5

Effective Theory of Preactivations at Initialization

> *We believe this realm of work to be immensely important and rich, but we expect its growth to require a degree of critical analysis that its more romantic advocates have always been reluctant to pursue . . . .*
>
> Minsky and Papert in the prologue to their 1988 expanded edition of *Perceptrons* [43].

The key defining feature of deep learning is the stacking of components on top of each other in order to get a *deep* neural network architecture. Despite the empirical preference for deeper networks, it's not at all obvious *why* deep is good for learning. For a fixed number of neurons per layer, deep implies many more parameters, and often in deep learning more parameters lead to better performance. But there are other ways to include more parameters. For instance, why not just have a single hidden layer that is very wide? In fact, in the strict infinite-width limit, such a single-layer model has the same number of parameters as any deeper MLP: infinity.

The proper way to think about the effects of depth is not to just count the number of model parameters but instead to ask what happens when we add an additional layer to our MLP. In §4, we developed a formalism to address exactly this question through recursions for observable quantities of interest, enabling us to compute how the distributions of initial network outputs change upon adding a layer. What we need, then, is a tool to effectively extract the explicit depth dependence from these recursions.

Building on the effective theory formalism developed in §4, in this chapter we'll extend the criticality and fluctuation analyses performed in §3 to MLPs with any nonlinear activation function. Enlightened by the success of the previous chapter in finding simplification in the wide regime ($n \gg 1$), we now seek additional simplicity in the limit of large depth ($L \gg 1$).[1] We'll first analyze the limit of an infinite number of neurons per layer and then back off this limit to consider networks of large but finite width and

[1] What this means – in light of the discussion in §3.4 – is that we take the limit of large width *first* and then look at the limit of large depth.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

depth. The result will be explicit expressions for the two-point and four-point correlators of preactivations in these asymptotic limits.[2]

This will let us address the question of what happens to input signals as they propagate through the many layers of deep neural networks at initialization (§5.1). We'll come to understand that the order-one values of the initialization hyperparameters – i.e., the bias variances $C_b^{(\ell)}$ and the rescaled weight variances $C_W^{(\ell)}$ – have pronounced qualitative effects on the behavior of the observables, just as we saw in §3 for deep linear networks. In particular, we'll explain how neural-network behavior becomes increasingly sensitive to these initialization hyperparameters with increasing depth.[3]

Such a tuning brings a network to *criticality*, a term we borrow from statistical physics used to describe self-similar systems. To this end, we give a general prescription for tuning initialization hyperparameters to their critical values for a given activation function and network architecture (§5.2 and §5.3). In the process, we also identify some activation functions that don't allow for criticality. We will also see that certain activation functions behave very similarly to each other when tuned to criticality, highlighting an important connection to the notion of *universality* in statistical physics.

The study of finite-width corrections at criticality then leads us to one of the main results of this chapter, an *emergent scale* given by the aspect ratio of the network depth to the network width, L/n (§5.4). This aspect ratio ultimately serves as the *cutoff* of our effective theory, controlling the region of validity of our effective theory as well as determining the strength and importance of the finite-width corrections to the infinite-width description. At one end of the spectrum, we find that the shorter and fatter networks are, the more and more they behave like their infinite-width counterparts. At the other end, skinny and tall networks become increasingly dominated by non-Gaussian *fluctuations* due to *interactions* between neurons. Overall, this serves to generalize our fluctuation analysis for deep linear networks in §3.3.

Lastly, we'll conclude the chapter by addressing and resolving a subtlety that arises in our criticality analysis for nonsmooth activation functions such as the `ReLU` (§5.5).

5.1 Criticality Analysis of the Kernel

For the bulk of this chapter, our goal is to extend the notion of criticality discussed in §3 to deep MLPs with general activation functions $\sigma(z)$. Our starting point is the kernel recursion

[2] Despite the asymptotic nature of these solutions, we note many of these tools were developed to study the strong interactions, where a parameter that is 3 in practice is taken to be infinity. Thus, sometimes even $3 \sim \infty$, since $1/3 \ll 1$ can be made to work as perturbative parameter [44].

[3] The initial part of this analysis was first carried out in a series of papers, [45–47], using a different set of techniques that are ultimately equivalent to ours in the infinite-width limit. Extending this analysis, we'll identify two very general conditions according to the *principle of criticality* that let us determine the correct order-one values for the initialization hyperparameters. In §10.3, we'll see that the need for these conditions can also be understood by demanding that fully-trained networks generalize well.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

$$K^{(\ell+1)}_{\alpha\beta} = C_b + C_W \left\langle \sigma_\alpha \sigma_\beta \right\rangle_{K^{(\ell)}} \,, \tag{5.1}$$

derived in the previous chapter. As a reminder, the *kernel* $K^{(\ell)}_{\alpha\beta}$, defined by (4.116), is the infinite-width limit of the mean metric $G^{(\ell)}_{\alpha\beta}$. Here, in order to restrict the number of distinct hyperparameters, we have set the bias variance $C^{(\ell)}_b = C_b$ and the rescaled weight variance $C^{(\ell)}_W = C_W$ to be layer-independent. The initial condition for this recursion is given by the first-layer kernel

$$K^{(1)}_{\alpha\beta} = C_b + C_W \left(\frac{1}{n_0} \sum_{i=1}^{n_0} x_{i;\alpha} x_{i;\beta} \right) , \tag{5.2}$$

set by the inner products of inputs $\sum_{i=1}^{n_0} x_{i;\alpha} x_{i;\beta}$. Our goal is to analyze how the kernel changes as a function of layer and, as we'll see, this analysis at the level of the kernel is sufficient to pin down the critical initialization hyperparameters to leading order in 1/width.

For deep linear networks, the Gaussian expectation of the activations with respect to the kernel is just given by the kernel, $\left\langle \sigma_\alpha \sigma_\beta \right\rangle_{K^{(\ell)}} = \left\langle z_\alpha z_\beta \right\rangle_{K^{(\ell)}} = K^{(\ell)}_{\alpha\beta}$, and the recursion equation was simple enough for us to obtain a full solution. Stepping outside the realm of the `linear` activation function, however, the kernel recursion acquires two new complications that require some care: *(i)* the expectation value $\left\langle \sigma_\alpha \sigma_\beta \right\rangle_{K^{(\ell)}}$ will be a nonlinear function of the kernel and *(ii)* for two distinct inputs $\alpha \neq \beta$, the recursion mixes off-diagonal components $K^{(\ell)}_{\alpha\beta}$ with diagonal components $K^{(\ell)}_{\alpha\alpha}$ and $K^{(\ell)}_{\beta\beta}$.

Let's illustrate this with a `quadratic` activation function $\sigma(z) = z^2$. As should be second nature by now, evaluating (5.1) requires two pairs of Wick contractions, giving

$$\begin{aligned} K^{(\ell+1)}_{\alpha\beta} &= C_b + C_W \left\langle z_\alpha^2 z_\beta^2 \right\rangle_{K^{(\ell)}} \\ &= C_b + C_W \left(K^{(\ell)}_{\alpha\alpha} K^{(\ell)}_{\beta\beta} + 2 K^{(\ell)}_{\alpha\beta} K^{(\ell)}_{\alpha\beta} \right) . \end{aligned} \tag{5.3}$$

Thus, unlike deep linear networks, for quadratic activations $K^{(\ell+1)}_{\alpha\beta}$ depends not only on $K^{(\ell)}_{\alpha\beta}$ but also on $K^{(\ell)}_{\alpha\alpha}$ and $K^{(\ell)}_{\beta\beta}$, requiring us to solve three coupled nonlinear recursion equations for $\alpha \neq \beta$. This mixing is generic for nonlinear activation functions.

For practitioners, this is good news: the off-diagonal elements of the kernel are related to the generalization ability of the network. For deep linear networks, the lack of mixing via nonlinearity suggests an *inductive bias* that limits such networks' ability to develop nontrivial correlations for pairs of samples. While mixing is a benefit in practice, it's an obstacle in theory, albeit a surmountable one. Since the kernel recursion mixes at most two inputs, it is sufficient to analyze the case with a single input and the case with two distinct inputs. We shall now perform these analyses in turn, with an eye toward deriving the general conditions for criticality.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

A Single Input

Each diagonal component of the kernel can be solved self-consistently by itself. Specifically, labeling a single input by $\alpha = 0$,

$$\begin{aligned} K_{00}^{(\ell+1)} &= C_b + C_W \left\langle \sigma(z_0)\, \sigma(z_0) \right\rangle_{K^{(\ell)}} \\ &= C_b + C_W g\Big(K_{00}^{(\ell)}\Big) \,. \end{aligned} \tag{5.4}$$

Here, we introduced a helper function

$$g(K) \equiv \left\langle \sigma(z)\, \sigma(z) \right\rangle_K \equiv \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\; e^{-\frac{z^2}{2K}}\, \sigma(z)\, \sigma(z), \tag{5.5}$$

to emphasize that the expectation $\langle \sigma(z_0)\sigma(z_0)\rangle_{K^{(\ell)}}$ is a function only of a single component of the kernel, $K_{00}^{(\ell)}$. By focusing our attention first on the single-input kernel, we can deal with the nonlinearity before confronting the mixing of kernel components.

In particular, the single-input recursion (5.4) is really telling us how the average magnitude of the preactivations for the input

$$K_{00}^{(\ell)} = \mathbb{E}\left[\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \Big(z_{i;0}^{(\ell)}\Big)^2\right] \tag{5.6}$$

changes as a function of layer ℓ, with the initial condition for the recursion set by (5.2).[4] For the same reasons we considered in §3.2, we would like that the kernel $K_{00}^{(\ell)}$ neither exponentially explodes nor exponentially vanishes. However, such exponential behavior is generic, and so for most choices of initialization hyperparameters (C_b, C_W), the kernel will either explode exponentially toward a *trivial* fixed point at infinity or collapse exponentially onto a trivial fixed point at a finite value $K_{00}^{\star}$. Thus, our first criticality condition is to mitigate this exploding or collapsing kernel problem for the single input.

There is an ancient technique used to analyze nonlinear recursions such as the single-input kernel recursion (5.4): linearization around a fixed point. Namely, we first identify a fixed point of the recursion, i.e., a value $K_{00}^{\star}$ that satisfies

$$K_{00}^{\star} = C_b + C_W g(K_{00}^{\star}) \,, \tag{5.7}$$

and then expand the kernel around it as

$$K_{00}^{(\ell)} = K_{00}^{\star} + \Delta K_{00}^{(\ell)} . \tag{5.8}$$

This expansion for the single-input recursion (5.4) results in the linearized recursion

$$\Delta K_{00}^{(\ell+1)} = \chi_{\parallel}(K_{00}^{\star})\, \Delta K_{00}^{(\ell)} + O\Big(\Delta^2\Big) \,, \tag{5.9}$$

[4] N.B. in §5.1–§5.3, as we are studying the infinite-width limit, we will omit "$+\,O(1/n)$" from equations for simplicity of notation. In §5.4 we will account for the $+\,O(1/n)$ corrections.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

where we introduced the **parallel susceptibility**

$$\begin{aligned}\chi_{\parallel}(K) &\equiv C_W g'(K) \\ &= C_W \frac{d}{dK}\left[\frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}}\sigma(z)\,\sigma(z)\right] \\ &= \frac{C_W}{2K^2}\left\langle \sigma(z)\,\sigma(z)\left(z^2-K\right)\right\rangle_K .\end{aligned} \tag{5.10}$$

The susceptibility $\chi_{\parallel}(K)$ characterizes how susceptible the kernel is to perturbations around the fixed point, hence the name: the kernel value exponentially expands away from or contracts toward the fixed-point value, according to whether $\chi_{\parallel}(K^{\star}_{00}) > 1$ or $\chi_{\parallel}(K^{\star}_{00}) < 1$. (The label *parallel* will be explained at the very end of this section.)

Thus, we see that in order to mitigate this *exploding and vanishing kernel problem* for a single input at linear order, we require the tuning of initialization hyperparameters (C_b, C_W) such that

$$\chi_{\parallel}(K^{\star}_{00}) = 1, \tag{5.11}$$

with the fixed-point value $K^{\star}_{00}$ defined implicitly through the fixed-point equation (5.7). As we shall detail later, criticality can happen in three ways depending on the choice of activation functions.

- First, as we saw for deep linear networks, the single-input kernel can be perfectly preserved as $K^{(\ell)}_{00} = K^{(1)}_{00} = C_b + C_W\left(\sum_i x_{i;0}x_{i;0}/n_0\right)$, resulting in a line of fixed points parametrized by input norms $\sum_i x_{i;0}x_{i;0}$. We will see this happening in §5.2 for *scale-invariant* activation functions due to the absence of higher-order corrections $O(\Delta^{p>1})$ in (5.9).

- Second, the kernel can slowly decay toward a fixed point $K^{\star}_{00} = 0$ for all input norms, with a power law $K^{(\ell)}_{00} \sim 1/\ell^q$ with $0 < q \le 1$. We will see this happening in §5.3.3 for a class of activation functions that include `tanh` and `sin`, due to the presence of $O(\Delta^{p>1})$ corrections in (5.9).

- Third, criticality can happen with a power-law decay toward a nonzero fixed-point value $K^{\star}_{00} \neq 0$. We will see this happening in §5.3.4 for the `SWISH` and `GELU` activation functions.

In all these cases, when initialization hyperparameters are tuned to criticality, we call $K^{\star}_{00} = K^{\star}_{00}\left(C_b^{\text{critical}}, C_W^{\text{critical}}\right)$ a **nontrivial fixed point**, distinguishing it from trivial fixed points for generic hyperparameters around which perturbations behave exponentially.

Two Inputs

Now, given two distinct inputs, let's label them with sample indices $\alpha = \pm$. For such a pair of inputs, we have three distinct kernel components to consider: $K^{(\ell)}_{++}$, $K^{(\ell)}_{--}$, and

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

$K^{(\ell)}_{+-} = K^{(\ell)}_{-+}$. The single-input analysis can be directly applied to determine the layer dependence of the diagonal components $K^{(\ell)}_{++}$ and $K^{(\ell)}_{--}$, so to complete our analysis we need to extract the layer dependence of the off-diagonal component $K^{(\ell)}_{+-}$, given solutions for the diagonal pieces. Such an analysis will yield a second criticality condition that, together with (5.11), will pin down the critical initialization hyperparameters $(C_b, C_W)^{\text{critical}}$ for a given activation function.

Ultimately, our approach will be to linearize around the degenerate limit where both inputs coincide identically, i.e., $x_{i;+}, x_{i;-} \to x_{i;0}$. In such a limit, all the ways of pairing up the two inputs are the same, and so all the components of the full kernel matrix must take the same value, i.e., $K^{(\ell)}_{++}, K^{(\ell)}_{--}, K^{(\ell)}_{+-} \to K^{(\ell)}_{00}$. Thus, each recursion degenerates to the same single-input recursion (5.4), which we know has a fixed-point value $K^{\star}_{00}$. This means that the coincident-limit solution,

$$\begin{pmatrix} K^{(\ell)}_{++} & K^{(\ell)}_{+-} \\ K^{(\ell)}_{-+} & K^{(\ell)}_{--} \end{pmatrix} = \begin{pmatrix} K^{\star}_{00} & K^{\star}_{00} \\ K^{\star}_{00} & K^{\star}_{00} \end{pmatrix} = K^{\star}_{00} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{5.12}$$

must also be a fixed point of the full kernel recursion for the two inputs.

There are three different kinds of perturbations that we need to consider in order to understand the approach of the full kernel matrix to this degenerate fixed point. The first kind corresponds to the perturbation $\Delta K^{(\ell)}_{00}$ that appeared in our single-input analysis, which controls how the average of the kernel's four components approaches the fixed-point value $K^{\star}_{00}$. Next, in backing off the coincident limit, the single input splits into two distinct inputs, $x_{i;0} \to x_{i;+}, x_{i;-}$. Then, we could imagine separating these two inputs so that they become endowed with different magnitudes, i.e., $\sum_i x^2_{i;+} \neq \sum_i x^2_{i;-}$, and follow the expected evolution of this difference through the network:

$$R^{(\ell)} \equiv \mathbb{E}\left[\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(z^{(\ell)}_{i;+}\right)^2\right] - \mathbb{E}\left[\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(z^{(\ell)}_{i;-}\right)^2\right] = K^{(\ell)}_{++} - K^{(\ell)}_{--}. \tag{5.13}$$

Such a perturbation is actually still covered by the single-input analysis, since the evolutions of the diagonal components $K^{(\ell)}_{++}$ and $K^{(\ell)}_{--}$ are mutually independent of each other, with their approach to the fixed point simply controlled by the single-input recursion.

