^{\alpha}\sigma_{j;\alpha}^{(\ell)} + i\mathcal{J}_{ij}\right). \tag{8.47}$$

Performing these Gaussian integrals, we find

$$\begin{aligned}\int&\left[\prod_{i,\alpha}\frac{d\Lambda_i{}^{\alpha}\, dz_{i;\alpha}^{(\ell+1)}}{2\pi}\right]\mathcal{O}\left(z^{(\ell+1)}\right)\exp\left[-\sum_{i,\alpha_1,\alpha_2}\Lambda_i{}^{\alpha_1}\Lambda_i{}^{\alpha_2}\left(\frac{C_b^{(\ell+1)}}{2} + \frac{C_W^{(\ell+1)}}{2n_\ell}\sum_j \sigma_{\alpha_1;j}^{(\ell)}\sigma_{\alpha_2;j}^{(\ell)}\right)\right. \\ &\left. + i\sum_{i,\alpha}\Lambda_i{}^{\alpha}\left(z_{i;\alpha}^{(\ell+1)} - \frac{C_W^{(\ell+1)}}{n_\ell}\sum_j \mathcal{J}_{ij}\sigma_{j;\alpha}^{(\ell)}\right) + \frac{C_W^{(\ell+1)}}{2n_\ell}\sum_{i,j}\mathcal{J}_{ij}^2\right].\end{aligned} \tag{8.48}$$

Just as in our previous Hubbard–Stratonoviching (4.20), the *stochastic metric* (4.70),

$$\widehat{G}_{\alpha_1\alpha_2}^{(\ell+1)} \equiv C_b^{(\ell+1)} + C_W^{(\ell+1)}\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\sigma_{j;\alpha_1}^{(\ell)}\sigma_{j;\alpha_2}^{(\ell)}, \tag{8.49}$$

appears in the quadratic term of the Hubbard–Stratonovich variables $\Lambda_i{}^{\alpha}$, while the linear term is slightly modified by a shifting of the preactivations with the subtraction of the quantity

$$\widehat{\mathcal{M}}_{i;\alpha} \equiv C_W^{(\ell+1)}\left(\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\mathcal{J}_{ij}\sigma_{j;\alpha}^{(\ell)}\right). \tag{8.50}$$

Completing the squares with the Hubbard–Stratonovich variables and integrating them out, we get

$$\frac{\exp\left(\frac{C_W^{(\ell+1)}}{2n_\ell}\sum_{i,j}\mathcal{J}_{ij}^2\right)}{\sqrt{\left|2\pi\widehat{G}^{(\ell+1)}\right|^{n_{\ell+1}}}}\int\left[\prod_{i,\alpha}dz_{i;\alpha}^{(\ell+1)}\right]\mathcal{O}\left(z^{(\ell+1)}\right) \tag{8.51}$$
$$\times\exp\left[-\frac{1}{2}\sum_i\sum_{\alpha_1,\alpha_2}\widehat{G}^{\alpha_1\alpha_2}_{(\ell+1)}\left(z_{i;\alpha_1}^{(\ell+1)}-\widehat{\mathcal{M}}_{i;\alpha_1}\right)\left(z_{i;\alpha_2}^{(\ell+1)}-\widehat{\mathcal{M}}_{i;\alpha_2}\right)\right].$$

Ignoring the quadratic source factor $\mathcal{J}_{ij}^2$ outside the integral, this is just a *Gaussian expectation* of $\mathcal{O}$ against the $(\ell+1)$-th-layer preactivation distribution with a mean $\widehat{\mathcal{M}}_{i;\alpha}$ and a variance $\widehat{G}^{(\ell+1)}_{\alpha_1\alpha_2}$. (Make sure you remember our general relativity convention: $\widehat{G}^{\alpha_1\alpha_2}_{(\ell+1)}$ is the inverse of $\widehat{G}^{(\ell+1)}_{\alpha_1\alpha_2}$.)

Now, let's compensate for this mean by shifting the dummy integration variable as $z_{i;\alpha}^{(\ell+1)}\to z_{i;\alpha}^{(\ell+1)}+\widehat{\mathcal{M}}_{i;\alpha_1}$, which yields a compact expression in terms of our *zero-mean* Gaussian expectation notation (4.45):

$$\exp\left(\frac{C_W^{(\ell+1)}}{2n_\ell}\sum_{i,j}\mathcal{J}_{ij}^2\right)\left\langle\!\left\langle\mathcal{O}\left(z^{(\ell+1)}+\widehat{\mathcal{M}}\right)\right\rangle\!\right\rangle_{\widehat{G}^{(\ell+1)}}. \tag{8.52}$$

Plugging this result (8.52) back into our interlayer correlation (8.42) and substituting back in for the mean shift (8.50), we arrive at a simple formula for our generating function:

$$\mathbb{E}\left[\mathcal{O}\left(z^{(\ell+1)}\right)e^{\sum_{i,j}\mathcal{J}_{ij}W_{ij}^{(\ell+1)}}\mathcal{Q}\left(z^{(\ell)},\widehat{H}^{(\ell)}\right)\right] \tag{8.53}$$
$$=\exp\left(\frac{C_W^{(\ell+1)}}{2n_\ell}\sum_{i,j}\mathcal{J}_{ij}^2\right)\mathbb{E}\left[\left\langle\!\!\left\langle\mathcal{O}\left(z_{i;\alpha}^{(\ell+1)}+C_W^{(\ell+1)}\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\mathcal{J}_{ij}\sigma_{j;\alpha}^{(\ell)}\right)\right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}}\mathcal{Q}\left(z^{(\ell)},\widehat{H}^{(\ell)}\right)\right].$$

After performing the Gaussian expectation over the $(\ell+1)$-th-layer preactivations $z_{i;\alpha}^{(\ell+1)}$ – which is typically trivial in all the concrete applications that we'll encounter – the expectation in (8.53) is only with respect to ℓ-th-layer variables.[8] This was our desired result.

To see how to use the generating function (8.53), let's work out some explicit examples. First, consider the case with no weight insertions. Setting the source to zero, $\mathcal{J}_{ij}=0$, we find

$$\mathbb{E}\left[\mathcal{O}\left(z^{(\ell+1)}\right)\mathcal{Q}\left(z^{(\ell)},\widehat{H}^{(\ell)}\right)\right]=\mathbb{E}\left[\left\langle\!\left\langle\mathcal{O}\left(z^{(\ell+1)}\right)\right\rangle\!\right\rangle_{\widehat{G}^{(\ell+1)}}\mathcal{Q}\left(z^{(\ell)},\widehat{H}^{(\ell)}\right)\right]. \tag{8.54}$$

This formula is not trivial and is not something we knew before: here we see that the correlation between preactivations in neighboring layers is given by first computing a

[8] Recall that the stochastic metric $\widehat{G}^{(\ell+1)}$, (8.49), depends only on the ℓ-th-layer preactivations.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

Gaussian expectation of the $(\ell+1)$-th-layer function against the stochastic metric and then taking the full expectation of the resulting ℓ-th-layer quantity.

Next, let's consider two weight insertions. Twice-differentiating the generating function (8.53) by the source as $\frac{d}{d\mathcal{J}_{i_3 j_3}} \frac{d}{d\mathcal{J}_{i_4 j_4}}$ and then setting the source to zero, we get

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{O}\left(z^{(\ell+1)}\right) W^{(\ell+1)}_{i_3 j_3} W^{(\ell+1)}_{i_4 j_4} \mathcal{Q}\left(z^{(\ell)}, \widehat{H}^{(\ell)}\right)\right] \qquad (8.55)\\
&= \delta_{i_3 i_4} \delta_{j_3 j_4} \frac{C_W^{(\ell+1)}}{n_\ell} \mathbb{E}\left[\langle\!\langle \mathcal{O} \rangle\!\rangle_{\widehat{G}^{(\ell+1)}} \mathcal{Q}\left(z^{(\ell)}, \widehat{H}^{(\ell)}\right)\right]\\
&\quad + \left(\frac{C_W^{(\ell+1)}}{n_\ell}\right)^2 \sum_{\beta_3,\beta_4,\gamma_3,\gamma_4} \mathbb{E}\Bigg[\left\langle\!\!\left\langle \left(z^{(\ell+1)}_{i_3;\beta_3} z^{(\ell+1)}_{i_4;\beta_4} - \delta_{i_3 i_4} \widehat{G}^{(\ell+1)}_{\beta_3\beta_4}\right) \mathcal{O} \right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}}\\
&\qquad\qquad \times \widehat{G}^{\beta_3\gamma_3}_{(\ell+1)} \widehat{G}^{\beta_4\gamma_4}_{(\ell+1)} \sigma^{(\ell)}_{j_3;\gamma_3} \sigma^{(\ell)}_{j_4;\gamma_4} \mathcal{Q}\left(z^{(\ell)}, \widehat{H}^{(\ell)}\right)\Bigg].
\end{aligned}
$$

Here, we used integration by parts to exchange the derivatives for a projection as

$$
\left\langle\!\!\left\langle \frac{\partial^2 \mathcal{O}}{\partial z^{(\ell+1)}_{i_3;\gamma_3} \partial z^{(\ell+1)}_{i_4;\gamma_4}} \right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}} = \sum_{\beta_3,\beta_4} \widehat{G}^{\beta_3\gamma_3}_{(\ell+1)} \widehat{G}^{\beta_4\gamma_4}_{(\ell+1)} \left\langle\!\!\left\langle \left(z^{(\ell+1)}_{i_3;\beta_3} z^{(\ell+1)}_{i_4;\beta_4} - \delta_{i_3 i_4} \widehat{G}^{(\ell+1)}_{\beta_3\beta_4}\right) \mathcal{O} \right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}} . \qquad (8.56)
$$

Intuitively, the first term in (8.55) comes from forming a Wick contraction with the two weight insertions, while the second term comes from two pairs of Wick contractions, each between an inserted weight and a weight hidden inside the $z^{(\ell+1)}$'s in $\mathcal{O}$.

Thusly, with the *intra*layer formula (8.41) recalled and the *inter*layer formulae (8.54) and (8.55) derived, we are as ready as we'll ever be to recursively analyze the joint statistics of the NTK and preactivations in deeper layers. This concludes our *inter*lude.

8.3.1 NTK Mean

Taking the expectation of the stochastic NTK forward equation (8.12), we get

$$
\begin{aligned}
\mathbb{E}\left[\widehat{H}^{(\ell+1)}_{i_1 i_2;\alpha_1\alpha_2}\right] &= \delta_{i_1 i_2}\left[\lambda_b^{(\ell+1)} + \lambda_W^{(\ell+1)}\left(\frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right]\right)\right] \qquad (8.57)\\
&\quad + \delta_{i_1 i_2} C_W^{(\ell+1)} \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2} \widehat{H}^{(\ell)}_{jj;\alpha_1\alpha_2}\right],
\end{aligned}
$$

where, as is now familiar, on the second line we used the independence of the $(\ell+1)$-th-layer weights from the ℓ-th-layer preactivations, and then immediately evaluated the weight expectation. Immediately, we see that the NTK mean is diagonal in neural indices at any network depth.

Given that, let's decompose the ℓ-th-layer NTK into a mean and fluctuation as

$$
\widehat{H}^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2} \equiv \delta_{i_1 i_2} H^{(\ell)}_{\alpha_1\alpha_2} + \widehat{\Delta H}^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2}, \qquad (8.58)
$$

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

where we have denoted the ℓ-th-layer NTK mean as $\delta_{i_1 i_2} H^{(\ell)}_{\alpha_1\alpha_2}$. As before, we have separated the part of the mean that encodes the sample dependence and symbolized it without a hat. Substituting this decomposition into our expression (8.57) for the NTK mean, we see that the $(\ell+1)$-th-layer mean obeys a recursion

$$\begin{aligned} H^{(\ell+1)}_{\alpha_1\alpha_2} = \lambda_b^{(\ell+1)} + \lambda_W^{(\ell+1)} \left(\frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1} \sigma^{(\ell)}_{j;\alpha_2} \right] \right) + C_W^{(\ell+1)} H^{(\ell)}_{\alpha_1\alpha_2} \left(\frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1} \sigma'^{(\ell)}_{j;\alpha_2} \right] \right) \\ + C_W^{(\ell+1)} \left(\frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1} \sigma'^{(\ell)}_{j;\alpha_2} \widehat{\Delta H}^{(\ell)}_{jj;\alpha_1\alpha_2} \right] \right), \end{aligned} \tag{8.59}$$

depending on both the mean and fluctuation in the previous layer ℓ.

To the leading order in $1/n$, the first two expectation values on the right-hand side of (8.59) are given by Gaussian expectations

$$\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1} \sigma^{(\ell)}_{j;\alpha_2} \right] = \langle \sigma_{\alpha_1} \sigma_{\alpha_2} \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right), \tag{8.60}$$

$$\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1} \sigma'^{(\ell)}_{j;\alpha_2} \right] = \langle \sigma'_{\alpha_1} \sigma'_{\alpha_2} \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right). \tag{8.61}$$

To see this, note that the first expectation is just the leading Gaussian contribution (4.61) with the non-Gaussian coupling v suppressed as $\sim 1/n$ as per (4.82), and that the evaluation of the second expectation proceeds identically to the first regardless of whether the activation has a derivative or not. Meanwhile, the final expectation on the second line of (8.59) involves an NTK–preactivation cross correlation, which is also suppressed in the large-width limit:

$$\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1} \sigma'^{(\ell)}_{j;\alpha_2} \widehat{\Delta H}^{(\ell)}_{jj;\alpha_1\alpha_2} \right] = O\left(\frac{1}{n}\right). \tag{8.62}$$

We will prove this shortly in the next subsection in (8.71).

Assembling these leading contributions, the NTK mean recursion simplifies to

$$H^{(\ell+1)}_{\alpha_1\alpha_2} = \lambda_b^{(\ell+1)} + \lambda_W^{(\ell+1)} \langle \sigma_{\alpha_1} \sigma_{\alpha_2} \rangle_{G^{(\ell)}} + C_W^{(\ell+1)} \langle \sigma'_{\alpha_1} \sigma'_{\alpha_2} \rangle_{G^{(\ell)}} H^{(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n}\right). \tag{8.63}$$

If you're in the habit of marking up your book, feel free to draw a box around this formula.