Finally, rather than considering the difference of the squares, we could consider the square of the difference

$$D^{(\ell)} \equiv \mathbb{E}\left[\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(z^{(\ell)}_{i;+} - z^{(\ell)}_{i;-}\right)^2\right] = K^{(\ell)}_{++} + K^{(\ell)}_{--} - 2K^{(\ell)}_{+-}, \tag{5.14}$$

where to get the expression on the right-hand side we expanded the binomial and then used the definition of the kernel components. This quantity measures the magnitude of the difference between the two inputs after being passed through ℓ layers of the network and can be nonvanishing even when such inputs themselves have the same magnitudes, i.e., $\sum_i x^2_{i;+} = \sum_i x^2_{i;-}$. Importantly, this distance measure $D^{(\ell)}$ depends

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

on the off-diagonal component of the kernel, $K_{+-}^{(\ell)}$, and so we expect that analyzing this perturbation will give something new. As we will see, the approach of this third perturbation $D^{(\ell)}$ to the coincident fixed point with $D^{(\ell)}=0$ will yield a second criticality condition. Together with the single-input criticality condition (5.11), this will be sufficient to completely determine the critical initialization hyperparameters.

Let's translate the above discussion into math. To do so, we will find it convenient to project the full kernel matrix into the following basis:

$$K_{\alpha_1\alpha_2}^{(\ell)} = \begin{pmatrix} K_{++}^{(\ell)} & K_{+-}^{(\ell)} \\ K_{-+}^{(\ell)} & K_{--}^{(\ell)} \end{pmatrix} = K_{[0]}^{(\ell)}\gamma_{\alpha_1\alpha_2}^{[0]} + K_{[1]}^{(\ell)}\gamma_{\alpha_1\alpha_2}^{[1]} + K_{[2]}^{(\ell)}\gamma_{\alpha_1\alpha_2}^{[2]}, \tag{5.15}$$

where we've introduced symmetric matrices

$$\gamma_{\alpha_1\alpha_2}^{[0]} \equiv \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \gamma_{\alpha_1\alpha_2}^{[1]} \equiv \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \gamma_{\alpha_1\alpha_2}^{[2]} \equiv \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{5.16}$$

In this basis, the components of the kernel are

$$K_{[0]}^{(\ell)} = \frac{1}{4}\left[K_{++}^{(\ell)} + K_{--}^{(\ell)} + 2K_{+-}^{(\ell)}\right] = \mathbb{E}\left[\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(\frac{z_{i;+}^{(\ell)} + z_{i;-}^{(\ell)}}{2}\right)^2\right], \tag{5.17}$$

$$K_{[1]}^{(\ell)} = \frac{1}{2}\left[K_{++}^{(\ell)} - K_{--}^{(\ell)}\right] = \frac{1}{2}R^{(\ell)}, \tag{5.18}$$

$$K_{[2]}^{(\ell)} = \frac{1}{4}\left[K_{++}^{(\ell)} + K_{--}^{(\ell)} - 2K_{+-}^{(\ell)}\right] = \frac{1}{4}D^{(\ell)}. \tag{5.19}$$

This basis was strategically chosen so that both $K_{[1]}^{(\ell)}$ and $K_{[2]}^{(\ell)}$ correspond to our two natural distance measures of two distinct inputs – the difference in the magnitudes $R^{(\ell)}$ and the magnitude of the difference $D^{(\ell)}$ – and both vanish in the coincident limit. The remaining component, $K_{[0]}^{(\ell)}$, measures the overall average magnitude or the magnitude of the center of mass of the two ℓ-th-layer preactivations.

This basis has two additional nice properties that will later prove useful: *(i)* both $K_{[0]}^{(\ell)}$ and $K_{[2]}^{(\ell)}$ are even (invariant) under the parity swap of two inputs $+ \leftrightarrow -$, while $K_{[1]}^{(\ell)}$ is odd, changing its sign $K_{[1]}^{(\ell)} \to -K_{[1]}^{(\ell)}$ as $+ \leftrightarrow -$ and *(ii)* the $\gamma^{[a]}$ matrices are orthogonal.[5]

[5] This symmetry decomposition mirrors tensorial decomposition used by physicists to organize particles by their spin. Operationally, we can project out the components of any 2×2 matrix $M_{\alpha\beta}$ into components $M_{[a]}$ in the $\gamma^{[a]}$ basis by tracing over the sample indices and normalizing:

$$M_{[a]} = \frac{\sum_{\alpha,\beta} M_{\alpha\beta}\gamma_{\beta\alpha}^{[a]}}{\sum_{\alpha,\beta}\gamma_{\alpha\beta}^{[a]}\gamma_{\beta\alpha}^{[a]}}, \tag{5.20}$$

with $\alpha,\beta \in \{+,-\}$. One can easily check that the $\gamma_{\alpha\beta}^{[a]}$ matrices themselves are orthogonal under this inner product (5.20).

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Now, let's discuss perturbations around the coincident fixed point (5.12) in terms of this new basis. These perturbations have a very natural interpretation in terms of two infinitesimally-separated input points, $x_{i;+}$ and $x_{i;-}$, perturbed around a **midpoint input** $x_{i;0} \equiv (x_{i;+} + x_{i;-})/2$ as

$$x_{i;\pm} = x_{i;0} \pm \frac{1}{2}\delta x_i . \tag{5.21}$$

The dynamics of such preactivations $z^{(\ell)}_{i;\pm} \equiv z^{(\ell)}_i(x_\pm)$ then encode the evolution of these perturbed signals through the network as a function of layer depth ℓ. The coincident limit corresponds to $\delta x_i \to 0$ and, as we back off from this limit, we should be able to expand the $K^{(\ell)}_{[a]}$ components around their coincident-limit values $K^{(\ell)}_{[0]} = K^{(\ell)}_{00}$ and $K^{(\ell)}_{[1]} = K^{(\ell)}_{[2]} = 0$. This results in an expansion

$$K^{(\ell)}_{[0]} = K^{(\ell)}_{00} + \delta\delta K^{(\ell)}_{[0]} + O\left(\delta^4\right), \tag{5.22}$$

$$K^{(\ell)}_{[1]} = \delta K^{(\ell)}_{[1]} + \delta\delta\delta K^{(\ell)}_{[1]} + O\left(\delta^5\right), \tag{5.23}$$

$$K^{(\ell)}_{[2]} = \delta\delta K^{(\ell)}_{[2]} + \delta\delta\delta\delta K^{(\ell)}_{[2]} + O\left(\delta^6\right), \tag{5.24}$$

where the order of the kernel perturbation in δx is denoted by a preceding δ^p, and we used the even/odd behavior of the components $K^{(\ell)}_{[a]}$ under the parity symmetry $+ \leftrightarrow -$ to limit which terms appear in each expansion.[6] Next, we will use these expansions to determine whether the behavior of the leading perturbations $\delta K^{(\ell)}_{[1]}$ and $\delta\delta K^{(\ell)}_{[2]}$ around their fixed-point values are exponential, power-law, or constant.

To do so, we'll need to expand the original kernel recursion (5.1) order by order in δ. Before embarking on such an algebraic journey, let's think about what we should expect. At the zeroth order in δ, we'll just recover the recursion for the single-input kernel (5.4)

$$K^{(\ell+1)}_{00} = C_b + C_W g\left(K^{(\ell)}_{00}\right). \tag{5.25}$$

This should demystify our notational choice in the single-input analysis, as $K^{(\ell)}_{00}$ simply represents the kernel for a single input x_0 corresponding to the midpoint of a pair of inputs x_+, x_-, as per (5.21). Going forward, we will call $K^{(\ell)}_{00}$ the **midpoint kernel**.[7]

Next, at first order in δ, we will get the recursion

$$\delta K^{(\ell+1)}_{[1]} = \chi_{\parallel}\left(K^{(\ell)}_{00}\right) \delta K^{(\ell)}_{[1]} . \tag{5.26}$$

[6] These expansions are valid when the activation function $\sigma(z)$ is sufficiently smooth. For nonsmooth activation functions such as the ReLU, these expansions are more complicated, though still analyzable. We will consider these subtleties in more detail in §5.5.

[7] We note that $K^{(\ell)}_{[0]}$ is the kernel for the midpoint of the layer-ℓ preactivations, $(z^{(\ell)}_{i;+} + z^{(\ell)}_{i;-})/2$, which is not quite the same as the midpoint kernel $K^{(\ell)}_{00}$ for the preactivations of the midpoint input x_0 propagated to layer ℓ. The difference is expressed in (5.22) and will turn out to be negligible for quantities at leading order in the δ expansion.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

This is to be expected. On account of the parity symmetry, $\delta K_{[1]}^{(\ell+1)}$ can only be proportional to $\delta K_{[1]}^{(\ell)}$ at this order, and the proportionality factor must be none other than the parallel susceptibility, because $K_{[1]}^{(\ell)} = \frac{1}{2}\left[K_{++}^{(\ell)} - K_{--}^{(\ell)}\right]$ behaves in the same way as the single-input kernels: if the single-input kernels behave exponentially, then this difference should as well.

Lastly, at the second order in δ, we expect a recursion of the form

$$\delta\delta K_{[2]}^{(\ell+1)} = [\text{something}]\, \delta\delta K_{[2]}^{(\ell)} + [\text{something}']\left(\delta K_{[1]}^{(\ell)}\right)^2 + [\text{something}'']\, \delta\delta K_{[0]}^{(\ell)}, \quad (5.27)$$

which is the most general form it can take given the even parity symmetry of $\delta\delta K_{[2]}^{(\ell)}$, and where the [somethings] can be functions of the single-input kernel $K_{00}^{(\ell)}$. In the rest of this subsection we will derive the form of [something] and [something′], while also showing that [something″] vanishes due to the orthogonality of the $\gamma_{\alpha\beta}^{[a]}$ matrices.

Already at this heuristic level of the analysis, the bootstrapping nature of the system of equations should be clear. First, we find a solution for the midpoint kernel $K_{00}^{(\ell)}$, which then bootstraps the layer dependence of $\delta K_{[1]}^{(\ell)}$ through (5.26), the solution of which in turn feeds into (5.27) and together with $K_{00}^{(\ell)}$ bootstraps the layer dependence of $\delta\delta K_{[2]}^{(\ell)}$. In other words, rather than confronting three coupled nonlinear recursions, we can solve decoupled recursions one by one.

Deriving Bootstrapped Recursions

Now let's embark on our algebraic journey. Our goal is to expand the kernel recursion (5.1) up to the second order in δ. This requires us to evaluate $\langle\sigma(z_+)\sigma(z_+)\rangle_{K^{(\ell)}}$, $\langle\sigma(z_-)\sigma(z_-)\rangle_{K^{(\ell)}}$, and $\langle\sigma(z_+)\sigma(z_-)\rangle_{K^{(\ell)}}$ to that order, all of which are two-dimensional Gaussian integrals. Rather than treating all of these Gaussian integrals separately, we instead will evaluate the Gaussian expectation of an arbitrary function $\langle F(z_+, z_-)\rangle_{K^{(\ell)}}$ and then will plug in $F(z_+, z_-) = \sigma(z_+)\sigma(z_+)$, $\sigma(z_-)\sigma(z_-)$, or $\sigma(z_+)\sigma(z_-)$. Moreover, we will find that this general expression $\langle F(z_+, z_-)\rangle_{K^{(\ell)}}$ will come in handy in later chapters.

In order to evaluate this Gaussian expectation, it is natural to write the integral in the eigenbasis of the kernel rather than in the (z_+, z_-) coordinates. Denote such orthonormal eigenvectors by $\{\hat{e}^u, \hat{e}^w\}$, which satisfy the eigenvalue equations

$$\sum_{\beta=\pm} K_{\alpha\beta}^{(\ell)} \hat{e}_\beta^u = \lambda_u \hat{e}_\alpha^u\,, \qquad \sum_{\beta=\pm} K_{\alpha\beta}^{(\ell)} \hat{e}_\beta^w = \lambda_w \hat{e}_\alpha^w, \quad (5.28)$$

with eigenvalues λ_u and λ_w, respectively. Transforming to coordinates (u, w) defined via

$$z_\alpha(u, w) = u\hat{e}_\alpha^u + w\hat{e}_\alpha^w, \quad (5.29)$$

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

the Gaussian expectation becomes

$$\langle F(z_+, z_-)\rangle_{K^{(\ell)}} = \frac{\int du dw \ \exp\Big(-\frac{u^2}{2\lambda_u} - \frac{w^2}{2\lambda_w}\Big) \, F\Big(z_+(u,w)\,,\ z_-(u,w)\Big)}{\int du dw \ \exp\Big(-\frac{u^2}{2\lambda_u} - \frac{w^2}{2\lambda_w}\Big)} . \tag{5.30}$$

As we discussed in §1.3, this equation expresses the idea that the (u, w)-coordinate basis diagonalizes the kernel such that the distribution factorizes as $p(z_+, z_-) = p(u)p(w)$. The integral in the denominator represents the normalization factor of this factorized Gaussian expectation.

Now, we need to actually determine the eigenvalues λ_u and λ_w and eigenvectors $\{\hat{e}^u, \hat{e}^w\}$. We'll start our perturbative eigen-analysis by taking the by-now-familiar coincidental limit, $\delta \to 0$. As we discussed around (5.22), in this limit the kernel is degenerate:

$$K^{(\ell)}_{\alpha\beta} = K^{(\ell)}_{00}\gamma^{[0]}_{\alpha\beta} = K^{(\ell)}_{00}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} , \tag{5.31}$$

with the $\gamma^{[0]}_{\alpha\beta}$ component equal to the midpoint kernel $K^{(\ell)}_{[0]} = K^{(\ell)}_{00}$, and the other components vanishing. Such a matrix has the normalized eigenvectors

$$\hat{e}^u_\alpha = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} , \qquad \text{and} \qquad \hat{e}^w_\alpha = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix} , \tag{5.32}$$

with eigenvalues $\lambda_u = 2K^{(\ell)}_{00}$ and $\lambda_w = 0$, respectively.[8]

Next, let's back off from the coincidental limit and look back at the δ expansions (5.22)–(5.24) for $K^{(\ell)}_{[0,1,2]}$ around the midpoint kernel. With similar expansions for the eigenvectors $\hat{e}^{u,w}_{\pm}$ and eigenvalues $\lambda_{u,w}$, we can solve the eigenvalue equations (5.28) order by order.[9] Carrying out such expansions (in the margins or – if this isn't your personal copy of our book – in a private notebook) and solving (5.28) to second order, we find normalized eigenvectors

$$\hat{e}^u_\alpha = \begin{pmatrix} \hat{e}^u_+ \\ \hat{e}^u_- \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 + \frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}} - \frac{1}{8}\left(\frac{\delta K^{(\ell)}_{[1]}}{K^{(\ell)}_{00}}\right)^2 \\ 1 - \frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}} - \frac{1}{8}\left(\frac{\delta K^{(\ell)}_{[1]}}{K^{(\ell)}_{00}}\right)^2 \end{pmatrix} + O\Big(\delta^3\Big) , \tag{5.33}$$

$$\hat{e}^w_\alpha = \begin{pmatrix} \hat{e}^w_+ \\ \hat{e}^w_- \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 - \frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}} - \frac{1}{8}\left(\frac{\delta K^{(\ell)}_{[1]}}{K^{(\ell)}_{00}}\right)^2 \\ -1 - \frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}} + \frac{1}{8}\left(\frac{\delta K^{(\ell)}_{[1]}}{K^{(\ell)}_{00}}\right)^2 \end{pmatrix} + O\Big(\delta^3\Big) ,$$

[8] Here, the zero eigenvalue for w signifies that the matrix is degenerate. This implies that the distribution for the w coordinate is given by a Dirac delta function, $p(w) = \delta(w)$, indicating that there's really only one input in this limit.

[9] For physicists, note that this is second-order time-independent perturbation theory from quantum mechanics.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

and corresponding eigenvalues

$$\lambda_u = 2K^{(\ell)}_{00} + 2\delta\delta K^{(\ell)}_{[0]} + \frac{\left(\delta K^{(\ell)}_{[1]}\right)^2}{2K^{(\ell)}_{00}} + O\left(\delta^4\right), \tag{5.34}$$

$$\lambda_w = 2\delta\delta K^{(\ell)}_{[2]} - \frac{\left(\delta K^{(\ell)}_{[1]}\right)^2}{2K^{(\ell)}_{00}} + O\left(\delta^4\right).$$

Even if you don't have a private notebook, it's easy to check on a scrap of paper that (5.33) and (5.34) solve (5.28) to $O(\delta^2)$.

Now, having solved the eigenproblem, we can implement the change of coordinates. Before doing so, notice that the u coordinate is closely related to the coordinate z_0, the preactivation corresponding to the midpoint input. This makes it very natural to use z_0 as a coordinate instead of u. We can implement this by rescaling u as

$$\frac{u^2}{2\lambda_u} = \frac{z_0^2}{2K^{(\ell)}_{00}}, \tag{5.35}$$

changing variables in the integral (5.30) so that the Gaussian integral over u becomes a Gaussian integral over z_0 with a variance given by the midpoint kernel $K^{(\ell)}_{00}$. With this rescaling, the full coordinate transformation becomes

$$z_{\pm}(z_0, w) = z_0 \left[1 \pm \left(\frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}} \right) + \left(\frac{\delta\delta K^{(\ell)}_{[0]}}{2K^{(\ell)}_{00}} \right) + O\left(\delta^3\right) \right] + \frac{w}{\sqrt{2}} \left[\pm 1 + O(\delta) \right]. \tag{5.36}$$

Here, we can truncate the term in the square brackets multiplying w at $O(1)$, since the w coordinate has zero mean and a variance $\lambda_w = O(\delta^2)$. This means that, when performing the w integration, terms proportional to w^0 will be $O(1)$, terms proportional to w^2 will be $O(\delta^2)$, higher-order terms will be subleading, and, of course, all the odd terms will vanish. By contrast, the z_0 coordinate has zero mean and a variance $K^{(\ell)}_{00} = O(1)$, so we actually need keep terms up to $O(\delta^2)$.

Next, we need to plug this expression (5.36) into our arbitrary function

$$F(z_+, z_-) = F\Big(z_+(z_0, w),\ z_-(z_0, w)\Big), \tag{5.37}$$

now viewed as a function of the two independent Gaussian variables z_0 and w, and perform the integration over them. To do so, first we need to Taylor-expand the function in both δ and w around $F(z_0, z_0)$. This gives

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

$$\begin{aligned}F(z_+,z_-) &= F(z_0,z_0) + z_0\left(\frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}}\right)(\partial_+ - \partial_-)\,F + z_0\left(\frac{\delta\delta K^{(\ell)}_{[0]}}{2K^{(\ell)}_{00}}\right)(\partial_+ + \partial_-)\,F \\ &\quad + z_0^2\left(\frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}}\right)^2 \frac{(\partial_+ - \partial_-)^2\,F}{2} + \frac{w^2}{2}\frac{(\partial_+ - \partial_-)^2\,F}{2} \\ &\quad + (\text{odd in } w) + O\Big(\delta^3, w^2\delta, w^4\Big)\,,\end{aligned} \tag{5.38}$$

with the abbreviation $\partial_+^p\partial_-^q F \equiv \partial_+^p\partial_-^q F(z_+,z_-)|_{z_+=z_-=z_0}$. The Gaussian integral over w is simple to perform: we just replace w^2 with its variance λ_w (5.34). Finally, we will express our final answer in terms of single-variable Gaussian expectations over the variable z_0 – which, as you should recall, has a variance given by the scalar midpoint kernel $K^{(\ell)}_{00}$ – giving

$$\begin{aligned}\langle F(z_+,z_-)\rangle_{K^{(\ell)}} &= \langle F(z_0,z_0)\rangle_{K^{(\ell)}_{00}} + \left(\frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}}\right)\langle z_0\,(\partial_+ - \partial_-)\,F\rangle_{K^{(\ell)}_{00}} \\ &\quad + \left(\frac{\delta\delta K^{(\ell)}_{[0]}}{2K^{(\ell)}_{00}}\right)\langle z_0\,(\partial_+ + \partial_-)\,F\rangle_{K^{(\ell)}_{00}} \\ &\quad + \frac{1}{2}\left\langle\left[\delta\delta K^{(\ell)}_{[2]} + \left(\frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}}\right)^2\Big(z_0^2 - K^{(\ell)}_{00}\Big)\right](\partial_+ - \partial_-)^2\,F\right\rangle_{K^{(\ell)}_{00}} + O\Big(\delta^3\Big)\,.\end{aligned} \tag{5.39}$$

This completes our computation of this general expectation.

In order to apply this formula to evaluate the expectations $\langle\sigma(z_\alpha)\sigma(z_\beta)\rangle_{K^{(\ell)}}$ in the kernel recursion (5.1), recall the definitions of gamma matrices in (5.16) and note

$$[\sigma(z_\alpha)\sigma(z_\beta)]\,|_{z_+=z_-=z_0} = \sigma(z_0)\sigma(z_0)\gamma^{[0]}_{\alpha\beta}, \tag{5.40}$$

$$\{(\partial_+ - \partial_-)\,[\sigma(z_\alpha)\sigma(z_\beta)]\}\,|_{z_+=z_-=z_0} = 2\sigma'(z_0)\sigma(z_0)\gamma^{[1]}_{\alpha\beta}, \tag{5.41}$$

$$\{(\partial_+ + \partial_-)\,[\sigma(z_\alpha)\sigma(z_\beta)]\}\,|_{z_+=z_-=z_0} = 2\sigma'(z_0)\sigma(z_0)\gamma^{[0]}_{\alpha\beta}, \tag{5.42}$$

$$\Big\{(\partial_+ - \partial_-)^2\,[\sigma(z_\alpha)\sigma(z_\beta)]\Big\}\,|_{z_+=z_-=z_0} = 2\sigma''(z_0)\sigma(z_0)\gamma^{[0]}_{\alpha\beta} + 2\sigma'(z_0)\sigma'(z_0)\gamma^{[2]}_{\alpha\beta}. \tag{5.43}$$

Plugging these individually into our general expression (5.39), we get

$$\begin{aligned}&\langle\sigma(z_\alpha)\sigma(z_\beta)\rangle_{K^{(\ell)}} \\ &= \Big[\langle\sigma(z_0)\sigma(z_0)\rangle_{K^{(\ell)}_{00}} + O\Big(\delta^2\Big)\Big]\,\gamma^{[0]}_{\alpha\beta} \\ &\quad + \left[\left(\frac{\delta K^{(\ell)}_{[1]}}{K^{(\ell)}_{00}}\right)\langle z_0\sigma'(z_0)\sigma(z_0)\rangle_{K^{(\ell)}_{00}}\right]\gamma^{[1]}_{\alpha\beta} \\ &\quad + \left[\delta\delta K^{(\ell)}_{[2]}\,\langle\sigma'(z_0)\sigma'(z_0)\rangle_{K^{(\ell)}_{00}} + \left(\frac{\delta K^{(\ell)}_{[1]}}{2K^{(\ell)}_{00}}\right)^2\Big\langle\Big(z_0^2 - K^{(\ell)}_{00}\Big)\,\sigma'(z_0)\sigma'(z_0)\Big\rangle_{K^{(\ell)}_{00}}\right]\gamma^{[2]}_{\alpha\beta}.\end{aligned} \tag{5.44}$$

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

The coefficients of the matrix $\langle\sigma(z_\alpha)\sigma(z_\beta)\rangle_{K^{(\ell)}}$ in the $\gamma^{[a]}_{\alpha\beta}$ basis can be simply read off from the above expression. Therefore, plugging this into the right-hand side of the full kernel recursion (5.1), we can expand the left-hand side of that equation in this basis as

$$K^{(\ell+1)}_{\alpha\beta} = K^{(\ell+1)}_{[0]}\gamma^{[0]}_{\alpha\beta} + K^{(\ell+1)}_{[1]}\gamma^{[1]}_{\alpha\beta} + K^{(\ell+1)}_{[2]}\gamma^{[2]}_{\alpha\beta} \tag{5.45}$$

and equate both sides to find recursions for each component in this basis. These are given just below.

Summary

Just above, we explained how to derive the recursions

$$K^{(\ell+1)}_{00} = C_b + C_W\, g\Big(K^{(\ell)}_{00}\Big)\,, \tag{5.46}$$

$$\delta K^{(\ell+1)}_{[1]} = \chi_{\parallel}\Big(K^{(\ell)}_{00}\Big)\,\delta K^{(\ell)}_{[1]}\,, \tag{5.47}$$

$$\delta\delta K^{(\ell+1)}_{[2]} = \chi_{\perp}\Big(K^{(\ell)}_{00}\Big)\,\delta\delta K^{(\ell)}_{[2]} + h\Big(K^{(\ell)}_{00}\Big)\Big(\delta K^{(\ell)}_{[1]}\Big)^2\,. \tag{5.48}$$

Here the by-now familiar helper function (5.5) is defined as

$$g(K) = \langle\sigma(z)\,\sigma(z)\rangle_K\,, \tag{5.49}$$

the parallel susceptibility that we already encountered in (5.10) is given by

$$\chi_{\parallel}(K) = C_W g'(K) = \frac{C_W}{2K^2}\Big\langle\sigma(z)\,\sigma(z)\,\Big(z^2-K\Big)\Big\rangle_K = \frac{C_W}{K}\langle z\,\sigma'(z)\,\sigma(z)\rangle_K\,, \tag{5.50}$$

the **perpendicular susceptibility** is newly introduced as

$$\chi_{\perp}(K) \equiv C_W\,\langle\sigma'(z)\,\sigma'(z)\rangle_K\,, \tag{5.51}$$

and the helper function that generates perturbations $\delta\delta K^{(\ell+1)}_{[2]}$ from perturbations $\delta K^{(\ell)}_{[1]}$ is given by

$$h(K) \equiv \frac{C_W}{4K^2}\Big\langle\sigma'(z)\,\sigma'(z)\,\Big(z^2-K\Big)\Big\rangle_K = \frac{1}{2}\frac{d}{dK}\chi_{\perp}(K). \tag{5.52}$$

In the last steps of (5.50) and (5.52), we made use of the following identity for the single-variable Gaussian expectation:

$$\begin{aligned}\frac{d}{dK}\left[\frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty}dz\,e^{-\frac{z^2}{2K}}F(z)\right] &= \frac{1}{2K^2}\left[\frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty}dz\,e^{-\frac{z^2}{2K}}F(z)(z^2-K)\right] \\ &= \frac{1}{2K}\left[\frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty}dz\,e^{-\frac{z^2}{2K}}z\frac{d}{dz}F(z)\right]\,,\end{aligned} \tag{5.53}$$

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

where to go from the first line to the second line we integrated by parts. These three recursions (5.46)–(5.48) are sufficient to completely fix the initialization hyperparameters and tune the network to criticality.

The first equation (5.46) is a recursion for the midpoint kernel $K_{00}^{(\ell)}$. To analyze this equation, we look for a fixed-point value $K_{00}^{\star}$ satisfying $K_{00}^{\star} = C_b + C_W g(K_{00}^{\star})$ and then linearize around such a fixed point as $K_{00}^{(\ell)} = K_{00}^{\star} + \Delta K_{00}^{(\ell)}$. Doing so, we see that $\Delta K_{00}^{(\ell+1)} = \chi_{\parallel}(K_{00}^{\star})\,\Delta K_{00}^{(\ell)} + O(\Delta^2)$ and realize that the parallel susceptibility $\chi_{\parallel}(K_{00}^{\star})$ governs the growth/decay of deviations $\Delta K_{00}^{(\ell)}$ from the fixed-point value $K_{00}^{\star}$.

The second equation (5.47) is the first equation (5.46) in disguise, since the $\delta K_{[1]}^{(\ell)}$ component is the leading difference in magnitude of preactivations for two inputs: $R^{(\ell)} = \left(K_{++}^{(\ell)} - K_{--}^{(\ell)}\right)/2$. As such, the same susceptibility $\chi_{\parallel}\left(K_{00}^{(\ell)}\right)$ governs its growth/decay. Another perspective is that the $\delta K_{[1]}^{(\ell)}$ component can be generated by considering a perturbation $\delta x_i \propto x_{i;0}$ that is parallel to the original input $x_{i;0}$, creating a difference in the norm of the two inputs. This deviation is naturally measured by $R^{(\ell)}$, and setting $\chi_{\parallel}(K_{00}^{\star}) = 1$ ensures that such a perturbation neither exponentially explodes nor exponentially vanishes. And that, after a long-winded journey, explains why we called this susceptibility *parallel*.

This third recursion (5.48) is something new, controlling the layer dependence of the magnitude of the difference of the two inputs $D^{(\ell)} = 4\delta\delta K_{[2]}^{(\ell)} + O(\delta^4)$. Such a perturbation in layer $\ell+1$ is sourced by two types of perturbations in layer ℓ, as exhibited by the two terms on the right-hand side of (5.48). One term $\propto \left(\delta K_{[1]}^{(\ell)}\right)^2$ is generated by preactivations in the ℓ-th layer with different norms. The other term $\propto \delta\delta K_{[2]}^{(\ell)}$ is generated by preactivations in the ℓ-th layer with a nonzero difference $D^{(\ell)}$ and is present even if the preactivations have the same norm. Such same-norm perturbations in the infinitesimal regime correspond to perturbations of the input that are *perpendicular* to the midpoint input, i.e., $\sum_{i=1}^{n_0} x_{i;0}\,\delta x_i = 0$. The perpendicular susceptibility $\chi_{\perp}(K_{00}^{\star})$ determines the dynamics of such perpendicular perturbations.[10] As a nonzero distance $D^{(\ell)}$ is essential for being able to compare and contrast the two inputs $x_{i;\pm}$ after being propagated to layer ℓ, we need to ensure that this quantity is well behaved. To avoid exponential behavior, we will demand $\chi_{\perp}(K_{00}^{\star}) = 1$.

[10] An alternative view is that, for a given instantiation of the network, this perpendicular susceptibility $\chi_{\perp}\left(K_{00}^{(\ell)}\right)$ controls changes of the preactivations with respect to changes in the input. To see this, note that the distance $D^{(\ell)}$ can be rewritten to leading order in the perturbation as

$$D^{(\ell)} = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\mathbb{E}\left[\left(z_{i;+}^{(\ell)} - z_{i;-}^{(\ell)}\right)^2\right] = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\mathbb{E}\left[\left(\sum_{j=1}^{n_0}\frac{dz_{i;0}^{(\ell)}}{dx_{j;0}}\delta x_j\right)^2\right] + O\left(\delta^4\right). \tag{5.54}$$

This makes quantity $\chi_{\perp}(K_{00}^{\star})$ of interest for controlling the infamous *exploding and vanishing gradient problem*, a perspective that we will make more concrete in §9.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Taken all together, our general notion of criticality requires the following two conditions to hold:

$$\chi_{\parallel}(K_{00}^{\star}) = 1 \,, \qquad \chi_{\perp}(K_{00}^{\star}) = 1 \,, \tag{5.55}$$

with the fixed-point value of the midpoint kernel $K_{00}^{\star}$ implicitly defined via

$$K_{00}^{\star} = C_b + C_W g(K_{00}^{\star}) \,. \tag{5.56}$$

These conditions are sufficient to ensure that the entire kernel matrix is preserved to leading order, namely that

$$\Delta K_{00}^{(\ell+1)} = \Delta K_{00}^{(\ell)} + O\!\left(\Delta^2\right) , \qquad K_{[1]}^{(\ell+1)} = K_{[1]}^{(\ell)} + O\!\left(\delta^3\right) , \qquad K_{[2]}^{(\ell+1)} = K_{[2]}^{(\ell)} + O\!\left(\delta^4\right) . \tag{5.57}$$

This generalizes the notion of criticality that we discussed for deep linear networks in §3. Over the next two sections, we will give a prescription for finding these critical initialization hyperparameters $(C_b, C_W)^{\text{critical}}$ for any nonlinear activation function.[11]

5.2 Criticality for Scale-Invariant Activations

Now, let's extend our criticality analysis to *scale-invariant* activation functions by applying the formalism that we just developed. Recall from §2.2 that a scale-invariant activation function satisfies

$$\sigma(\lambda z) = \lambda \sigma(z) \,, \tag{5.58}$$

for any positive rescaling $\lambda > 0$, and always takes the form

$$\sigma(z) = \begin{cases} a_+ z \,, & z \ge 0, \\ a_- z \,, & z < 0. \end{cases} \tag{5.59}$$

As a reminder, this class of activations includes the `linear` activation – by setting $a_+ = a_- = 1$ – and the `ReLU` – by setting $a_+ = 1$ and $a_- = 0$.

These activation functions are particularly simple in that the criticality conditions, (5.55), $\chi_{\parallel}\!\left(K_{00}^{(\ell)}\right) = \chi_{\perp}\!\left(K_{00}^{(\ell)}\right) = 1$, can be solved exactly. To start, we can easily compute

[11] Note that the criticality conditions, (5.55), further underscore the need for an ensemble. In §2.3, we motivated the initialization distribution by pointing out that the *zero initialization* $b_i^{(\ell)} = W_{ij}^{(\ell)} = 0$ doesn't break the permutation symmetry among the n_ℓ neurons of a layer. Here we see more generally that any zero-mean deterministic (i.e., $C_W = 0$) distribution for the weights – which includes the zero initialization – cannot satisfy $\chi_{\parallel} = \chi_{\perp} = 1$, since both susceptibilities (5.50) and (5.51) are proportional to C_W. Such a zero-weight initialization will always suffer from an exponential decay toward a trivial fixed point at $K_{00}^{\star} = C_b$.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

$g(K)$, (5.49), which reduces to two Gaussian integrals on half the real line times an even polynomial, yielding

$$g(K) = A_2 K, \tag{5.60}$$

where we have introduced an activation-dependent constant

$$A_2 \equiv \frac{a_+^2 + a_-^2}{2}. \tag{5.61}$$

From (5.50), we see that we can find $\chi_\parallel(K)$ by differentiating this expression with respect to K and multiplying by C_W. Inspecting (5.51), we see that to get $\chi_\perp(K)$, we can perform two more simple Gaussian integrals on half the real line. Together, we find that both susceptibilities are equal and independent of $K_{00}^{(\ell)}$:

$$\chi_\parallel\left(K_{00}^{(\ell)}\right) = \chi_\perp\left(K_{00}^{(\ell)}\right) = A_2 C_W \equiv \chi. \tag{5.62}$$

Lastly $h(K)$, (5.52), identically vanishes because it is a derivative of $\chi_\perp(K)$.