8.3.2 NTK–Preactivation Cross Correlations

Next, let's evaluate a cross-correlation expectation of a very general form,

$$\mathbb{E}\left[\mathcal{O}\left(z^{(\ell+1)}\right) \widehat{\Delta H}^{(\ell+1)}_{i_3 i_4;\alpha_3\alpha_4} \right]. \tag{8.64}$$

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

For instance, setting $\mathcal{O} = z^{(\ell+1)}_{i_1;\alpha_1} z^{(\ell+1)}_{i_2;\alpha_2}$ gives the elementary cross correlation (8.43), while setting $\mathcal{O} = \sigma'^{(\ell)}_{i_1;\alpha_1}\sigma'^{(\ell)}_{i_2;\alpha_2}$ gives the subleading cross correlation (8.62) that just appeared in (and then immediately disappeared from) our recursion for the NTK mean.

To begin, simply substitute the NTK forward equation (8.12) into the cross correlator (8.64), which yields

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{O}\Big(z^{(\ell+1)}\Big)\,\widehat{\Delta H}^{(\ell+1)}_{i_3i_4;\alpha_3\alpha_4}\right] = \mathrm{Cov}\Big[\mathcal{O}\Big(z^{(\ell+1)}\Big)\,,\,\widehat{H}^{(\ell+1)}_{i_3i_4;\alpha_3\alpha_4}\Big] \qquad (8.65)\\
&= \delta_{i_3i_4}\,\lambda^{(\ell+1)}_W \frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\left\{\mathbb{E}\left[\mathcal{O}\Big(z^{(\ell+1)}\Big)\,\sigma^{(\ell)}_{j;\alpha_3}\sigma^{(\ell)}_{j;\alpha_4}\right] - \mathbb{E}\left[\mathcal{O}\Big(z^{(\ell+1)}\Big)\right]\mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_3}\sigma^{(\ell)}_{j;\alpha_4}\right]\right\}\\
&\quad + \sum_{j_3,j_4=1}^{n_\ell}\left\{\mathbb{E}\left[\mathcal{O}\Big(z^{(\ell+1)}\Big)\,W^{(\ell+1)}_{i_3j_3}W^{(\ell+1)}_{i_4j_4}\sigma'^{(\ell)}_{j_3;\alpha_3}\sigma'^{(\ell)}_{j_4;\alpha_4}\widehat{H}^{(\ell)}_{j_3j_4;\alpha_3\alpha_4}\right]\right.\\
&\qquad\qquad \left. - \mathbb{E}\left[\mathcal{O}\Big(z^{(\ell+1)}\Big)\right]\mathbb{E}\left[W^{(\ell+1)}_{i_3j_3}W^{(\ell+1)}_{i_4j_4}\right]\mathbb{E}\left[\sigma'^{(\ell)}_{j_3;\alpha_3}\sigma'^{(\ell)}_{j_4;\alpha_4}\widehat{H}^{(\ell)}_{j_3j_4;\alpha_3\alpha_4}\right]\right\}.
\end{aligned}
$$

Now putting the freshly-derived interlayer formulae (8.54) and (8.55) to use, this cross correlator becomes

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{O}\Big(z^{(\ell+1)}\Big)\,\widehat{\Delta H}^{(\ell+1)}_{i_3i_4;\alpha_3\alpha_4}\right] \qquad (8.66)\\
&= \delta_{i_3i_4}\frac{\lambda^{(\ell+1)}_W}{C^{(\ell+1)}_W}\,\mathbb{E}\left[\left\langle\!\!\left\langle \mathcal{O}\Big(z^{(\ell+1)}\Big)\right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}}\widehat{\Delta G}^{(\ell+1)}_{\alpha_3\alpha_4}\right]\\
&\quad + \delta_{i_3i_4}\frac{C^{(\ell+1)}_W}{n_\ell}\sum_{j=1}^{n_\ell}\left\{\mathbb{E}\left[\left\langle\!\!\left\langle \mathcal{O}\Big(z^{(\ell+1)}\Big)\right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{j;\alpha_4}\widehat{H}^{(\ell)}_{jj;\alpha_3\alpha_4}\right]\right.\\
&\qquad\qquad \left. - \mathbb{E}\left[\left\langle\!\!\left\langle \mathcal{O}\Big(z^{(\ell+1)}\Big)\right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}}\right]\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{j;\alpha_4}\widehat{H}^{(\ell)}_{jj;\alpha_3\alpha_4}\right]\right\}\\
&\quad + \left(\frac{C^{(\ell+1)}_W}{n_\ell}\right)^2\sum_{j_3,j_4=1}^{n_\ell}\sum_{\beta_3,\beta_4,\gamma_3,\gamma_4}\mathbb{E}\Bigg[\left\langle\!\!\left\langle\Big(z^{(\ell+1)}_{i_3;\beta_3}z^{(\ell+1)}_{i_4;\beta_4} - \delta_{i_3i_4}\widehat{G}^{(\ell+1)}_{\beta_3\beta_4}\Big)\,\mathcal{O}\Big(z^{(\ell+1)}\Big)\right\rangle\!\!\right\rangle_{\widehat{G}^{(\ell+1)}}\\
&\qquad\qquad \times\,\widehat{G}^{\beta_3\gamma_3}_{(\ell+1)}\widehat{G}^{\beta_4\gamma_4}_{(\ell+1)}\sigma^{(\ell)}_{j_3;\gamma_3}\sigma^{(\ell)}_{j_4;\gamma_4}\sigma'^{(\ell)}_{j_3;\alpha_3}\sigma'^{(\ell)}_{j_4;\alpha_4}\widehat{H}^{(\ell)}_{j_3j_4;\alpha_3\alpha_4}\Bigg].
\end{aligned}
$$

Here, for the first term, we also recalled the definition of the metric fluctuation (4.74). From this general expression, we can already learn two important lessons.

First, setting $\mathcal{O} = z^{(\ell+1)}_{i_1;\alpha_1} z^{(\ell+1)}_{i_2;\alpha_2}$, we get an expression for the elementary $(\ell+1)$-th-layer cross correlation in terms of ℓ-th-layer variables:

$$
\begin{aligned}
&\mathbb{E}\left[z^{(\ell+1)}_{i_1;\alpha_1}z^{(\ell+1)}_{i_2;\alpha_2}\widehat{\Delta H}^{(\ell+1)}_{i_3i_4;\alpha_3\alpha_4}\right]\\
&= \delta_{i_1i_2}\delta_{i_3i_4}\left\{\frac{\lambda^{(\ell+1)}_W}{C^{(\ell+1)}_W}\mathbb{E}\left[\widehat{\Delta G}^{(\ell+1)}_{\alpha_1\alpha_2}\widehat{\Delta G}^{(\ell+1)}_{\alpha_3\alpha_4}\right] + C^{(\ell+1)}_W\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\mathbb{E}\left[\widehat{\Delta G}^{(\ell+1)}_{\alpha_1\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{j;\alpha_4}\widehat{H}^{(\ell)}_{jj;\alpha_3\alpha_4}\right]\right\}
\end{aligned}
$$

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

$$
\begin{aligned}
&+ \delta_{i_1 i_3}\delta_{i_2 i_4}\left(\frac{C_W^{(\ell+1)}}{n_\ell}\right)^2 \sum_{j,k=1}^{n_\ell} \mathbb{E}\left[\sigma_{j;\alpha_1}^{(\ell)}\sigma_{k;\alpha_2}^{(\ell)}\sigma_{j;\alpha_3}^{\prime(\ell)}\sigma_{k;\alpha_4}^{\prime(\ell)}\widehat{H}_{jk;\alpha_3\alpha_4}^{(\ell)}\right] \\
&+ \delta_{i_1 i_4}\delta_{i_2 i_3}\left(\frac{C_W^{(\ell+1)}}{n_\ell}\right)^2 \sum_{j,k=1}^{n_\ell} \mathbb{E}\left[\sigma_{j;\alpha_2}^{(\ell)}\sigma_{k;\alpha_1}^{(\ell)}\sigma_{j;\alpha_3}^{\prime(\ell)}\sigma_{k;\alpha_4}^{\prime(\ell)}\widehat{H}_{jk;\alpha_3\alpha_4}^{(\ell)}\right] \\
\equiv\ & \frac{1}{n_\ell}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}D_{\alpha_1\alpha_2\alpha_3\alpha_4}^{(\ell+1)} + \delta_{i_1 i_3}\delta_{i_2 i_4}F_{\alpha_1\alpha_3\alpha_2\alpha_4}^{(\ell+1)} + \delta_{i_1 i_4}\delta_{i_2 i_3}F_{\alpha_1\alpha_4\alpha_2\alpha_3}^{(\ell+1)}\right], \qquad (8.67)
\end{aligned}
$$

where on the final line we decomposed the cross correlation into two tensors with sample indices only, just as we did for the second layer in (8.37). Equating the first expression with the second, we see that these tensors are defined by the following ℓ-th-layer expectations:

$$
\begin{aligned}
\frac{1}{n_\ell}D_{\alpha_1\alpha_2\alpha_3\alpha_4}^{(\ell+1)} \equiv\ & \frac{\lambda_W^{(\ell+1)}}{C_W^{(\ell+1)}}\mathbb{E}\left[\widehat{\Delta G}_{\alpha_1\alpha_2}^{(\ell+1)}\widehat{\Delta G}_{\alpha_3\alpha_4}^{(\ell+1)}\right] \qquad (8.68) \\
&+ C_W^{(\ell+1)}\left(\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\mathbb{E}\left[\widehat{\Delta G}_{\alpha_1\alpha_2}^{(\ell+1)}\sigma_{j;\alpha_3}^{\prime(\ell)}\sigma_{j;\alpha_4}^{\prime(\ell)}\widehat{H}_{jj;\alpha_3\alpha_4}^{(\ell)}\right]\right), \\
\frac{1}{n_\ell}F_{\alpha_1\alpha_3\alpha_2\alpha_4}^{(\ell+1)} \equiv\ & \left(C_W^{(\ell+1)}\right)^2\left(\frac{1}{n_\ell^2}\sum_{j,k=1}^{n_\ell}\mathbb{E}\left[\sigma_{j;\alpha_1}^{(\ell)}\sigma_{k;\alpha_2}^{(\ell)}\sigma_{j;\alpha_3}^{\prime(\ell)}\sigma_{k;\alpha_4}^{\prime(\ell)}\widehat{H}_{jk;\alpha_3\alpha_4}^{(\ell)}\right]\right). \qquad (8.69)
\end{aligned}
$$

We'll come back to evaluate these last two expressions – and thereby derive recursions for both cross correlation tensors – after we reveal the second lesson.

Second, we can start again from the general cross correlator (8.66) and push our calculation a little bit further to leading order in $1/n$. The key step is perturbatively expanding the Gaussian expectation – just as we expanded the stochastic Gaussian distribution (4.56) before using the Schwinger–Dyson equation (4.55) – to get

$$
\begin{aligned}
&\left\langle\!\left\langle \mathcal{O}\left(z^{(\ell+1)}\right)\right\rangle\!\right\rangle_{\widehat{G}^{(\ell+1)}} \qquad (8.70) \\
&= \left\langle\!\left\langle \mathcal{O}\left(z^{(\ell+1)}\right)\right\rangle\!\right\rangle_{G^{(\ell+1)}} \\
&+ \frac{1}{2}\sum_{\beta_1,\beta_2,\gamma_1,\gamma_2}\left\langle\!\!\left\langle \sum_m \left(z_{m;\beta_1}^{(\ell+1)}z_{m;\beta_2}^{(\ell+1)} - G_{\beta_1\beta_2}^{(\ell+1)}\right)\mathcal{O}\left(z^{(\ell+1)}\right)\right\rangle\!\!\right\rangle_{G^{(\ell+1)}} G_{(\ell+1)}^{\beta_1\gamma_1}G_{(\ell+1)}^{\beta_2\gamma_2}\widehat{\Delta G}_{\gamma_1\gamma_2}^{(\ell+1)} + O\left(\Delta^2\right).
\end{aligned}
$$

Plugging this back into (8.66), picking up the (leading-order) pieces, and using the definitions (8.68) and (8.69), we get

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{O}\left(z^{(\ell)}\right)\widehat{\Delta H}^{(\ell)}_{i_3i_4;\alpha_3\alpha_4}\right] \qquad (8.71)\\
&= \delta_{i_3i_4}\frac{1}{n_{\ell-1}}\left[\frac{1}{2}\sum_{\beta_1,\beta_2,\gamma_1,\gamma_2}\left\langle\!\!\left\langle\sum_{m=1}^{n_\ell}\left(z^{(\ell)}_{m;\beta_1}z^{(\ell)}_{m;\beta_2}-G^{(\ell)}_{\beta_1\beta_2}\right)\mathcal{O}\left(z^{(\ell)}\right)\right\rangle\!\!\right\rangle_{G^{(\ell)}} G^{\beta_1\gamma_1}_{(\ell)}G^{\beta_2\gamma_2}_{(\ell)}\right]D^{(\ell)}_{\gamma_1\gamma_2\alpha_3\alpha_4}\\
&\quad+\frac{1}{n_{\ell-1}}\sum_{\beta_1,\beta_2,\gamma_1,\gamma_2}\left\langle\!\!\left\langle\left(z^{(\ell)}_{i_3;\beta_1}z^{(\ell)}_{i_4;\beta_2}-\delta_{i_3i_4}G^{(\ell)}_{\beta_1\beta_2}\right)\mathcal{O}\left(z^{(\ell)}\right)\right\rangle\!\!\right\rangle_{G^{(\ell)}} G^{\beta_1\gamma_1}_{(\ell)}G^{\beta_2\gamma_2}_{(\ell)}F^{(\ell)}_{\gamma_1\alpha_3\gamma_2\alpha_4}\\
&\quad+O\left(\frac{1}{n^2}\right),
\end{aligned}
$$

where we have also relabeled *layer indices* as $(\ell+1)\to\ell$ everywhere for ease of later substitutions.[9] This result illustrates that these more general cross correlations are governed by the same tensors, $D^{(\ell)}$ and $F^{(\ell)}$, as the elementary cross correlation (8.67). We can indeed compute all the cross correlators if we find and solve recursions for $D^{(\ell)}$ and $F^{(\ell)}$. It is this task that we turn to next.