With all that, we can write the general kernel recursions (5.46), (5.47), and (5.48) for scale-invariant activations as

$$K_{00}^{(\ell+1)} = C_b + \chi K_{00}^{(\ell)}\,, \tag{5.63}$$

$$\delta K_{[1]}^{(\ell+1)} = \chi \delta K_{[1]}^{(\ell)}, \tag{5.64}$$

$$\delta\delta K_{[2]}^{(\ell+1)} = \chi \delta\delta K_{[2]}^{(\ell)}. \tag{5.65}$$

These are quite simple to solve. Just as the initialization hyperparameter C_W governed the exploding and vanishing kernel problem in §3.2, the constant susceptibility $\chi = A_2 C_W$ governs the same problem here:

- If $\chi > 1$, all quantities explode exponentially in ℓ toward a trivial fixed point at infinity.
- If $\chi < 1$, the fixed-point value of the kernel is given by $K_{00}^\star = \frac{C_b}{1-\chi}$, and all perturbations around the fixed point vanish exponentially with ℓ.
- If $C_W = 1/A_2$ and $C_b = 0$, then the network is at criticality. Not only does every perturbation stay constant,[12] but also any value of $K_{00}^\star$ serves as a nontrivial fixed

[12] One caveat is in order. While the constancy of the preactivation norm $K_{[0]}^{(\ell)}$ and the parallel perturbation $K_{[1]}^{(\ell)}$ is exact, the constancy of $K_{[2]}^{(\ell)}$ is an artifact of our infinitesimal perturbation analysis. In fact, the finite-angle analysis of nonlinear scale-invariant activation functions in §5.5 describes how $K_{[2]}^{(\ell)}$ crosses over from near constancy for small ℓ to a power-law decay $\sim 1/\ell^2$ for large ℓ. In short, the preservation of the whole kernel matrix seen in §3.2 is a special property of the `linear` activation, and for nonlinear scale-invariant activation functions, there is a slow power-law decay of some observables. This power-law behavior is quite benign compared to exponential behavior and is typical at criticality.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

point, i.e., there is a *line of nontrivial fixed points.*[13] In particular, the value of the fixed point is given by

$$K^{\star}_{00} = \frac{1}{A_2}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} x_{i;0}^2\right). \tag{5.66}$$

- If $C_W = 1/A_2$ and $C_b > 0$, then $\delta K^{(\ell)}_{[1]}$ and $\delta\delta K^{(\ell)}_{[2]}$ stay constant at this infinitesimal level of analysis. However, $K^{(\ell)}_{00}$ grows linearly toward a nontrivial fixed point at infinity, with the rate set by C_b. Since the kernel does not exhibit any exponential behavior, such a network is at criticality in a broad sense. This **semi-criticality** results in a *line of semi-critical initialization hyperparameters* parameterized by C_b in the hyperparameter plane spanned by (C_b, C_W).

In conclusion, this study generalizes the analysis carried out for deep linear networks in §3.2 and identifies

$$(C_b, C_W)^{\text{critical}} = \left(0, \frac{1}{A_2}\right), \tag{5.67}$$

with $A_2 = (a_+^2 + a_-^2)/2$, as the critical initialization hyperparameters for scale-invariant activation functions.[14] For the `ReLU` activation function, this reproduces the *Kaiming initialization* $(C_b, C_W)^{\text{critical}} = (0, 2)$ [48].

5.3 Universality Beyond Scale-Invariant Activations

All of the activation functions treated in the last section shared a rather special property: scale invariance (5.58). This property gave rise to equal and kernel-independent parallel and perpendicular susceptibilities, $\chi_{\parallel}(K) = \chi$ and $\chi_{\perp}(K) = \chi$, all together enabling us to drastically simplify the criticality analysis for these activation functions.[15] Such an analysis showed that networks equipped with a scale-invariant activation function will behave similarly to each other under *representation group flow* at criticality.

In theoretical physics, systems at criticality that behave similarly under *renormalization group flow* are said to fall into the same **universality class**. The effective actions describing such systems converge under the iterative coarse-graining procedure, such that at long-range scales these systems share the same underlying mathematical model or *effective theory*, independent of the microscopic details of the particular system. This phenomenon is known as **universality** [49].

[13] For physicists, note that a similar line of fixed points often appears in scale-invariant field theories with exactly marginal deformations.

[14] We see here that our simplification of $C_b = 0$ for deep linear networks in §3 was completely warranted, *ex post facto*.

[15] The kernel-independence property follows directly from scale invariance, as any dependence would have introduced a scale into the problem.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

This motivates the use of the same term, universality class, to characterize activation functions that share the same limiting behavior under representation group flow, thus furthering the connection between *RG flow and RG flow* that we began developing in §4.6. Activation functions that form a universality class will have an identical effective description after flowing through many layers, meaning that the effective theory describing the preactivation distribution becomes independent of the fine details of the particular activation function. The power of universality is that a *single* effective theory enables us to understand criticality for the many different activation functions within the same universality class.

Clearly, all the scale-invariant activation functions form a universality class. However, the simplifications that enabled us to easily analyze this **scale-invariant universality class**, e.g., the kernel independence of the susceptibilities, do not hold for other activation functions. For activation functions such as the `sigmoid`, `tanh`, or `SWISH`, we'll need to develop a much more general algorithm to find critical initialization hyperparameters. In §5.3.1, we'll illustrate how this algorithm works, and then we'll analyze specific activation functions in §5.3.2, §5.3.3, and §5.3.4.

5.3.1 General Strategy

Let's start with some recollections. As discussed most recently in §5.1, for a generic choice of initialization hyperparameters C_b and C_W, the kernel recursion for a single input x_0,

$$K_{00}^{(\ell+1)} = C_b + C_W g\Big(K_{00}^{(\ell)}\Big) , \tag{5.68}$$

admits a fixed-point solution satisfying

$$K_{00}^{\star} = C_b + C_W g(K_{00}^{\star}) , \tag{5.69}$$

where the helper function

$$g(K) \equiv \langle \sigma(z)\sigma(z) \rangle_K \tag{5.70}$$

is understood as a function of the kernel value K. Our goal is to find critical initialization hyperparameters whose associated fixed-point value $K_{00}^{\star} = K_{00}^{\star}(C_b, C_W)$ gives rise to $\chi_{\parallel}(K_{00}^{\star}) = \chi_{\perp}(K_{00}^{\star}) = 1$.

How do we actually find these critical values? Conceptually, the most obvious route – illustrated in Figure 5.1 for the `tanh` activation function – is the following procedure:

1. For each value of C_b and C_W, with $C_b \geq 0$ and $C_W \geq 0$, find a fixed-point value of the kernel $K_{00}^{\star} = K_{00}^{\star}(C_b, C_W)$, implicitly defined via $K_{00}^{\star} = C_b + C_W g_0(K_{00}^{\star})$ with the constraint $K_{00}^{\star} \geq 0$.

2. With $K_{00}^{\star}(C_b, C_W)$, evaluate both $\chi_{\parallel}(K_{00}^{\star})$ and $\chi_{\perp}(K_{00}^{\star})$, scanning over values in the (C_b, C_W) plane until the criticality conditions $\chi_{\parallel} = 1$ and $\chi_{\perp} = 1$ are both met.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

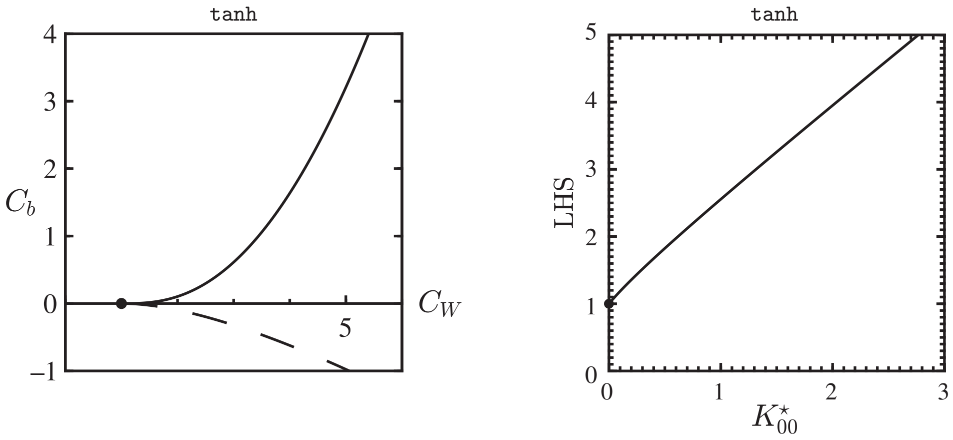

Figure 5.1 Two algorithms to pin down a nontrivial fixed point, illustrated here for the `tanh` activation function. **Left:** the lines defined by the conditions $\chi^\star_\perp = 1$ (solid) and $\chi^\star_\parallel = 1$ (dashed) are shown in the hyperparameter plane (C_W, C_b) for the `tanh` activation function. The intersection of these two lines gives the critical initialization hyperparameters $(C_W, C_b) = (1, 0)$. **Right:** the left-hand side of the condition (5.73) is plotted as a function of $K^\star_{00}$. The plotted line hits unity as $K^\star_{00} \to 0$.

This algorithm, however, is practically cumbersome to carry out for general activation functions, both numerically and analytically. In order to obtain a more implementation-friendly algorithm, let's reshuffle the logic a bit. First, note that for a candidate fixed-point value $K^\star_{00}$, setting

$$C_W = \left[\langle \sigma'(z)\sigma'(z)\rangle_{K^\star_{00}}\right]^{-1}, \tag{5.71}$$

$$C_b = K^\star_{00} - \frac{\langle \sigma(z)\sigma(z)\rangle_{K^\star_{00}}}{\langle \sigma'(z)\sigma'(z)\rangle_{K^\star_{00}}}, \tag{5.72}$$

satisfies both the fixed-point equation $K^\star_{00} = C_b + C_W g_0(K^\star_{00})$ and the first criticality condition $\chi_\perp(K^\star_{00}) = 1$. The second criticality condition $\chi_\parallel(K^\star_{00}) = 1$ is then tantamount to $\chi_\perp(K^\star_{00}) / \chi_\parallel(K^\star_{00}) = 1$, which is simply the following ratio of expectations:

$$\left[\frac{2K^2 \langle \sigma'(z)\sigma'(z)\rangle_K}{\langle \sigma(z)\sigma(z)\left(z^2 - K\right)\rangle_K}\right]\Bigg|_{K=K^\star_{00}} = 1, \tag{5.73}$$

independent of the initialization hyperparameters C_W and C_b. Therefore, we can use the following simpler algorithm:

1. Scan over values of $K^\star_{00} \geq 0$ until (5.73) is satisfied.
2. Plug the resulting value of $K^\star_{00}$ into (5.71) and (5.72) to evaluate the critical initialization hyperparameters (and also make sure $C_b \geq 0$).

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

In Figure 5.1, the left-hand side of (5.73) is plotted as a function of $K_{00}^{\star}$ for the `tanh` activation function, which we see hits unity at $K_{00}^{\star} = 0$. Then, evaluating (5.71) and (5.72) in the limit $K_{00}^{\star} \to 0$ efficiently gives the critical initialization hyperparameters for `tanh`: $(C_W, C_b) = (1, 0)$.[16]

In passing, we note that scale-invariant activation functions trivially satisfy the condition (5.73) for any fixed-point value $K_{00}^{\star}$, since the susceptibilities are equal to the same kernel-independent constant, $\chi_{\parallel}(K) = \chi_{\perp}(K) = \chi$. It's easy to check that for this universality class, the above algorithm recovers the critical initialization hyperparameters (5.67) given in §5.2.

5.3.2 No Criticality: Sigmoid, Softplus, Nonlinear Monomials, etc.

For some activation functions, a nontrivial fixed point for the kernel does not exist. For example, consider the `sigmoid` activation function

$$\sigma(z) = \frac{1}{1 + e^{-z}} . \tag{5.74}$$

The condition (5.73) is plotted for this activation in Figure 5.2. While this condition is satisfied at $K_{00}^{\star} = 0$, evaluating (5.72) in this limit yields $C_b = -\left(\frac{\sigma(0)}{\sigma'(0)}\right)^2 < 0$. Since the variance of the bias cannot be negative, this is unphysical.[17] Thus, the `sigmoid` cannot be tuned to criticality and should not be used.[18]

Next let's consider the `softplus` activation function

$$\sigma(z) = \log(1 + e^{z}) , \tag{5.75}$$

which, as a reminder, is a smooth approximation of the `ReLU`. Plotting the condition (5.73) in Figure 5.2, we see that it cannot be satisfied for any $K_{00}^{\star} \geq 0$. Thus, in contrast to the `ReLU`, the `softplus` cannot be tuned to criticality. This supports the lore in the community that the `ReLU` is superior to the `softplus`, despite their similarity and the `softplus`' smoothness.

As we will see in the next subsection, the real problem with these activation functions is that they do not cross zero at $z = 0$. There is an easy fix, namely, setting

$$\sigma(0) = 0 , \tag{5.76}$$

by an appropriate constant shift for each activation. With such a shift, the `sigmoid` turns into the `tanh`, albeit with the preactivation and activation each scaled by a half.

[16] Even though the fixed-point value of the midpoint kernel is zero, this is a *nontrivial* fixed point. In particular, we will see in §5.3.3 that kernels with a nontrivial fixed point at $K_{00}^{\star} = 0$ form a *universality class*, characterized by a benign power-law decay in ℓ. In practice, the power-law behavior means that for any finite depth, the kernel will remain finite.

[17] The limiting value of $C_b = -\left(\frac{\sigma(0)}{\sigma'(0)}\right)^2$ hints that the conditions $\sigma(0) = 0$ and $\sigma'(0) \neq 0$ may be necessary constraints for an activation function to have a nontrivial fixed point.

[18] Similarly, as a nonsmooth limit of a logistic function, the `perceptron` activation function is even worse and doesn't merit discussion.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

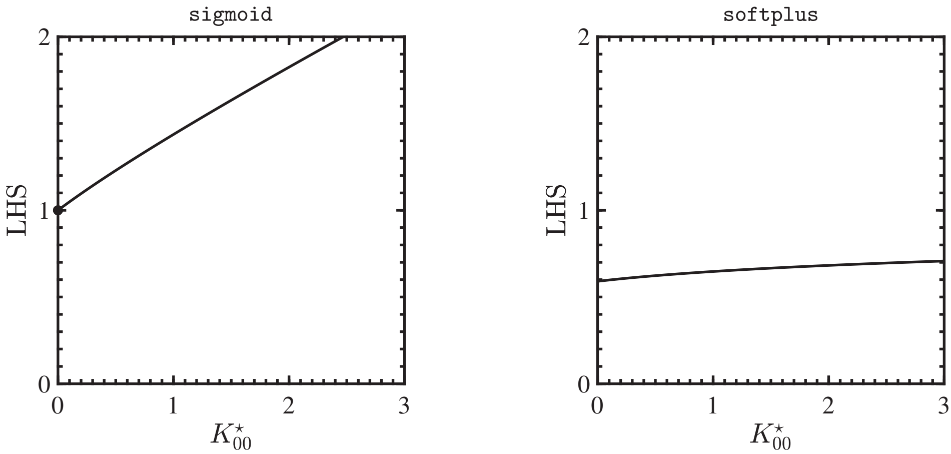

Figure 5.2 The left-hand side of the condition (5.73) is plotted as a function of $K^{\star}_{00}$ for the `sigmoid` activation function (left) and the `softplus` activation function (right). For the `sigmoid`, the plotted line hits unity as $K^{\star}_{00} \to 0$, but the associated critical initialization hyperparameters (C_b, C_W) are unphysical because $C_b < 0$. For the `softplus`, the plotted line does not hit unity. These activation functions cannot be tuned to criticality.

Such a scaled `tanh` indeed admits a critical initialization, which is easy to check after reading the discussion in the next subsection.

With that in mind, let's see what happens for activation functions that cross zero nonlinearly. For simplicity, take any nonlinear `monomial` activation function

$$\sigma(z) = z^p\,, \quad p = 2, 3, 4, \ldots . \tag{5.77}$$

In this case, direct Gaussian integration translates the condition (5.73) into the constraint

$$\frac{p}{2p-1} = 1, \tag{5.78}$$

which cannot be satisfied for nonlinear monomials, since $p \neq 1$. Thus, such nonlinear monomials also shouldn't be used in deep networks. More importantly, in addition to $\sigma(0) = 0$, criticality seems to require the condition

$$\sigma'(0) \neq 0, \tag{5.79}$$

which we will investigate more generally in the next subsection.

The impossibility of criticality for all of the activation functions discussed in this subsection means that their use should be discouraged. While the problem is somewhat mitigated for shallow networks – since there are fewer layers for the exponential behavior to damage the signals – as networks become deeper and deeper, criticality becomes more and more essential.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

5.3.3 $K^{\star} = 0$ Universality Class: tanh, sin, etc.

In §5.3.1, we learned through a numerical investigation that `tanh` has a nontrivial fixed point at $K^{\star}_{00} = 0$. In addition, in the last subsection §5.3.2, our analysis suggested that the conditions $\sigma(0) = 0$ and $\sigma'(0) \neq 0$ are important for any smooth activation to have a nontrivial fixed point.

In this subsection, we will connect these two observations. In particular, in the vicinity of $K^{\star}_{00} = 0$, we can analytically analyze the kernel recursions (5.46)–(5.48) by Taylor-expanding around $K^{\star}_{00} = 0$ and directly integrating the Gaussian expectations. This analysis will show that the conditions $\sigma(0) = 0$ and $\sigma'(0) \neq 0$ are together both necessary and sufficient for a smooth activation function to have a nontrivial fixed point at $K^{\star}_{00} = 0$, leading to the definition of our second universality class.

Let's use the following notation for the Taylor coefficients of any analytic activation function:

$$\sigma(z) = \sum_{p=0}^{\infty} \frac{\sigma_p}{p!} z^p . \tag{5.80}$$

Plugging this expansion into the definition of the helper function (5.70) and performing the Gaussian integral, we find

$$g(K) = \langle \sigma(z)\sigma(z) \rangle_K = \sigma_0^2 + \left(\sigma_1^2 + 2\sigma_0\sigma_2\right) K + O\left(K^2\right) . \tag{5.81}$$

From this we see that the fixed point of the recursion for the midpoint kernel,

$$K^{\star}_{00} = C_b + C_W g(K^{\star}_{00}) , \tag{5.82}$$

has a solution at $K^{\star}_{00} = 0$ if and only if $C_b = C_W \sigma_0^2 = 0$. Recalling that $C_W = 0$ violates the criticality conditions, we must pick $\sigma_0 = 0$. Henceforth we will assume that this choice has been made.

Continuing on with $\sigma_0 = 0$ and $C_b = 0$ in mind, inserting the expansion (5.80) into our expressions for the susceptibilities, (5.50) and (5.51), and performing the Gaussian integrals, we find

$$C_W g(K) = \left(C_W \sigma_1^2\right) \left[K + a_1 K^2 + a_2 K^3 + O\left(K^4\right)\right] , \tag{5.83}$$

$$\chi_{\parallel}(K) = \left(C_W \sigma_1^2\right) \left[1 + 2a_1 K + 3a_2 K^2 + O\left(K^3\right)\right] , \tag{5.84}$$

$$\chi_{\perp}(K) = \left(C_W \sigma_1^2\right) \left[1 + b_1 K + O\left(K^2\right)\right] , \tag{5.85}$$

where here we have also expanded $g(K)$ to higher order in the kernel, and the coefficients a_1, a_2, and b_1 are given by the following combinations of Taylor coefficients of the activation function:

$$a_1 \equiv \left(\frac{\sigma_3}{\sigma_1}\right) + \frac{3}{4}\left(\frac{\sigma_2}{\sigma_1}\right)^2 , \tag{5.86}$$

$$a_2 \equiv \frac{1}{4}\left(\frac{\sigma_5}{\sigma_1}\right) + \frac{5}{8}\left(\frac{\sigma_4}{\sigma_1}\right)\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{5}{12}\left(\frac{\sigma_3}{\sigma_1}\right)^2 , \tag{5.87}$$

$$b_1 \equiv \left(\frac{\sigma_3}{\sigma_1}\right) + \left(\frac{\sigma_2}{\sigma_1}\right)^2 . \tag{5.88}$$

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

It's easy to check that, e.g., for `tanh` these coefficients take the values $a_1 = -2$, $a_2 = 17/3$, $b_1 = -2$. Now, examining expansions (5.84) and (5.85), we see that to satisfy the criticality conditions $\chi_{\parallel}(K^{\star}_{00} = 0) = 1$ and $\chi_{\perp}(K^{\star}_{00} = 0) = 1$, we must set $C_W = 1/\sigma_1^2$. To ensure a finite variance, we also see that the activation function must have $\sigma_1 \neq 0$.

Thus, for any smooth activation function to have a nontrivial fixed point at $K^{\star}_{00} = 0$, it is necessary and sufficient that $\sigma(z)$ satisfy

$$\sigma_0 = 0\,, \qquad \sigma_1 \neq 0. \tag{5.89}$$

For such an activation, the critical initialization hyperparameters are then given by

$$(C_b, C_W)^{\text{critical}} = \left(0, \frac{1}{\sigma_1^2}\right). \tag{5.90}$$

Just to emphasize this a bit, any activation with these conditions (5.89) initialized with (5.90) will have a nontrivial fixed point at $K^{\star}_{00} = 0$. The set of activation functions that vanish at the origin with a nonzero first derivative make up the **$K^{\star} = 0$ universality class**. The canonical class member is the `tanh` activation function, though there are obviously a very large number of members in this class, e.g., the `sin` activation function is a member too.

Having determined the critical initialization hyperparameters, let's now try to understand the behavior of the kernel for the $K^{\star} = 0$ universality class. We will see that when tuned to criticality, the activations satisfying (5.89) all behave similarly under RG flow, with the large-depth behavior of the kernel depending only on the first few Taylor coefficients of $\sigma(z)$.

Deep Asymptotic Analysis for the Midpoint Kernel

Recalling the expansion $K^{(\ell)}_{00} = K^{\star}_{00} + \Delta K^{(\ell)}_{00}$ around the fixed point and considering the expansion (5.83) for $g(K)$, the midpoint kernel recursion at $K^{\star}_{00} = 0$ criticality becomes

$$\Delta K^{(\ell+1)}_{00} = \Delta K^{(\ell)}_{00} + a_1 \left(\Delta K^{(\ell)}_{00}\right)^2 + a_2 \left(\Delta K^{(\ell)}_{00}\right)^3 + O\!\left(\left(\Delta K^{(\ell)}_{00}\right)^4\right). \tag{5.91}$$

Since the whole point of criticality is to alleviate exponential behavior, we expect a gentler decay back to the $K^{(\ell)}_{00} = 0$ fixed point. With that in mind, let's plug a power-law ansatz $\Delta K^{(\ell)}_{00} \sim \left(\frac{1}{\ell}\right)^{p_0}$ into (5.91). Noting that $\left(\frac{1}{\ell+1}\right)^{p_0} = \frac{1}{\ell^{p_0}}\left[1 - \frac{p_0}{\ell} + O\!\left(\frac{1}{\ell^2}\right)\right]$ and matching the leading terms on both sides, we get a solution

$$\Delta K^{(\ell)}_{00} = \left[\frac{1}{(-a_1)}\right]\frac{1}{\ell} + \cdots. \tag{5.92}$$

Thus, the behavior at criticality is a mild power-law decay, with a **critical exponent** $p_0 = 1$. Such an exponent is said to be *universal* for the $K^{\star} = 0$ universality class, since it is completely independent of the details of the particular activation function.

Importantly, for this asymptotic solution to be consistent, we must have $(-a_1) > 0$ to ensure the positivity of the kernel. If instead we had $(-a_1) < 0$, then the asymptotic

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

solution (5.92) would be negative, making it invalid. In this case the fixed point would be unstable, exponentially repelling the kernel away from $K^\star_{00} = 0$.[19] We will see in the next subsection that the `SWISH` and `GELU` activation functions exhibit such an instability near $K^\star_{00} = 0$.

Moreover, in the last subsection we suggested that an activation function that doesn't satisfy $\sigma(0) = 0$ could be potentially salvaged with a constant shift. In particular, perhaps the `softplus` could be saved by subtracting a constant $\log(2)$ so that $\sigma(0) = 0$? However, in this case we'd have $(-a_1) < 0$, and the kernel will get repelled from the only candidate nontrivial fixed point at $K^\star_{00} = 0$. And since $\chi_{\parallel}(K) > 1$ away from $K = 0$, the midpoint kernel will diverge exponentially. Thus, despite this attempt, we see that the `softplus` cannot be saved.

Returning to our solution (5.92), we can actually do quite a bit better than "..." for the subleading asymptotic analysis. As a first guess to improve our ansatz, let's include a subleading $1/\ell^2$ term in $\Delta K^{(\ell)}_{00}$. However, if we try to match terms on both sides of (5.91), we find that there's no way of canceling the $1/\ell^3$ terms. What we can do instead is to also add $\log(\ell)/\ell^2$ with an independent coefficient to our ansatz. This generates an additional $1/\ell^3$ term, allowing for a consistent solution. Generally for any of the observables $\mathcal{O}^{(\ell)}$ that we will consider, the correct **scaling ansatz** for the large-ℓ asymptotic expansion is of the form

$$\begin{aligned}\mathcal{O}^{(\ell)} &= \left(\frac{1}{\ell}\right)^{p_{\mathcal{O}}}\left[c_{0,0} + c_{1,1}\left(\frac{\log \ell}{\ell}\right) + c_{1,0}\left(\frac{1}{\ell}\right) + c_{2,2}\left(\frac{\log^2 \ell}{\ell^2}\right) + \cdots\right] \\ &= \left(\frac{1}{\ell}\right)^{p_{\mathcal{O}}}\left[\sum_{s=0}^{\infty}\sum_{q=0}^{s} c_{s,q}\left(\frac{\log^q \ell}{\ell^s}\right)\right], \end{aligned} \tag{5.93}$$

where the critical exponent $p_{\mathcal{O}}$ is expected to be universal for a given class, while the constants $c_{s,q}$ will depend on the details of a particular activation function. Carrying this process forward for $\mathcal{O}^{(\ell)} = \Delta K^{(\ell)}_{00}$, we can systematically determine the subleading behavior of the kernel perturbation as

$$\begin{aligned}\Delta K^{(\ell)}_{00} = &\left[\frac{1}{(-a_1)}\right]\frac{1}{\ell} + \left[\frac{-(a_2 - a_1^2)}{a_1^3}\right]\frac{\log\left(\frac{\ell}{\ell_0}\right)}{\ell^2} \\ &+ \left[\frac{-\left(a_2 - a_1^2\right)^2}{a_1^5}\right]\frac{\left[\log\left(\frac{\ell}{\ell_0}\right)\right]^2}{\ell^3} + \left[\frac{\left(a_2 - a_1^2\right)^2}{a_1^5}\right]\frac{\log\left(\frac{\ell}{\ell_0}\right)}{\ell^3} + O\left(\frac{1}{\ell^3}\right),\end{aligned} \tag{5.94}$$

and with enough effort, this asymptotic expansion can be refined to arbitrary degree by including the higher-order corrections according to the scaling ansatz (5.93) described above.

[19] Generically, $(-a_1) < 0$ implies that $\chi_{\parallel} > 1$ away from $K^\star_{00} = 0$, which repels the midpoint kernel first with a power law and then exponentially. However, the semi-criticality that we discussed in §5.2 for scale-invariant activations was exceptional. For this universality class, $a_1 = 0$, and hence growth toward the fixed point at infinity is governed by a power law.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Here, the constant ℓ_0 is undetermined by this large-ℓ asymptotic analysis and nontrivially depends on the input norm through

$$K_{00}^{(1)} = \frac{1}{\sigma_1^2}\frac{1}{n_0}\sum_{i=1}^{n_0} x_{i;0}^2, \tag{5.95}$$

which sets the initial condition (5.2) for the kernel recursion (5.1) when the rescaled weight variance is set to criticality, $C_W = 1/\sigma_1^2$. To get a sense of what this means, let's assume that $\chi_{\parallel}(K)$ is monotonically decreasing for $K \geq 0$ with $\chi_{\parallel}(0) = 1$ – as is true for `tanh` – and consider what happens when an input $x_{i;0}$ has a very large magnitude. Such a large-norm input will lead to a large value for the first-layer midpoint kernel, $K_{00}^{(1)} \gg 1$. In the range $0 < k_\sharp < K_{00}^{(\ell)}$, for some constant $k_\sharp$, the kernel $K_{00}^{(\ell)}$ will decay quicker than $\chi_{\parallel}(k_\sharp)^\ell$, with $\chi_{\parallel}(k_\sharp) < 1$, until it enters the power-law regime near $K_{00}^\star = 0$. The undetermined constant ℓ_0 is a remnant of this complicated crossover behavior, capturing the leading *data dependence* of the midpoint kernel.

Additionally, the asymptotic expansion for the midpoint kernel (5.94) has a nice interpretation under RG flow. While the critical exponent of the falloff $p_0 = 1$ is generic for the universality class, we see that the coefficients of the terms do depend on the details of the activation function, albeit only the first few Taylor coefficients. In fact, for larger and larger ℓ, the dependence is on fewer and fewer of the coefficients, with the leading term only depending on a_1, (5.86). In this asymptotic limit, any activation function in the $K^\star = 0$ universality class with the same first three Taylor coefficients around zero will be completely indistinguishable. Thus, from the representation group flow perspective, one of the results of having a deeper network is to make the particular details of the activation function more and more irrelevant.

Lastly, let us note for all aspiring "activation designers" out there that we can engineer critical exponents other than $p_0 = 1$ by fine-tuning the Taylor coefficients of the activation function. For example, if we set $a_1 = 0$ by balancing σ_3 and σ_2, then the kernel approaches a $K_{00}^\star = 0$ nontrivial fixed point with a $1/\sqrt{\ell}$ power law decay so long as $(-a_2) > 0$. The need for such tuning indicates that the $\sim 1/\ell$ behavior is generic for activation functions in the $K^\star = 0$ universality class.[20]

Deep Asymptotic Analysis for Parallel Perturbations

Next, let's solve the $\delta K_{[1]}^{(\ell)}$ recursion for parallel perturbations. Plugging the expansion (5.84) for $\chi_{\parallel}(K)$ into the recursion (5.47), we get an algebraic equation

$$\delta K_{[1]}^{(\ell+1)} = \left[1 + 2a_1\Delta K_{00}^{(\ell)} + 3a_2\left(\Delta K_{00}^{(\ell)}\right)^2 + O\!\left(\left(\Delta K_{00}^{(\ell)}\right)^3\right)\right]\delta K_{[1]}^{(\ell)}. \tag{5.96}$$

[20] More precisely, we should have defined the $K^\star = 0$ universality class with the requirement $a_1 \neq 0$. This in turn would lead us to define a whole family of universality classes labeled by the degree of fine tuning of the a_1, a_2, etc., or equivalently labeled by the value of the critical exponent p_0.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Then, plugging in the large-ℓ solution for $\Delta K_{00}^{(\ell)}$ (5.94) and a large-ℓ asymptotic expansion for $\delta K_{[1]}^{(\ell)}$ based on our scaling ansatz (5.93), we can solve the resulting equation by matching the terms on both sides:

$$\delta K_{[1]}^{(\ell)} = \frac{\delta_{\parallel}}{\ell^2}\left[1 + \frac{2a_1\left(a_2 - a_1^2\right)}{a_1^3}\frac{\log\left(\frac{\ell}{\ell_0}\right)}{\ell} + O\left(\frac{1}{\ell}\right)\right]. \tag{5.97}$$

Inspecting our solution, we identify our second critical exponent for the $K^\star = 0$ universality class: $p_{\parallel} = 2$, corresponding to the $1/\ell^2$ falloff of $\delta K_{[1]}^{(\ell)}$. The particular value of this exponent is to be expected. As noted before, the parallel perturbation is just a difference of single-input kernels for two inputs with differing norms, $K_{[1]}^{(\ell)} = \left(K_{++}^{(\ell)} - K_{--}^{(\ell)}\right)/2$. The leading $1/\ell^2$ scaling occurs because the diagonal components $K_{++}^{(\ell)}$ and $K_{--}^{(\ell)}$ are governed by the same asymptotic behavior up to order $\log(\ell)/\ell^2$, including the same coefficients. Thus, the leading difference appears at order $1/\ell^2$, due to different input-dependent constants ℓ_+ and ℓ_- in expansions analogous to (5.94) for $K_{++}^{(\ell)}$ and $K_{--}^{(\ell)}$, with the undetermined constant $\delta_{\parallel} \propto \log(\ell_+/\ell_-)$. In this way, this constant explicitly carries the data dependence of the parallel perturbation.