D-recursion

Starting from our expression for $D^{(\ell+1)}$ (8.68) and substituting in the definition of the stochastic metric (8.49), we get

$$
\begin{aligned}
D^{(\ell+1)}_{\alpha_1\alpha_2\alpha_3\alpha_4} &= C_W^{(\ell+1)}\frac{1}{n_\ell}\sum_{j,k=1}^{n_\ell}\mathrm{Cov}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2},\,\lambda_W^{(\ell+1)}\sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}+C_W^{(\ell+1)}H^{(\ell)}_{\alpha_3\alpha_4}\sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right]\\
&\quad+\left(C_W^{(\ell+1)}\right)^2\frac{1}{n_\ell}\sum_{j,k=1}^{n_\ell}\mathrm{Cov}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2},\,\sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right]+O\left(\frac{1}{n}\right),
\end{aligned}
\tag{8.72}
$$

where we have again decomposed the ℓ-th-layer NTK into a mean and fluctuation piece. We see that there are two types of terms here: covariances on a single neuron $j=k$ and covariances between pairs of neurons $j\neq k$.

For the single-neuron contribution with $j=k$, at the leading order, we find the same contribution that we found for the second layer (8.38):[10]

$$
C_W^{(\ell+1)}\left[\left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}}-\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\rangle_{G^{(\ell)}}\left\langle\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}}\right]+O\left(\frac{1}{n}\right). \tag{8.73}
$$

Here, the auxiliary stochastic matrix $\widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}$ is defined as

$$
\widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\equiv\lambda_W^{(\ell+1)}\sigma^{(\ell)}_{\alpha_1}\sigma^{(\ell)}_{\alpha_2}+C_W^{(\ell+1)}H^{(\ell)}_{\alpha_1\alpha_2}\sigma'^{(\ell)}_{\alpha_1}\sigma'^{(\ell)}_{\alpha_2}, \tag{8.74}
$$

[9] You might worry about the summation $\sum_{m=1}^{n_\ell}$ inside the first Gaussian expectation in (8.71). However, due to Gaussian factorization, this expectation stays of order one so long as the observable $\mathcal{O}$ depends on only a finite number of neurons.

[10] The subleading $O(1/n)$ piece includes a contribution from the non-Gaussian part of the distribution as well as a cross correlation contribution from the previous layer.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

which simply generalizes the second-layer definition (8.34). As a reminder, the unhatted matrix $H^{(\ell)}_{\alpha_1\alpha_2}$ is the NTK mean, which is not a random variable and can safely be taken outside the Gaussian expectation $\langle\cdot\rangle_{G^{(\ell)}}$. This means that, as a stochastic variable, $\widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}$ depends only on the ℓ-th-layer preactivations.

Next, for the pairs-of-neurons contribution to (8.72) with $j \neq k$, the first term can be evaluated by the intralayer formula (8.41) and yields

$$\frac{n_\ell}{4n_{\ell-1}} C_W^{(\ell+1)} \sum_{\gamma_1,\gamma_2,\gamma_3,\gamma_4} V_{(\ell)}^{(\gamma_1\gamma_2)(\gamma_3\gamma_4)} \left\langle \left(z_{\gamma_1} z_{\gamma_2} - G^{(\ell)}_{\gamma_1\gamma_2}\right) \sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}} \left\langle \left(z_{\gamma_3} z_{\gamma_4} - G^{(\ell)}_{\gamma_3\gamma_4}\right) \widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}} . \tag{8.75}$$

Meanwhile, the covariance in the second term can be unrolled as

$$\begin{aligned}
&\mathrm{Cov}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right] \qquad (8.76)\\
&= \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right] - \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\right]\mathbb{E}\left[\sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right]\\
&= \frac{1}{2n_{\ell-1}} \sum_{\beta_1,\beta_2,\gamma_1,\gamma_2} \left\langle \left(z_{\beta_1}z_{\beta_2} - G^{(\ell)}_{\beta_1\beta_2}\right)\sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}} \left\langle \sigma'_{\alpha_3}\sigma'_{\alpha_4}\right\rangle_{G^{(\ell)}} G^{\beta_1\gamma_1}_{(\ell)} G^{\beta_2\gamma_2}_{(\ell)} D^{(\ell)}_{\gamma_1\gamma_2\alpha_3\alpha_4} + O\!\left(\frac{1}{n^2}\right),
\end{aligned}$$

where in the last line we used the cross-correlation formula (8.71) with the observables $\mathcal{O} = \sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2}\sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}$ and $\mathcal{O} = \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}$, respectively. Combining these contributions with the first piece (8.73), we get our desired recursion:

$$\begin{aligned}
&D^{(\ell+1)}_{\alpha_1\alpha_2\alpha_3\alpha_4} \qquad (8.77)\\
&= C_W^{(\ell+1)}\left(\left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}} - \left\langle\sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}}\left\langle\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}}\right)\\
&\quad + \frac{n_\ell}{4n_{\ell-1}} C_W^{(\ell+1)} \sum_{\gamma_1,\gamma_2,\gamma_3,\gamma_4} V_{(\ell)}^{(\gamma_1\gamma_2)(\gamma_3\gamma_4)} \left\langle\left(z_{\gamma_1}z_{\gamma_2} - G^{(\ell)}_{\gamma_1\gamma_2}\right)\sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}} \left\langle\left(z_{\gamma_3}z_{\gamma_4} - G^{(\ell)}_{\gamma_3\gamma_4}\right)\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}}\\
&\quad + \frac{n_\ell}{2n_{\ell-1}}\left(C_W^{(\ell+1)}\right)^2 \sum_{\beta_1,\beta_2,\gamma_1,\gamma_2} D^{(\ell)}_{\gamma_1\gamma_2\alpha_3\alpha_4} \left\langle\left(z_{\beta_1}z_{\beta_2} - G^{(\ell)}_{\beta_1\beta_2}\right)\sigma_{\alpha_1}\sigma_{\alpha_2}\right\rangle_{G^{(\ell)}} G^{\beta_1\gamma_1}_{(\ell)} G^{\beta_2\gamma_2}_{(\ell)} \left\langle\sigma'_{\alpha_3}\sigma'_{\alpha_4}\right\rangle_{G^{(\ell)}}\\
&\quad + O\!\left(\frac{1}{n}\right).
\end{aligned}$$

Interestingly, we see that at leading order, the D-type cross correlation in layer $(\ell+1)$ mixes D-type correlations from layer ℓ with the four-point vertex $V^{(\ell)}$ but does not mix with the F-type cross correlations or any part of the NTK variance.

F-recursion

Starting from our expression for $F^{(\ell+1)}$ (8.69) and decomposing the NTK into a mean and fluctuation, we get

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

$$
\begin{aligned}
F^{(\ell+1)}_{\alpha_1\alpha_3\alpha_2\alpha_4} = {} & \left(C_W^{(\ell+1)}\right)^2 \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_3}\sigma^{(\ell)}_{j;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_4}\right] H^{(\ell)}_{\alpha_3\alpha_4} \\
& + \left(C_W^{(\ell+1)}\right)^2 \frac{1}{n_\ell}\sum_{j,k=1}^{n_\ell} \mathbb{E}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_3}\sigma^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right].
\end{aligned}
\tag{8.78}
$$

At leading order, the first term simply becomes a single-neuron Gaussian expectation; the second term can be evaluated with the cross-correlation formula (8.71), where the diagonal sum with $j = k$ is of order $O(1/n)$ and can be neglected, while the off-diagonal sum with $j \neq k$ yields the term involving $F^{(\ell)}$. All together, this gives

$$
\begin{aligned}
F^{(\ell+1)}_{\alpha_1\alpha_3\alpha_2\alpha_4} = {} & \left(C_W^{(\ell+1)}\right)^2 \langle \sigma_{\alpha_1}\sigma_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(\ell)}} H^{(\ell)}_{\alpha_3\alpha_4} \\
& + \frac{n_\ell}{n_{\ell-1}}\left(C_W^{(\ell+1)}\right)^2 \sum_{\beta_1,\beta_2,\gamma_1,\gamma_2} \langle \sigma_{\alpha_1}\sigma'_{\alpha_3} z_{\beta_1}\rangle_{G^{(\ell)}} \langle \sigma_{\alpha_2}\sigma'_{\alpha_4} z_{\beta_2}\rangle_{G^{(\ell)}} G^{\beta_1\gamma_1}_{(\ell)} G^{\beta_2\gamma_2}_{(\ell)} F^{(\ell)}_{\gamma_1\alpha_3\gamma_2\alpha_4} \\
& + O\!\left(\frac{1}{n}\right).
\end{aligned}
\tag{8.79}
$$

As with the D-recursion before, the first term was present in the second-layer $F^{(2)}$ (8.39), while the second term is a direct consequence of having a fluctuating NTK in the previous layer ℓ. Additionally, we see that at leading order, the F-type cross correlation doesn't mix at all with any of our other finite-width tensors.

Finally, before moving on to discuss the NTK variance, let us note that the recursions for both $D^{(\ell)}$ and $F^{(\ell)}$ – combined with the initial condition $D^{(1)} = F^{(1)} = 0$ from the first layer where the NTK is deterministic – ensure that they each stay of order one. Given the factor of $1/n_\ell$ in the decomposition of the cross correlation into these tensors (8.67) and our "second lesson" encapsulated by the cross-correlation formula (8.71), this means that any and all cross correlations are suppressed in the $1/n$ expansion and vanish identically in the strict infinite-width limit.

8.3.3 NTK Variance

Now let's finally *slay the beast* that is the NTK variance. Similar to the NTK–preactivation cross correlation, the NTK-variance calculation in deeper layers differs from the second-layer calculation due to nontrivial intralayer correlations (8.41) in the previous layer and due to the fluctuating NTK (8.58).

The NTK variance is given by the expected magnitude of the NTK fluctuation

$$
\mathbb{E}\left[\widehat{\Delta H}^{(\ell+1)}_{i_1i_2;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell+1)}_{i_3i_4;\alpha_3\alpha_4}\right] = \operatorname{Cov}\left[\widehat{H}^{(\ell+1)}_{i_1i_2;\alpha_1\alpha_2},\, \widehat{H}^{(\ell+1)}_{i_3i_4;\alpha_3\alpha_4}\right].
\tag{8.80}
$$

To begin our calculation, let us plug the NTK forward equation (8.12) into this defining expression and then integrate out the weights $W^{(\ell+1)}$, which is easy since they are independent random variables. Although there are many terms, the algebra is mostly straightforward:

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\Delta H}^{(\ell+1)}_{i_1 i_2;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell+1)}_{i_3 i_4;\alpha_3\alpha_4}\right] \qquad (8.81)\\
&= \delta_{i_1 i_2}\delta_{i_3 i_4}\frac{1}{n_\ell^2}\sum_{j,k=1}^{n_\ell}\Bigg\{\left(\lambda_W^{(\ell+1)}\right)^2 \mathrm{Cov}\Big[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\Big]\\
&\qquad + \left(\lambda_W^{(\ell+1)}\right) C_W^{(\ell+1)} \mathrm{Cov}\Big[\sigma^{(\ell)}_{j;\alpha}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{H}^{(\ell)}_{kk;\alpha_3\alpha_4}\Big]\\
&\qquad + \left(\lambda_W^{(\ell+1)}\right) C_W^{(\ell+1)} \mathrm{Cov}\Big[\sigma'^{(\ell)}_{j;\alpha}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{H}^{(\ell)}_{jj;\alpha\alpha_2},\, \sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\Big]\\
&\qquad + \left(C_W^{(\ell+1)}\right)^2 \mathrm{Cov}\Big[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{H}^{(\ell)}_{jj;\alpha_1\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{H}^{(\ell)}_{kk;\alpha_3\alpha_4}\Big]\Bigg\}\\
&+ \delta_{i_1 i_3}\delta_{i_2 i_4}\left(C_W^{(\ell+1)}\right)^2\frac{1}{n_\ell^2}\sum_{j,k=1}^{n_\ell}\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\widehat{H}^{(\ell)}_{jk;\alpha_1\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right]\\
&+ \delta_{i_1 i_4}\delta_{i_2 i_3}\left(C_W^{(\ell+1)}\right)^2\frac{1}{n_\ell^2}\sum_{j,k=1}^{n_\ell}\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\widehat{H}^{(\ell)}_{jk;\alpha_1\alpha_2}\sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{j;\alpha_4}\widehat{H}^{(\ell)}_{kj;\alpha_3\alpha_4}\right].
\end{aligned}
$$

In obtaining these last three terms, you should have made three distinct pairings for the two pairs of Wick contractions of the four $W^{(\ell+1)}$'s, one paring within the same NTK and two pairings across the NTKs.

An inspection of the pattern of neural indices in the Kronecker deltas from (8.81) suggests that we should again decompose the NTK variance into two tensors as

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{i_3 i_4;\alpha_3\alpha_4}\right] \qquad (8.82)\\
&\equiv \frac{1}{n_{\ell-1}}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}A^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4}B^{(\ell)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}B^{(\ell)}_{\alpha_1\alpha_4\alpha_2\alpha_3}\right],
\end{aligned}
$$

just as we did for the second layer before in (8.31). Here, a factor of $1/n_{\ell-1}$ was pulled out in anticipation that the overall variance would be $O(1/n)$ just as it was for the second layer. For now you can think of this parameterization as an ansatz; we will soon recursively show that $A^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$ and $B^{(\ell)}_{\alpha_1\alpha_3\alpha_2\alpha_4}$ stay of order one as the network width increases.

Now, let's work out the layer recursions for $A^{(\ell)}$ and $B^{(\ell)}$.

B-recursion

We'll start with B-recursion because it's simpler. Considering (8.81) with the decomposition (8.82) in mind, we see that $B^{(\ell+1)}$ is given by the following ℓ-th-layer expectation:

$$
B^{(\ell+1)}_{\alpha_1\alpha_3\alpha_2\alpha_4} = \left(C_W^{(\ell+1)}\right)^2\frac{1}{n_\ell}\sum_{j,k=1}^{n_\ell}\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\widehat{H}^{(\ell)}_{jk;\alpha_1\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right]. \qquad (8.83)
$$

As should now be familiar, the double summation in (8.83) splits into two types of terms, diagonal ones with $j = k$ and off-diagonal ones with $j \neq k$.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

For the diagonal part, the leading contribution is from the NTK mean

$$\begin{aligned}
&\left(C_W^{(\ell+1)}\right)^2 \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{j;\alpha_4}\right] H^{(\ell)}_{\alpha_1\alpha_2}H^{(\ell)}_{\alpha_3\alpha_4} + O\left(\frac{1}{n}\right) \qquad (8.84)\\
&= \left(C_W^{(\ell+1)}\right)^2 \langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(\ell)}} H^{(\ell)}_{\alpha_1\alpha_2}H^{(\ell)}_{\alpha_3\alpha_4} + O\left(\frac{1}{n}\right),
\end{aligned}$$

which is analogous to what we found in the second layer (8.33).[11]

For the off-diagonal part of (8.83), the NTK mean vanishes, and the leading contribution is from the NTK fluctuation

$$\left(C_W^{(\ell+1)}\right)^2 \frac{1}{n_\ell}\sum_{\substack{j,k=1\\ j\neq k}}^{n_\ell} \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right]. \qquad (8.85)$$

The expectation is already $O(\Delta^2)$ from the two NTK fluctuations inside it, and thus, neglecting higher-order correlations of order $O(\Delta^3)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right] \qquad (8.86)\\
&= \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right]\mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right] + O\left(\frac{1}{n^2}\right),
\end{aligned}$$

where the detailed explanation for such a factorization is given in this footnote.[12] Then, using the decomposition (8.82) for the NTK variance and similar logic as (4.63) to evaluate the four-point correlator of off-diagonal activations, we get

[11] Again, the subleading $O(1/n)$ piece includes a contribution from the non-Gaussian distribution as well as a cross correlation contribution from the previous layer.