Deep Asymptotic Analysis for Perpendicular Perturbations

Finally, let's conclude our analysis by solving the $\delta\delta K_{[2]}^{(\ell)}$ recursion for perpendicular perturbations. Let's begin by plugging the expansion (5.85) for $\chi_{\perp}(K)$ into the recursion (5.48). Since we want to focus on perpendicular perturbations with $\sum_{i=1}^{n_0} x_{i;0}\,\delta x_i = 0$, we will also turn off parallel perturbations by setting $\delta K_{[1]}^{(\ell)} = 0$. Putting this all together gives an algebraic equation

$$\delta\delta K_{[2]}^{(\ell+1)} = \left[1 + b_1 \Delta K_{00}^{(\ell)} + O\left(\left(\Delta K_{00}^{(\ell)}\right)^2\right)\right]\delta\delta K_{[2]}^{(\ell)}. \tag{5.98}$$

Plugging in the large-ℓ asymptotic solution for $\Delta K_{00}^{(\ell)}$ and solving the resulting equation with another large-ℓ asymptotic expansion for $\delta\delta K_{[2]}^{(\ell)}$ based on our scaling ansatz (5.93), we get

$$\delta\delta K_{[2]}^{(\ell)} = \frac{\delta^2}{\ell^{\frac{b_1}{a_1}}}\left[1 + \frac{b_1\left(a_2 - a_1^2\right)}{a_1^3}\frac{\log\left(\frac{\ell}{\ell_0}\right)}{\ell} + O\left(\frac{1}{\ell}\right)\right], \tag{5.99}$$

where δ^2 is another unfixed constant undetermined by the large-ℓ solution, in this case related nontrivially to the magnitude of the difference of the inputs: $\sum_{i=1}^{n_0}\left(x_{i;+} - x_{i;-}\right)^2$. Here we see that the presumptive critical exponent, $p_{\perp} \equiv b_1/a_1$, depends mildly on the details of the activation function.

However, note that something nice happens for odd activation functions such as `tanh` and `sin`. In this case, we see from (5.86) and (5.88) that $a_1 = b_1$, giving us a bona fide

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

critical exponent, $p_\perp = 1$, when restricting the universality class to odd activations. This means that perpendicular perturbations decay with the same power in ℓ as the midpoint kernel decays to the fixed point, $\sim 1/\ell$. Thus, at criticality the ratio $K^{(\ell)}_{[2]}/K^{(\ell)}_{[0]}$ is fixed at the leading order, preserving the angles between nearby perpendicular inputs. Importantly, this ensures that the relationship between input points is conserved under the RG flow, even if the signals propagate through a very deep network.

Furthermore, the milder falloff of the perpendicular perturbations suggests that they are in some sense more important than the parallel ones. This is because with enough depth, the $K^{(\ell)}_{[1]}$ component will become subleading to the $K^{(\ell)}_{[0]}$ and the $K^{(\ell)}_{[2]}$ components, due to the $1/\ell^2$ scaling of the former compared to the $1/\ell$ scaling of the latter two. For this reason, we are going to ignore these parallel perturbations of the kernel going forward.

5.3.4 Half-Stable Universality Classes: SWISH, etc. and GELU, etc.

In this final subsection, we consider two other semi-popular activation functions in order to explore nontrivial fixed points away from zero, $K^\star_{00} \neq 0$.

- The `SWISH` activation function is defined as

$$\sigma(z) = \frac{z}{1 + e^{-z}}. \tag{5.100}$$

Similar to the intuition for the `softplus`, the `SWISH` is intended as a smooth version of the `ReLU`. Following our general algorithm in §5.3.1 for finding the critical initialization hyperparameters, we actually find two nontrivial fixed points for the

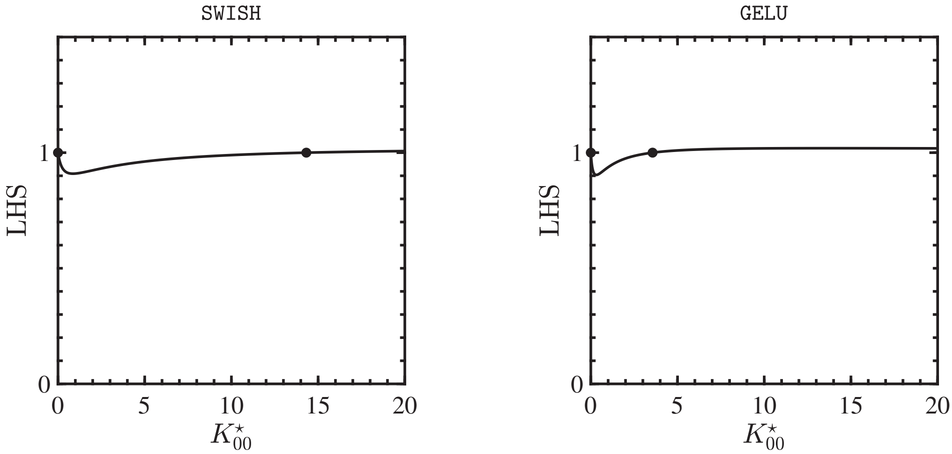

Figure 5.3 The left-hand side of the condition (5.73) is plotted as a function of $K^\star_{00}$ for the `SWISH` activation function (left) and the `GELU` activation function (right). For both activation functions, the plotted line hits unity (black dots) at $K^\star_{00} = 0$ as well as at a nonzero half-stable nontrivial fixed point $K^\star_{00} \neq 0$.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

kernel, see Figure 5.3. In particular, the condition (5.73) is met at $K_{00}^{\star} = 0$ with $(C_b, C_W) = (0, 4)$ and at $K_{00}^{\star} \approx 14.32017362$ with

$$(C_b, C_W) \approx (0.55514317, 1.98800468) . \tag{5.101}$$

For the $K_{00}^{\star} = 0$ nontrivial fixed point, one can check that $(-a_1) < 0$, and hence it's unstable. For the $K_{00}^{\star} \approx 14.3$ nontrivial fixed point, we expand the midpoint kernel recursion as $K_{00}^{(\ell)} = K_{00}^{\star} + \Delta K_{00}^{(\ell)}$, yielding

$$\Delta K_{00}^{(\ell+1)} = \Delta K_{00}^{(\ell)} + \tilde{a}_1 \left(\Delta K_{00}^{(\ell)}\right)^2 + O\!\left(\left(\Delta K_{00}^{(\ell)}\right)^3\right) , \tag{5.102}$$

with $(-\tilde{a}_1) \approx -2.84979219 \cdot 10^{-6}$.

Here, the large-ℓ asymptotic analysis around the finite fixed point is identical to the case of $K_{00}^{\star} = 0$, resulting in

$$\Delta K_{00}^{(\ell)} \sim \left[\frac{1}{(-\tilde{a}_1)}\right] \frac{1}{\ell} . \tag{5.103}$$

However, the interpretation is slightly different, because the fixed-point value $K_{00}^{\star} \approx 14.3$ is nonvanishing. In particular, this implies that when $K_{00}^{(\ell)} < K_{00}^{\star}$ the kernel is attracted to the fixed point, while when $K_{00}^{(\ell)} > K_{00}^{\star}$ the kernel is repelled.[21] Hence, this fixed point is **half-stable**, and so the activation function is perhaps half-useful. In practice, however, $|\tilde{a}_1|$ is small enough that the SWISH behaves in an almost scale-invariant manner around $K_{00}^{(\ell)} \sim K_{00}^{\star} \approx 14.3$.

- The GELU activation is defined as

$$\sigma(z) = \frac{z}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right] , \tag{5.104}$$

and as a reminder is another smoothed ReLU. Following our recipe for criticality, the condition (5.73) is again met twice, at $K_{00}^{\star} = 0$ with $(C_b, C_W) = (0, 4)$ and at $K_{00}^{\star} = \frac{3+\sqrt{17}}{2}$ with

$$(C_b, C_W) \approx (0.17292239, 1.98305826) , \tag{5.105}$$

see Figure 5.3. Similar to the SWISH, the fixed point at $K_{00}^{\star} = 0$ is unstable with $(-a_1) = -6/\pi < 0$, and the fixed point at $K_{00}^{\star} = \frac{3+\sqrt{17}}{2}$ is half-stable, in this case with $(-\tilde{a}_1) \approx (1.43626419) \cdot 10^{-4}$. Note that the sign of $\tilde{a}_1$ here differs from the sign for the SWISH. Thus, this time, when $K_{00}^{(\ell)} > K_{00}^{\star}$ the midpoint kernel is attracted to the fixed point, while when $K_{00}^{(\ell)} < K_{00}^{\star}$ it is repelled.[22] Note that the absolute value $|\tilde{a}_1|$ is bigger for the GELU than for the SWISH, meaning that it behaves less scale-invariantly and looks less like the ReLU.

[21] With the half-critical initialization hyperparameters for the SWISH (5.101), there is a *trivial* fixed point at $K_{00}^{\star} \approx 14.5$ that exponentially attracts the midpoint kernel when $K_{00}^{(\ell)} > 14.3$.

[22] With the half-critical initialization hyperparameters for the GELU (5.105), there is a *trivial* fixed point at $K_{00}^{\star} \approx 3.2$ that exponentially attracts the midpoint kernel when $K_{00}^{(\ell)} < \frac{3+\sqrt{17}}{2} \approx 3.6$.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Unlike the shifted `softplus`, which admits only an unstable nontrivial fixed point at $K^\star_{00} = 0$, here the nonmonotonicity of the `GELU` and `SWISH` activation functions gave rise to half-stable nontrivial fixed points at $K^\star_{00} \neq 0$. They are both representatives of *half-stable universality classes*. For both of these `ReLU`-like activations functions, the critical initialization hyperparameters for the $K^\star_{00} \neq 0$ half-stable nontrivial fixed points are very similar to the critical `ReLU` initialization $(C_b, C_W) = (0, 2)$; the activations in each of these classes really are just small perturbations of the `ReLU`. At the same time, the fact that there's a fixed point at a particular kernel value $K^\star_{00} \neq 0$ indicates – however weakly – the introduction of a particular scale. This is one way to see that these universality classes break scale invariance.

In summary, despite being `ReLU`-like and also smooth, both the `SWISH` and the `GELU` are likely inferior to the `ReLU` itself.

5.4 Fluctuations

Now that we fully understand how to tune infinite-width networks to criticality, let's back off this large-n limit to analyze the behavior of realistic networks. Specifically, we're going to extend the finite-width analysis that we performed for deep linear networks in §3.3 to MLPs with nonlinear activation functions. Before diving in, let's review the motivation for carrying out such an analysis.

First, note that practitioners only use a single network rather than an ensemble of networks.[23] As we have discussed, sometimes a single instantiation will generically deviate from the mean. Therefore, in order to understand what *typically* happens in a single instantiation for an observable of interest, we have to compute not only the mean but also instantiation-to-instantiation fluctuations around the mean. As we explained in §3.3, such fluctuations are generically finite-width effects, controlled by the $1/n$-suppressed four-point vertex $V^{(\ell)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}$. If fluctuations are large, then a single instantiation can behave poorly, despite being sampled from an initialization distribution tune to criticality.

Second, we saw in §4.3 that the infinite-width ℓ-th-layer preactivation distribution factorizes as

$$p\left(z^{(\ell)}_1, \ldots, z^{(\ell)}_{n_\ell}\middle|\mathcal{D}\right) = p\left(z^{(\ell)}_1\middle|\mathcal{D}\right)\cdots p\left(z^{(\ell)}_{n_\ell}\middle|\mathcal{D}\right) + O\left(\frac{1}{n_\ell}\right), \tag{5.106}$$

where the distributions $p\left(z^{(\ell)}_i\middle|\mathcal{D}\right)$ on each neuron are given by statistically independent Gaussian distributions. (To emphasize the neural dependence here, we have included neural indices while suppressing sample indices.) Recalling our discussion of interactions and statistical independence in §1.3, this means that intralayer correlations among neurons are entirely finite-width phenomena. Later, we will show how this lack of interactions connects to the fact that the representations of an infinite-width network cannot

[23] Actually in some cases practitioners can use ensembles of networks, though the computational cost of such models grows in proportion to the number of networks in the ensemble.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

evolve during gradient-based learning. Thus, understanding these finite-width effects is a prerequisite to understanding how practical networks actually learn from input data.[24]

Third, finite-width corrections can modify the mean value of observables. As we saw in §4.5, at finite width all observables in principle receive an infinite series of subleading corrections. For instance, a possible finite-width NLO correction to the metric, $G^{\{1\}(\ell)}_{\alpha_1\alpha_2}$, can shift the infinite-width metric, $G^{\{0\}(\ell)}_{\alpha_1\alpha_2} \equiv K^{(\ell)}_{\alpha_1\alpha_2}$, a.k.a. the kernel. Such a finite-width correction could potentially ruin criticality, since our derivation of the critical initialization hyperparameters depended explicitly on the infinite-width fixed-point value of the kernel.[25]

There will be two main takeaways from this section.

- First, we will find that the leading finite-width fluctuations scale with the depth-to-width ratio of the network, L/n. We saw the importance of this *emergent scale* for the `linear` activation function in §3.3; here, we see that it persists very generally for nonlinear activation functions. In the language of §4.6, this means that finite-width corrections are *relevant* under representation group flow and that deeper networks deviate more and more from the simple infinite-width limit. This emphasizes the importance of including such corrections when analyzing such networks and – taking into account the fact that overly deep networks suffer from overwhelming fluctuations – suggests that our perturbative effective theory works best in the regime where practical networks also work best.

- Second, the NLO metric $G^{\{1\}(\ell)}_{\alpha_1\alpha_2}$ is subdominant to the kernel $K^{(\ell)}_{\alpha_1\alpha_2}$ as long as an appropriate $O(1/n)$ correction is made to C_W. This means that the NLO metric vanishes in the interpolating limit – $n, L \to \infty$, with L/n fixed – and thus can safely be neglected for most wide networks of reasonable depths.

A Single Input, Reloaded

In order to illustrate the important qualitative effects of finite width, we will again specialize to just a single input. The reason for this choice can be best understood by progressing through another twofold list:

(i) Once the two initialization hyperparameters, C_b and C_W, are tuned to criticality at leading order by the one- and two-input analysis of the kernel, the only additional tuning comes from the single-input analysis of the NLO metric $G^{\{1\}(\ell)}_{\alpha_1\alpha_2}$. Therefore, the multi-input solutions for the vertex and NLO metric do not add anything to the criticality analysis.

(ii) The most interesting part of the two-input vertex is a component that gives variance of the input-output Jacobian of the network. (As we described in footnote 10,

[24] We'll go into more detail about the role that these correlations play in the inductive bias of MLPs in §6 and then connect these interactions to representation learning in §11.

[25] In §5.4.1 we will show that the NLO metric $G^{\{1\}(\ell)}_{\alpha_1\alpha_2} = 0$ vanishes for the scale-invariant universality class, which is why we didn't discuss this type of correction for deep linear networks in §3.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

the mean value of this Jacobian is captured by the $K_{[2]}^{(\ell)}$ component of the kernel.) However, the would-be analysis of this input-output variance will be subsumed by our analysis of the variance of the neural tangent kernel in §8, which more directly gives the variance of gradients relevant for training.

In the rest of this section we'll omit the $\alpha = 0$ sample indices, since such notation is unnecessarily cumbersome when considering only a single input. We'll also simplify things further by picking all the hidden-layer widths to be equal:

$$n_1 = n_2 = \cdots = n_{L-1} \equiv n. \tag{5.107}$$

In addition to being a sensible choice, this means notationally that we don't have to carry around factors of $n_\ell/n_{\ell-1}$ everywhere. With these decisions in mind, the relevant recursions from §4 become

$$K^{(\ell+1)} = C_b + C_W g\left(K^{(\ell)}\right), \tag{5.108}$$

$$V^{(\ell+1)} = \chi_\parallel^2\left(K^{(\ell)}\right) V^{(\ell)} + C_W^2\left[\left\langle \sigma^4(z)\right\rangle_{K^{(\ell)}} - \left\langle \sigma^2(z)\right\rangle_{K^{(\ell)}}^2\right], \tag{5.109}$$

$$G^{\{1\}(\ell+1)} = \chi_\parallel\left(K^{(\ell)}\right) G^{\{1\}(\ell)} + \frac{1}{8} j\left(K^{(\ell)}\right)\frac{V^{(\ell)}}{\left(K^{(\ell)}\right)^2}, \tag{5.110}$$

where the helper function $g(K)$ and the parallel susceptibility $\chi_\parallel(K)$ were defined in (5.5) and (5.50), and we have defined another helper function

$$j(K) \equiv C_W \left\langle \sigma(z)\,\sigma(z)\left[\left(\frac{z^2}{K}\right)^2 - 6\left(\frac{z^2}{K}\right) + 3\right]\right\rangle_K . \tag{5.111}$$

These three recursions can be solved for each universality class by mirroring our bootstrap analysis of $K_{00}^{(\ell)}$, $\delta K_{[1]}^{(\ell)}$, $\delta\delta K_{[2]}^{(\ell)}$ in §5.2 and §5.3.