[12] In greater detail, you can think of what we are doing here as separating $\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}$ into a mean $\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right]$ and a fluctuation, and – since the expectation already contains two fluctuations $\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}$ – the latter fluctuating piece contributes $O(\Delta^3)$ and thus can be neglected.

In alternate detail, we can view this expectation (8.86) as a correlator of three random variables, $\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}$, $\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}$, and $\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}$, and decompose it into one-point and two-point correlators as

$$\begin{aligned}
&\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right] \qquad (8.87)\\
&= \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right]\mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\right]\mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right]\\
&\quad + \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right]\mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right]\\
&\quad + \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\right]\mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right]\\
&\quad + \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right]\mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\right] + O\left(\frac{1}{n^2}\right),
\end{aligned}$$

where the $O(1/n^2)$ part contains the connected piece of the decomposition. Since the NTK fluctuation has mean zero, only the second term survives at this order.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

$$\left(C_W^{(\ell+1)}\right)^2 \frac{1}{n_\ell} \sum_{\substack{j,k=1\\ j\neq k}}^{n_\ell} \mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{k;\alpha_2}\sigma'^{(\ell)}_{j;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right] \mathbb{E}\left[\widehat{\Delta H}^{(\ell)}_{jk;\alpha_1\alpha_2}\widehat{\Delta H}^{(\ell)}_{jk;\alpha_3\alpha_4}\right] + O\!\left(\frac{1}{n}\right) \tag{8.88}$$
$$= \left(C_W^{(\ell+1)}\right)^2 \left\langle \sigma'_{\alpha_1}\sigma'_{\alpha_3}\right\rangle_{G^{(\ell)}} \left\langle \sigma'_{\alpha_2}\sigma'_{\alpha_4}\right\rangle_{G^{(\ell)}} \frac{n_\ell}{n_{\ell-1}} B^{(\ell)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + O\!\left(\frac{1}{n}\right).$$

Substituting both the diagonal contribution (8.84) and off-diagonal contribution (8.88) back into (8.83), we get the B-recursion:

$$B^{(\ell+1)}_{\alpha_1\alpha_3\alpha_2\alpha_4} = \left(C_W^{(\ell+1)}\right)^2 \Bigg[\left\langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\sigma'_{\alpha_3}\sigma'_{\alpha_4}\right\rangle_{G^{(\ell)}} H^{(\ell)}_{\alpha_1\alpha_2} H^{(\ell)}_{\alpha_3\alpha_4} \tag{8.89}$$
$$+ \left(\frac{n_\ell}{n_{\ell-1}}\right) \left\langle \sigma'_{\alpha_1}\sigma'_{\alpha_3}\right\rangle_{G^{(\ell)}} \left\langle \sigma'_{\alpha_2}\sigma'_{\alpha_4}\right\rangle_{G^{(\ell)}} B^{(\ell)}_{\alpha_1\alpha_3\alpha_2\alpha_4}\Bigg] + O\!\left(\frac{1}{n}\right).$$

As promised, we recursively see that $B^{(\ell)}$ is an order-one quantity. Additionally, we note that at leading order, this B-type NTK variance doesn't mix with any other finite-width tensors.

A-recursion

Let us now determine the A-recursion. Again equating our expression for the NTK variance (8.81) with the A/B-decomposition (8.82), we see that $A^{(\ell+1)}$ is given by the following ℓ-th-layer covariances:

$$A^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} = \frac{1}{n_\ell}\sum_{j,k=1}^{n_\ell} \Bigg\{ \left(\lambda_W^{(\ell+1)}\right)^2 \mathrm{Cov}\!\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right] \tag{8.90}$$
$$+ \lambda_W^{(\ell+1)} C_W^{(\ell+1)} \mathrm{Cov}\!\left[\sigma^{(\ell)}_{j;\alpha}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right]$$
$$+ \lambda_W^{(\ell+1)} C_W^{(\ell+1)} \mathrm{Cov}\!\left[\sigma'^{(\ell)}_{j;\alpha}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{H}^{(\ell)}_{jj;\alpha\alpha_2},\, \sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right]$$
$$+ \left(C_W^{(\ell+1)}\right)^2 \mathrm{Cov}\!\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{H}^{(\ell)}_{jj;\alpha_1\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right] \Bigg\}.$$

As we've now seen many times previously, our approach will be to divide up the double summation in (8.90) into two types of terms, diagonal terms on a single neuron with $j = k$ and off-diagonal terms on pairs of neurons with $j \neq k$.

As was the case for the B-recursion, the leading contribution from the diagonal part with $j = k$ comes from the NTK mean and matches what we found for the second layer (8.32):

$$\left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}} - \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\right\rangle_{G^{(\ell)}} \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}} + O\!\left(\frac{1}{n}\right), \tag{8.91}$$

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press

where the definition of the auxiliary stochastic tensor $\widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}$ was given in (8.74).[13]

This leaves us with the off-diagonal part of (8.90) with $j \neq k$. Here, there will be leading contributions both from the NTK mean and from the NTK fluctuations. The contributions from the mean are given by replacing $\widehat{H}^{(\ell)}_{i_1 i_2;\alpha_1\alpha_2} \to \delta_{i_1 i_2} H^{(\ell)}_{\alpha_1\alpha_2}$:

$$\begin{aligned}\frac{1}{n_\ell}\sum_{\substack{j,k=1\\ j\neq k}}^{n_\ell}\Bigg\{&\left(\lambda_W^{(\ell+1)}\right)^2 \mathrm{Cov}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right] \\ &+ \lambda_W^{(\ell+1)} C_W^{(\ell+1)} H^{(\ell)}_{\alpha_3\alpha_4}\mathrm{Cov}\left[\sigma^{(\ell)}_{j;\alpha}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right] \\ &+ \lambda_W^{(\ell+1)} C_W^{(\ell+1)} H^{(\ell)}_{\alpha_1\alpha_2}\mathrm{Cov}\left[\sigma'^{(\ell)}_{j;\alpha}\sigma'^{(\ell)}_{j;\alpha_2},\, \sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right] \\ &+ \left(C_W^{(\ell+1)}\right)^2 H^{(\ell)}_{\alpha_1\alpha_2} H^{(\ell)}_{\alpha_3\alpha_4}\mathrm{Cov}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right]\Bigg\}.\end{aligned} \tag{8.92}$$

All four of these covariances can be evaluated using the intralayer formula (8.41). After a little bit of algebra, this gives

$$\frac{n_\ell}{4n_{\ell-1}}\sum_{\gamma_1,\gamma_2,\gamma_3,\gamma_4} V_{(\ell)}^{(\gamma_1\gamma_2)(\gamma_3\gamma_4)}\left\langle\left(z_{\gamma_1}z_{\gamma_2}-G^{(\ell)}_{\gamma_1\gamma_2}\right)\widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\right\rangle_{G^{(\ell)}}\left\langle\left(z_{\gamma_3}z_{\gamma_4}-G^{(\ell)}_{\gamma_3\gamma_4}\right)\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}} \tag{8.93}$$

at leading order, where we again made use of the definition of $\widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}$, (8.74).

Finally, we're left with the off-diagonal contributions from the NTK fluctuations, which – if we write them out in excruciating detail – are given by

$$\begin{aligned}\frac{1}{n_\ell}\sum_{\substack{j,k=1\\ j\neq k}}^{n_\ell}\Bigg\{&\left(\lambda_W^{(\ell+1)}\right) C_W^{(\ell+1)}\mathrm{Cov}\left[\sigma^{(\ell)}_{j;\alpha_1}\sigma^{(\ell)}_{j;\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right] \\ &+ \left(\lambda_W^{(\ell+1)}\right) C_W^{(\ell+1)}\mathrm{Cov}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{\Delta H}^{(\ell)}_{jj;\alpha_1\alpha_2},\, \sigma^{(\ell)}_{k;\alpha_3}\sigma^{(\ell)}_{k;\alpha_4}\right] \\ &+ \left(C_W^{(\ell+1)}\right)^2 H^{(\ell)}_{\alpha_1\alpha_2}\mathrm{Cov}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right] \\ &+ \left(C_W^{(\ell+1)}\right)^2 H^{(\ell)}_{\alpha_3\alpha_4}\mathrm{Cov}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{\Delta H}^{(\ell)}_{jj;\alpha_1\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\right] \\ &+ \left(C_W^{(\ell+1)}\right)^2 \mathrm{Cov}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{\Delta H}^{(\ell)}_{jj;\alpha_1\alpha_2},\, \sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right]\Bigg\}.\end{aligned} \tag{8.94}$$

[13] Once again, the subleading $O(1/n)$ piece includes a contribution from the non-Gaussian distribution as well as a cross correlation contribution from the previous layer – *and* now also a contribution from the previous layer's NTK variance.

The last term, involving the two NTK fluctuations, can be evaluated similarly to how we evaluated such a term for the B-recursion in (8.86),[14] here giving

$$\left(C_W^{(\ell+1)}\right)^2 \langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\rangle_{G^{(\ell)}} \langle \sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(\ell)}} \frac{n_\ell}{n_{\ell-1}} A^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + O\left(\frac{1}{n}\right). \tag{8.96}$$

The remaining four covariances in (8.94) with only a single NTK fluctuation are identical in structure to (8.76), letting us leave the details of this to you and your roll of parchment.

At this point, let's review all the components of our expression for $A^{(\ell+1)}$: we have the diagonal contribution (8.91); and off-diagonal contributions from the NTK mean (8.93), from the covariance of two NTK fluctuations (8.96), and from the four covariances on your parchment. Assembling these components, we get the A-recursion:

$$\begin{aligned}
&A^{(\ell+1)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \\
&= \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}} - \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\right\rangle_{G^{(\ell)}} \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\right\rangle_{G^{(\ell)}} \\
&\quad + \frac{n_\ell}{4n_{\ell-1}} \sum_{\gamma_1,\gamma_2,\gamma_3,\gamma_4} V_{(\ell)}^{(\gamma_1\gamma_2)(\gamma_3\gamma_4)} \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\left(z_{\gamma_1}z_{\gamma_2} - G^{(\ell)}_{\gamma_1\gamma_2}\right)\right\rangle_{G^{(\ell)}} \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\left(z_{\gamma_3}z_{\gamma_4} - G^{(\ell)}_{\gamma_3\gamma_4}\right)\right\rangle_{G^{(\ell)}} \\
&\quad + \frac{n_\ell}{n_{\ell-1}} \left(C_W^{(\ell+1)}\right)^2 \langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\rangle_{G^{(\ell)}} \langle \sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(\ell)}} A^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} \\
&\quad + \frac{n_\ell}{n_{\ell-1}} \frac{C_W^{(\ell+1)}}{2} \sum_{\beta_1,\beta_2,\gamma_1,\gamma_2} \Bigg[\left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_1\alpha_2}\left(z_{\beta_1}z_{\beta_2} - G^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{G^{(\ell)}} G^{\beta_1\gamma_1}_{(\ell)} G^{\beta_2\gamma_2}_{(\ell)} D^{(\ell)}_{\gamma_1\gamma_2\alpha_3\alpha_4} \langle \sigma'_{\alpha_3}\sigma'_{\alpha_4}\rangle_{G^{(\ell)}} \\
&\qquad\qquad + \left\langle \widehat{\Omega}^{(\ell+1)}_{\alpha_3\alpha_4}\left(z_{\beta_1}z_{\beta_2} - G^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{G^{(\ell)}} G^{\beta_1\gamma_1}_{(\ell)} G^{\beta_2\gamma_2}_{(\ell)} D^{(\ell)}_{\gamma_1\gamma_2\alpha_1\alpha_2} \langle \sigma'_{\alpha_1}\sigma'_{\alpha_2}\rangle_{G^{(\ell)}} \Bigg] \\
&\quad + O\left(\frac{1}{n}\right).
\end{aligned} \tag{8.97}$$

As promised, we recursively see that $A^{(\ell)}$ is an order-one quantity. Additionally, we note with interest that, at leading order, this A-type contribution to the NTK variance at layer $(\ell+1)$ mixes with the four-point vertex $V^{(\ell)}$ and the cross correlation $D^{(\ell)}$ in layer ℓ, though not with $B^{(\ell)}$ or $F^{(\ell)}$.

This completes our analysis of all the finite-width effects for the NTK–preactivation joint distribution; both the leading NTK–preactivation cross correlations and the NTK variance scale as $\sim 1/n$ in the large-width expansion and vanish in the infinite-width limit.[15] These quantities are sufficient to fully characterize the leading finite-width effects of gradient-based learning.

[14] The only difference between this and the B-version before (8.86) is that here, since we're evaluating a covariance, the term

$$\mathbb{E}\left[\sigma'^{(\ell)}_{j;\alpha_1}\sigma'^{(\ell)}_{j;\alpha_2}\widehat{\Delta H}^{(\ell)}_{jj;\alpha_1\alpha_2}\right]\mathbb{E}\left[\sigma'^{(\ell)}_{k;\alpha_3}\sigma'^{(\ell)}_{k;\alpha_4}\widehat{\Delta H}^{(\ell)}_{kk;\alpha_3\alpha_4}\right] \tag{8.95}$$

is being subtracted. However, this term is of order $O\left(1/n^2\right)$ and can thus be neglected.

[15] Recalling our discussion from footnote 11 in §2.3, this means that the NTK *self-averages* in the strict infinite-width limit; in this limit, the particular value of the NTK in any instantiation of the network parameters is fixed and equal to the ensemble mean.

https://doi.org/10.1017/9781009023405.010 Published online by Cambridge University Press