5.4.1 Fluctuations for the Scale-Invariant Universality Class

Recall from §5.2 that the scale-invariant universality class contains any activation function of the form

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0,\\ a_- z\,, & z < 0,\end{cases} \tag{5.112}$$

with the `ReLU` $(a_+ = 1, a_- = 0)$ as the exemplar member to keep in mind. Also recall that for this class we evaluated the helper function as $g(K) = A_2 K$ and the parallel susceptibility as $\chi_\parallel = A_2 C_W \equiv \chi$, with the activation-dependent constant given by $A_2 \equiv (a_+^2 + a_-^2)/2$. The other terms in the new recursions (5.109) and (5.110) can similarly be evaluated by computing Gaussian integrals on the half-line, yielding

$$C_W^2\left[\left\langle \sigma^4(z)\right\rangle_K - \left\langle \sigma^2(z)\right\rangle_K^2\right] = C_W^2\left(3A_4 - A_2^2\right)K^2\,, \qquad j(K) = 0, \tag{5.113}$$

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

with a new activation-dependent constant

$$A_4 \equiv \frac{a_+^4 + a_-^4}{2}, \tag{5.114}$$

to pair with our other constant, A_2. With these expressions, the three recursions can be simplified as

$$K^{(\ell+1)} = C_b + \chi K^{(\ell)}, \tag{5.115}$$

$$V^{(\ell+1)} = \chi^2 \left(\frac{3A_4}{A_2^2} - 1\right)\left(K^{(\ell)}\right)^2 + \chi^2 V^{(\ell)}, \tag{5.116}$$

$$G^{\{1\}(\ell+1)} = \chi\, G^{\{1\}(\ell)}. \tag{5.117}$$

As a reminder, we already solved the kernel recursion in §5.2.

Things are now quite simple.

- First, remember from §4.1 that the first-layer preactivation distribution is always exactly Gaussian, implying that the first-layer two-point correlator is simply given in terms of the first-layer kernel $K^{(1)}$ to all orders in n:

  $$\mathbb{E}\left[z_i^{(1)} z_j^{(1)}\right] = \delta_{ij} K^{(1)}. \tag{5.118}$$

  This means that the first-layer NLO metric must vanish, $G^{\{1\}(1)} = 0$, and recursion (5.117) then tells us that the NLO metric will vanish in any subsequent layer. Thus, for activations in the scale-invariant universality class, we learn that the single-input metric does not get corrected at $O(1/n)$.

- Second, let's focus on criticality by setting $C_b = 0$ and $C_W = 1/A_2$. As discussed in §5.2, this setting of hyperparameters fixes the kernel to be an input-dependent layer-independent constant

  $$K^{(\ell)} = K^\star \equiv \frac{1}{A_2}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} x_i^2\right). \tag{5.119}$$

  In particular, this means that the critical exponent for the single-input kernel is given by $p_0 = 0$. Setting $\chi = 1$ and substituting this expression into (5.116), we find a linearly growing solution for the four-point vertex

  $$V^{(\ell)} = (\ell - 1)\left(\frac{3A_4}{A_2^2} - 1\right)(K^\star)^2. \tag{5.120}$$

  By inspection, we identify another critical exponent for the scale-invariant universality class: assuming $V^{(\ell)} \sim (1/\ell)^{p_V}$, then $p_V = -1$. This exponent encodes the linear growth of the vertex under RG flow. Of particular note, the coefficient in front of (5.120) evaluates to $\left(\frac{3A_4}{A_2^2} - 1\right) = 2$ for `linear` activations in contrast to $= 5$ for `ReLU` activations. Apparently the fluctuations in `ReLU` networks are significantly stronger than in deep linear networks. More generally, we conclude that the strength of such fluctuations is *not* universal.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

- Third, let's revisit semi-criticality by setting $C_W = 1/A_2$ but setting the bias variance to an arbitrary positive constant, $C_b > 0$. As we saw in §5.2, in this case the kernel grows linearly toward a nontrivial fixed point at infinity, $K^{(\ell)} \sim \ell$, i.e., $p_0 = -1$. Plugging such a solution into the vertex recursion (5.116), we see that the four-point vertex grows cubicly, $V^{(\ell)} \sim \ell^3$, i.e., $p_V = -3$. However, the appropriate dimensionless quantity – normalizing the vertex by the square of the kernel – still grows linearly in ℓ, i.e., $p_V - 2p_0 = -1$.[26] Thus, even for semi-criticality the universal ℓ/n-scaling of the finite-width corrections is preserved.

5.4.2 Fluctuations for the $K^\star = 0$ Universality Class

Let's now consider the $K^\star = 0$ universality class. As a reminder, this class contains all smooth activation functions that satisfy $\sigma(0) = 0$ and $\sigma'(0) \neq 0$, with `tanh` as the exemplar member to keep in mind. In §5.3.3, we determined that activations in this class have a nontrivial fixed point at $K^\star = 0$ and found that the associated critical initialization hyperparameters are given by $C_b = 0$ and $C_W = 1/\sigma_1^2$. For the rest of this subsection we will focus on such networks at criticality.

Mirroring our approach in §5.3.3 to solve the kernel recursions, we can evaluate the Gaussian expectations in the vertex recursion (5.109) and the NLO-metric recursion (5.110) by Taylor-expanding the activation around $z = 0$ and explicitly computing the Gaussian integrals. Keeping in mind the criticality condition $C_W = 1/\sigma_1^2$, this gives the following expressions:

$$\chi_\parallel(K) = 1 + 2a_1 K + 3a_2 K^2 + O\left(K^3\right), \tag{5.121}$$

$$C_W^2 \left[\left\langle \sigma^4(z)\right\rangle_K - \left\langle \sigma^2(z)\right\rangle_K^2\right] = 2K^2 + \left(-52a_1 + 60b_1\right) K^3 + O\left(K^4\right), \tag{5.122}$$

$$\frac{j(K)}{8K^2} = a_1 + 3a_2 K + O\left(K^2\right). \tag{5.123}$$

Here, the expression for $\chi_\parallel(K)$ is simply reprinted from §5.3.3. Similarly, to limit the amount of time you have to flip back and forth, let us also reprint the large-ℓ asymptotic expansion of the kernel perturbation originally given by (5.94):

$$\Delta K^{(\ell)} = \left[\frac{1}{(-a_1)}\right]\frac{1}{\ell} + \left[\frac{-(a_2 - a_1^2)}{a_1^3}\right]\frac{\log\left(\frac{\ell}{\ell_0}\right)}{\ell^2} + \left[\frac{-\left(a_2 - a_1^2\right)^2}{a_1^5}\right]\frac{\left[\log\left(\frac{\ell}{\ell_0}\right)\right]^2}{\ell^3} + \left[\frac{\left(a_2 - a_1^2\right)^2}{a_1^5}\right]\frac{\log\left(\frac{\ell}{\ell_0}\right)}{\ell^3} + O\left(\frac{1}{\ell^3}\right). \tag{5.124}$$

[26] To elaborate a bit more, first please reread footnote 15 in §1.3 on *dimensional analysis*. Now, if we give the preactivations a dimension $[z] = \zeta$, then we have for the kernel $[K] = \zeta^2$, while for the four-point vertex $[V] = \zeta^4$. Thus, the ratio V/K^2 is dimensionless.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Four-Point Vertex

Now, let's find a solution for the four-point vertex. Substituting (5.121) and (5.122) into the single-input vertex recursion (5.109) gives an algebraic equation

$$V^{(\ell+1)} = V^{(\ell)}\left[1 + 4a_1\Delta K^{(\ell)} + \left(6a_2 + 4a_1^2\right)\left(\Delta K^{(\ell)}\right)^2 + \cdots\right] + 2\left(\Delta K^{(\ell)}\right)^2 + \left(-52a_1 + 60b_1\right)\left(\Delta K^{(\ell)}\right)^3 + \cdots . \tag{5.125}$$

Using our scaling ansatz (5.93) for the large-ℓ asymptotic expansion

$$V^{(\ell)} = \left(\frac{1}{\ell}\right)^{p_V}\left[\# + \#'\frac{\log\ell}{\ell} + \frac{\#''}{\ell} + \cdots\right], \tag{5.126}$$

and (5.124) for $\Delta K^{(\ell)}$, and then matching terms, we find

$$V^{(\ell)} = \left[\frac{2}{3a_1^2}\right]\frac{1}{\ell} + \left[\frac{2(a_2 - a_1^2)}{3a_1^4}\right]\frac{\log\left(\frac{\ell}{\ell_0}\right)}{\ell^2} + \left[\frac{5a_2 + a_1(82a_1 - 90b_1)}{3a_1^4}\right]\frac{1}{\ell^2} + O\left(\frac{\log^2(\ell)}{\ell^3}\right), \tag{5.127}$$

where the constant scale ℓ_0 is same as the one in the $\Delta K^{(\ell)}$ expansion just above, again carrying the data dependence of the solution. We can also read off the critical exponent controlling the asymptotic falloff of the vertex for the $K^\star = 0$ universality class: $p_V = 1$.

Note that the value of the exponent $p_V = 1$ and the behavior of the four-point vertex $V^{(\ell)} \sim 1/\ell$ here are different from the value of the exponent $p_V = -1$ and the associated behavior $V^{(\ell)} \sim \ell$ that we found for the scale-invariant universality class. Also note that we saw this difference in the behavior of the kernel, $p_0 = 1$ vs. $p_0 = 0$, for the $K^\star = 0$ and scale-invariant classes, respectively. However, when instead considering the dimensionless quantity

$$\frac{V^{(\ell)}}{n\left(K^{(\ell)}\right)^2} \sim \frac{1}{n}\left(\frac{1}{\ell}\right)^{p_V - 2p_0} + \cdots , \tag{5.128}$$

we see that its scaling is consistent across both classes of activations:

$$p_V - 2p_0 = -1. \tag{5.129}$$

Thus, this **scaling law** holds across different universality classes. As the normalized quantity (5.128) controls leading finite-width corrections to observables – this was discussed in detail in §3.3 – such a law means that these corrections are always *relevant* under representation group flow.

Concretely, the normalized vertex function is given by

$$\frac{V^{(\ell)}}{n\left(K^{(\ell)}\right)^2} = \left(\frac{3A_4}{A_2^2} - 1\right)\frac{\ell}{n} + O\left(\frac{1}{n}\right) \tag{5.130}$$

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

for the scale-invariant universality class and

$$\frac{V^{(\ell)}}{n\left(K^{(\ell)}\right)^2} = \left(\frac{2}{3}\right)\frac{\ell}{n} + O\!\left(\frac{\log(\ell)}{n}\right) \tag{5.131}$$

for the $K^{\star} = 0$ universality class. Of practical relevance, this means that `ReLU` networks and `tanh` networks of the same depth and width will have a mostly similar sensitivity to such corrections. However, the $O(1)$ coefficient of this quantity *does* depend on the particular activation function: $= 5$ for `ReLU` and $= 2/3$ for `tanh`. In Appendix A, we'll analyze this a bit more using tools from *information theory* and see how it can lead to a preferred aspect ratio, L/n, that is different for specific choices of activation functions.

NLO Metric, Bare

Next, let's solve the NLO-metric recursion (5.110). Substituting in (5.121) for $\chi_{\parallel}(K)$ and (5.123) for $j(K)$, we get

$$G^{\{1\}(\ell+1)} = G^{\{1\}(\ell)}\left[1 + 2a_1\Delta K^{(\ell)} + \cdots\right] + V^{(\ell)}\left[a_1 + 3a_2\Delta K^{(\ell)} + \cdots\right]. \tag{5.132}$$

As should now be familiar, let's assume a large-ℓ scaling ansatz

$$G^{\{1\}(\ell)} = \#\left(\frac{1}{\ell}\right)^{p_1} + \cdots, \tag{5.133}$$

with p_1 as the associated critical exponent. Bootstrapping (5.132) by substituting in our previous solutions – (5.124) for $\Delta K^{(\ell)}$ and (5.127) for $V^{(\ell)}$ – we then insert our ansatz for $G^{\{1\}(\ell)}$ (5.133) and match terms to find

$$G^{\{1\}(\ell)} = -\left[\frac{1}{3(-a_1)}\right] + O\!\left(\frac{\log(\ell)}{\ell}\right). \tag{5.134}$$

This solution required us to set $p_1 = 0$ and gave a constant-in-ℓ leading contribution. Combining this with the kernel, we see that the finite-width-corrected two-point correlator

$$\mathbb{E}\left[z_i^{(\ell)} z_j^{(\ell)}\right] = \delta_{ij}\left[K^{(\ell)} + \frac{1}{n}G^{\{1\}(\ell)} + O\!\left(1/n^2\right)\right] \tag{5.135}$$

is given by

$$K^{(\ell)} + \frac{1}{n}G^{\{1\}(\ell)} = \left[\frac{1}{(-a_1)}\right]\left(\frac{1}{\ell} - \frac{1}{3n}\right) + \cdots. \tag{5.136}$$

This result is to be contrasted with the scale-invariant universality class, where the NLO metric vanished identically.

For the NLO metric, the appropriate dimensionless quantity to consider is the ratio between the correction term and the infinite-width term in the two-point correlator (5.135):

$$\frac{1}{n}\frac{G^{\{1\}(\ell)}}{K^{(\ell)}} \sim \frac{1}{n}\left(\frac{1}{\ell}\right)^{p_1 - p_0} + \cdots, \tag{5.137}$$

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

with the exponent $p_1 - p_0$ controlling the relative importance of this NLO correction. In this case we see that $p_1 - p_0 = -1$, meaning that the above ratio scales with the depth-to-width ratio ℓ/n. This again illustrates the perturbative cutoff of our effective theory, $\ell \lesssim n$. However, in this particular case such a scaling turns out to be an artifact of not properly tuning the initialization hyperparameters C_W at finite width, as we will see next.

NLO Metric, Renormalized

In §5.3.3, we learned how to find the critical initialization hyperparameters for the $K^\star = 0$ universality class, fixing the hyperparameters C_b and C_W using the infinite-width recursions for the kernel components. However, in §4.5 we explained that all of the observables computed in a large-n expansion receive an infinite series of subleading corrections in $1/n$. This suggests that we should have allowed further fine-tuning of the initialization hyperparameters at criticality by considering large-n expansions,

$$C_b^{(\ell)} = c_b^{(\ell)\{0\}} + \frac{c_b^{(\ell)\{1\}}}{n_{\ell-1}} + \frac{c_b^{(\ell)\{2\}}}{n_{\ell-1}^2} + \cdots \,, \tag{5.138}$$

$$C_W^{(\ell)} = c_W^{(\ell)\{0\}} + \frac{c_W^{(\ell)\{1\}}}{n_{\ell-1}} + \frac{c_W^{(\ell)\{2\}}}{n_{\ell-1}^2} + \cdots \,, \tag{5.139}$$

allowing us to adjust such hyperparameters order by order in $1/n$. Such an expansion could potentially give additional criticality conditions at each order in perturbation theory.

Considering the finite-width recursions (5.109) and (5.110), we see that such subleading tunings will not affect the leading-order result for observables that depend on the four-point vertex, since the leading contributions to such observables are already at $O(1/n)$. However, these tunings do affect the solution for the NLO metric, because the NLO metric is itself subleading.

Concretely, there is an additional contribution to the NLO-metric recursion (5.110) coming from inserting the expansions (5.138) and (5.139) into the kernel recursion (5.108). The terms proportional to $c_b^{(\ell)\{1\}}$ or $c_W^{(\ell)\{1\}}$ are now subleading and thus contribute to the NLO-metric recursion:

$$G^{\{1\}(\ell+1)} = \left[c_b^{(\ell)\{1\}} + c_W^{(\ell)\{1\}} g\!\left(K^{(\ell)}\right)\right] + \chi_\parallel^{(\ell)} G^{\{1\}(\ell)} + \frac{1}{8}\, j\!\left(K^{(\ell)}\right) \frac{V^{(\ell)}}{\left(K^{(\ell)}\right)^2}. \tag{5.140}$$

With this new "renormalized" perspective, we see that the analysis we did in the "bare" subsubsection before was just a particular choice of subleading corrections, $c_b^{(\ell)\{1\}} = c_W^{(\ell)\{1\}} = 0$. More generally, we really do have additional knobs to turn at this subleading order.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Substituting in (5.121) for $\chi_{\parallel}(K)$, (5.123) for $j(K)$, and (5.83) for $g(K)$, we find an algebraic equation

$$
\begin{aligned}
G^{\{1\}(\ell+1)} = c_b^{(\ell)\{1\}} + c_W^{(\ell)\{1\}}\sigma_1^2 \left[K^{(\ell)} + a_1\left(K^{(\ell)}\right)^2 + \cdots\right] \\
+ G^{\{1\}(\ell)}\left[1 + 2a_1K^{(\ell)} + \cdots\right] + V^{(\ell)}\left[a_1 + 3a_2K^{(\ell)} + \cdots\right].
\end{aligned}
\tag{5.141}
$$

Plugging in the solution for the kernel (5.124) and vertex (5.127) – making sure to include the subleading-in-ℓ terms in both – inserting our large-ℓ scaling ansatz for $G^{\{1\}(\ell)}$ (5.133) and matching terms, we find that the tunings

$$
c_b^{(\ell)\{1\}} = 0, \qquad c_W^{(\ell)\{1\}} = \frac{2}{3}c_W^{(\ell)\{0\}} = \frac{2}{3\sigma_1^2}
\tag{5.142}
$$

result in an asymptotically suppressed solution for the NLO metric:

$$
G^{\{1\}(\ell)} = \frac{2}{3}\left[\frac{3a_2 - a_1^2}{(-a_1)^3}\right]\frac{1}{\ell} + O\left(\frac{\log(\ell)}{\ell^2}\right),
\tag{5.143}
$$

with a critical exponent $p_1 = 1$. Specifically, the tuning of $c_b^{(\ell)\{1\}}$ was required to suppress a linear growing $\sim \ell$ contribution, while the tuning of $c_W^{(\ell)\{1\}}$ cancels the constant $O(1)$ piece we found before in (5.134).

- In a sense, we got lucky before in our bare analysis: redoing this analysis without a $c_b^{(\ell)\{1\}} = 0$ tuning, the dimensionless ratio (5.137) grows quadratically with depth and implies that the NLO metric dominates the kernel at $\ell \sim \sqrt{n}$. The fact that this subleading correction becomes parametrically large before reaching the ℓ/n perturbative cutoff of the effective theory really means that it's growing exponentially; $c_b^{(\ell)\{1\}} \neq 0$ eventually spoils criticality.

- In another sense, we got unlucky before: without the $c_W^{(\ell)\{1\}} = \frac{2}{3}c_W^{(\ell)\{0\}}$ tuning, the NLO metric is a leading ℓ/n correction. We see now that when properly handled, $p_1 - p_0 = 0$, and the dimensionless ratio (5.137) is $O(1)$ in depth at leading order. Such a correction is said to be *marginal* under the RG flow. This means that, while we'll always have to take into account the *relevant* four-point vertex corrections, we should be able to neglect NLO metric corrections as long as we respect the finite-width tunings (5.142).

Finally, the necessity of including such perturbative corrections to the critical initialization hyperparameters gives an alternate perspective on what can go wrong in practice when the network depth L approaches the network width n. Even for ensembles of such networks, the averaged quantities will require finer and finer tunings – e.g., (5.138) and (5.139) – in order for the effective theory describing the ensemble to reach criticality.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

For any reasonable value of n, such corrections will quickly become finer than the floating-point precision limit used to represent the hyperparameters. Thus, in practice it becomes essentially impossible to tune such large square networks to criticality.[27]

5.5 Finite-Angle Analysis for the Scale-Invariant Universality Class

In this section, we'll confront an important subtlety for activation functions in the scale-invariant universality class.

Recall that activation functions in this class take the form

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0, \\ a_- z\,, & z < 0 \end{cases} \tag{5.144}$$

and generally have a kink at the origin $z = 0$ (except for the *degenerate* member, the `linear` activation function, which has $a_+ = a_-$). In footnote 6 we first mentioned the existence of a subtlety after giving our δ expansions for the kernel (5.22)–(5.24), lightly questioning the validity of our expansions for nonsmooth $\sigma(z)$. In footnote 12, we then described the main consequence of this subtlety. In particular, we claimed that for nonlinear scale-invariant activation functions, the constant value – as a function of layer – of the perpendicular perturbation $\delta\delta K^{(\ell)}_{[2]}$ at criticality is an artifact of the perturbative δ expansion. To understand this claim properly, we'll need to work out the full nonperturbative kernel recursion for activation functions in this class. This in turn will let us see the aforementioned correction to the asymptotic large-ℓ behavior of the kernel component $\delta\delta K^{(\ell)}_{[2]}$.

For this analysis, it will be sufficient to focus on two inputs $x_{i;\pm}$ of the same norm. In our previous setup, we assumed that both inputs were nearby such that their difference $\delta x_i \equiv (x_{i;+} - x_{i;-})$ was perturbatively small, $\delta x_i \ll 1$; here, we will make no assumptions at all about their difference. Given the symmetries of the network evolution, the individual norms of the two preactivations corresponding to these inputs will also be equal:

$$K^{(\ell)}_d \equiv \mathbb{E}\left[\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(z^{(\ell)}_{i;+}\right)^2\right] = \mathbb{E}\left[\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(z^{(\ell)}_{i;-}\right)^2\right]. \tag{5.145}$$

Geometrically this means that our preactivations live together on an n_ℓ-dimensional sphere with radius $\sqrt{n_\ell K^{(\ell)}_d}$, and algebraically this means that the parallel component vanishes, $K^{(\ell)}_{[1]} = 0$, cf. (5.18). Going forward, we will call $K^{(\ell)}_d$ the **diagonal kernel**.[28]

[27] Note that this is an entirely different problem than the chaotic behavior at large depth that we described in §3.4 for deep linear networks. For the scale-invariant universality class, the NLO-metric correction vanishes, and therefore $c^{(\ell)\{1\}}_W = 0$.

[28] Perturbatively, the *diagonal kernel* $K^{(\ell)}_d$ is equal to the *midpoint kernel* $K^{(\ell)}_{00}$ – the kernel for the midpoint input $x_{i;0} \equiv (x_{i;+} + x_{i;-})/2$ – at leading order in the δ expansion, cf. (5.22)–(5.24).

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

The remaining dynamical variable is the polar angle between the preactivations. Therefore, we can decompose the two-input kernel matrix with the following parameterization:

$$K^{(\ell)}_{\alpha_1\alpha_2} = \begin{pmatrix} K^{(\ell)}_{++} & K^{(\ell)}_{+-} \\ K^{(\ell)}_{-+} & K^{(\ell)}_{--} \end{pmatrix} = K^{(\ell)}_{d} \begin{pmatrix} 1 & \cos\left(\psi^{(\ell)}\right) \\ \cos\left(\psi^{(\ell)}\right) & 1 \end{pmatrix}, \qquad \psi^{(\ell)} \in [0,\pi] . \tag{5.146}$$

The polar angle $\psi^{(\ell)}$ ranges from 0 – where the preactivations are coincident as $z_{i;+} = z_{i;-}$, making the kernel matrix degenerate – to π – where they're anti-correlated as $z_{i;+} = -z_{i;-}$. So far all we've done is fixed the norm of our two inputs to be equal and decomposed the kernel into a particular choice of coordinates; such a choice and parameterization can be applied to the analysis of any activation function. We'll now specialize to scale-invariant activation functions, for which class it's possible to derive a nonperturbative recursion for the polar angle.

RG Flow of the Polar Angle

The diagonal kernel follows the by-now familiar recursion for the single-input kernel (5.4)

$$K^{(\ell+1)}_{d} = C_b + C_W\, g\left(K^{(\ell)}_{d}\right) = C_b + A_2 C_W K^{(\ell)}_{d}, \tag{5.147}$$

where on the right-hand side we plugged in the explicit details for the scale-invariant universality class (5.63) and recalled $A_2 \equiv \left(a_+^2 + a_-^2\right)/2$. This part of the analysis carries over from §5.2. We recall here that we can readily solve the recursion for any choice of initialization hyperparameters, and in particular criticality is attained by setting $C_b = 0$ and $A_2 C_W = 1$, where the diagonal kernel stays exactly constant: $K^{(\ell)}_{d} = K^{(1)}_{d} \equiv K^{\star}_{d}$.

With the evolution of the magnitude determined, we now need to find a recursion for the polar angle $\psi^{(\ell)}$. Plugging our new decomposition (5.146) into the full kernel recursion (5.1), the off-diagonal component of the recursion becomes

$$K^{(\ell+1)}_{d} \cos\left(\psi^{(\ell+1)}\right) = C_b + C_W \left\langle \sigma(z_+)\,\sigma(z_-)\right\rangle_{K^{(\ell)}} . \tag{5.148}$$

In this parameterization, the Gaussian expectation reads

$$\left\langle \sigma(z_+)\,\sigma(z_-)\right\rangle_{K^{(\ell)}} \equiv \frac{\int dz_+ dz_-\ \sigma(z_+)\,\sigma(z_-)\, e^{-\frac{1}{2}\sum_{\alpha_1,\alpha_2=\pm} K^{\alpha_1\alpha_2}_{(\ell)} z_{\alpha_1} z_{\alpha_2}}}{2\pi K^{(\ell)}_{d} \sin\left(\psi^{(\ell)}\right)}, \tag{5.149}$$

where the denominator comes from evaluating the determinant $\sqrt{\left|2\pi K^{(\ell)}\right|}$. To make further progress, we need to evaluate this painful integral.

Nonperturbatively, these two kernels are very different. To see this most vividly, consider two antipodal inputs $x_{i;+} = -x_{i;-}$. Then, the midpoint input is the zero vector $x_{i;0} = 0$, and the midpoint kernel in the first layer is given by $K^{(1)}_{00} = C^{(1)}_{b}$. In contrast, the diagonal kernel is given by either of $K^{(1)}_{d} = C^{(1)}_{b} + (C^{(1)}_{W}/n_0)\sum_{i=1}^{n_0} x^2_{i;\pm}$.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Before working out the general case, let's focus on the `ReLU`. Setting $a_+ = 1$ and $a_- = 0$, we see that the argument of the Gaussian expectation is given by $\sigma(z_+)\,\sigma(z_-) = z_+ z_-$ when $z_+ > 0$ and $z_- > 0$ and vanishes otherwise. This means that the Gaussian expectation (5.149) is concentrated entirely in the first quadrant. In addition, noting that the integrand is invariant under parity, $(z_+, z_-) \to (-z_+, -z_-)$, we can niftily substitute the integral over the first quadrant for half the integral over the first and third quadrants. This lets us rewrite the above Gaussian expectation as

$$\langle \sigma(z_+)\,\sigma(z_-) \rangle_{K^{(\ell)}} = \frac{\frac{1}{2}\int dz_+ dz_- \big|_{z_+ z_- > 0}\; z_+ z_-\, e^{-\frac{1}{2}\sum_{\alpha_1,\alpha_2=\pm} K_{(\ell)}^{\alpha_1\alpha_2} z_{\alpha_1} z_{\alpha_2}}}{2\pi K_d^{(\ell)} \sin\left(\psi^{(\ell)}\right)} . \tag{5.150}$$

The above actually turns out to be the only nifty step of the derivation; everything else is just a Herculean sequence of coordinate changes.

There are three coordinate changes in said sequence:

$$\begin{aligned} z_\pm &= \frac{u \pm w}{\sqrt{2}} \qquad\qquad (5.151)\\ &= \sqrt{\frac{K_d^{(\ell)}\left[1+\cos\left(\psi^{(\ell)}\right)\right]}{2}}\; x \pm \sqrt{\frac{K_d^{(\ell)}\left[1-\cos\left(\psi^{(\ell)}\right)\right]}{2}}\; y \\ &= \sqrt{\frac{K_d^{(\ell)}\left[1+\cos\left(\psi^{(\ell)}\right)\right]}{2}}\; r\cos(\phi) \pm \sqrt{\frac{K_d^{(\ell)}\left[1-\cos\left(\psi^{(\ell)}\right)\right]}{2}}\; r\sin(\phi) . \end{aligned}$$

The first one diagonalizes the kernel so that the distribution factorizes $p(z_+, z_-) = p(u)p(w)$, the second one normalizes the coordinates with the kernel's eigenvalues, and the last one exchanges Cartesian coordinates for polar coordinates.[29] Accordingly, this lets us rewrite the sum in the exponential in (5.150) as

$$\sum_{\alpha_1,\alpha_2=\pm} K_{(\ell)}^{\alpha_1\alpha_2} z_{\alpha_1} z_{\alpha_2} = \frac{u^2}{K_d^{(\ell)}\left[1+\cos\left(\psi^{(\ell)}\right)\right]} + \frac{w^2}{K_d^{(\ell)}\left[1-\cos\left(\psi^{(\ell)}\right)\right]} = x^2 + y^2 = r^2 , \tag{5.152}$$

while the product in the integrand becomes

$$z_+ z_- = \frac{K_d^{(\ell)} r^2}{2}\left[\cos(2\phi) + \cos\left(\psi^{(\ell)}\right)\right] , \tag{5.153}$$

and the integral measure transforms as

$$dz_+ dz_- = K_d^{(\ell)} \sin\left(\psi^{(\ell)}\right) r\, dr\, d\phi . \tag{5.154}$$

[29] Unlike the perturbative calculations in (5.33) and (5.34), the diagonalization and normalization here are nonperturbatively exact. To reflect more on this, while we can always change coordinates as (5.151), we used the details of the `ReLU` in going from (5.149) to (5.150), establishing both the restricted domain of integration and the simplified form of the integrand, $\sigma(z_+)\,\sigma(z_-) \to z_+ z_-$, within that domain. For a general activation function, the resulting integral in the new coordinates (5.151) would still be difficult to evaluate, and we would have to resort to a perturbative expansion in $\psi^{(\ell)}$, ultimately analogous to the δ expansion, in order to make progress.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

Substituting (5.152)–(5.154) back into the Gaussian expectation (5.150), we get

$$\langle\sigma(z_+)\,\sigma(z_-)\rangle_{K^{(\ell)}} = \frac{K_d^{(\ell)}}{8\pi}\left[\int_0^\infty dr\; r^3 e^{-\frac{r^2}{2}}\right]\int_0^{2\pi} d\phi\Big|_{\cos(2\phi)+\cos\left(\psi^{(\ell)}\right)>0}\Big[\cos(2\phi)+\cos\left(\psi^{(\ell)}\right)\Big]\,. \tag{5.155}$$

The rest is now relatively straightforward. The radial integral can be evaluated by another change of the coordinate $s = r^2/2$:

$$\int_0^\infty dr\; r^3 e^{-\frac{r^2}{2}} = \int_0^\infty ds\; 2s\, e^{-s} = \Big[-2e^{-s} - 2s\,e^{-s}\Big]\Big|_0^\infty = 2. \tag{5.156}$$

For the angle integral, note that any function of $\cos(2\phi)$ gives the same contribution from the four intervals $\widetilde{\phi} \equiv 2\phi \in [0,\pi], [\pi,2\pi], [2\pi,3\pi], [3\pi,4\pi]$. Further, within that first interval the constraint $\cos\left(\widetilde{\phi}\right) > -\cos\left(\psi^{(\ell)}\right)$ can be simply expressed as $\widetilde{\phi} < \pi - \psi^{(\ell)}$. Together, this lets us write

$$\begin{aligned}
&\int_0^{2\pi} d\phi\Big|_{\cos(2\phi)+\cos\left(\psi^{(\ell)}\right)>0}\Big[\cos(2\phi)+\cos\left(\psi^{(\ell)}\right)\Big] \qquad (5.157)\\
&= 4\int_0^{\pi} \frac{d\widetilde{\phi}}{2}\Big|_{\cos(\widetilde{\phi})+\cos\left(\psi^{(\ell)}\right)>0}\Big[\cos(\widetilde{\phi})+\cos\left(\psi^{(\ell)}\right)\Big]\\
&= 2\int_0^{\pi-\psi^{(\ell)}} d\widetilde{\phi}\,\Big[\cos\left(\widetilde{\phi}\right)+\cos\left(\psi^{(\ell)}\right)\Big] = 2\sin\left(\psi^{(\ell)}\right) + 2\left(\pi-\psi^{(\ell)}\right)\cos\left(\psi^{(\ell)}\right).
\end{aligned}$$

Inserting (5.156) and (5.157) into (5.155), we finally arrive at an expression for the Gaussian expectation of `ReLU` activations:

$$\langle\sigma(z_+)\,\sigma(z_-)\rangle_{K^{(\ell)}} = \frac{K_d^{(\ell)}}{2\pi}\Big[\sin\left(\psi^{(\ell)}\right) + \left(\pi-\psi^{(\ell)}\right)\cos\left(\psi^{(\ell)}\right)\Big]\,. \tag{5.158}$$

Now, let's work out the painful integral (5.149) for an arbitrary scale-invariant activation function (5.144). In general, there are contributions from the first quadrant proportional to a_+^2 and similar contributions from the third quadrant proportional to a_-^2, in both cases with the constraint $z_+z_- > 0$ after our nifty trick. Then, there are also contributions from the second and fourth quadrants, both proportional to a_+a_- and with the constraint $z_+z_- < 0$. Following a very similar sequence of steps as we did before for the `ReLU`, we can evaluate the Gaussian expectation (5.149) as

$$\begin{aligned}
\langle\sigma(z_+)\,\sigma(z_-)\rangle_{K^{(\ell)}} =\; &\frac{K_d^{(\ell)}}{2\pi}(a_+^2+a_-^2)\int_0^{\pi-\psi^{(\ell)}} d\widetilde{\phi}\,\Big[\cos\left(\widetilde{\phi}\right)+\cos\left(\psi^{(\ell)}\right)\Big] \qquad (5.159)\\
&+\frac{K_d^{(\ell)}}{2\pi}(2a_+a_-)\int_{\pi-\psi^{(\ell)}}^{\pi} d\widetilde{\phi}\,\Big[\cos\left(\widetilde{\phi}\right)+\cos\left(\psi^{(\ell)}\right)\Big]\\
=\; &\frac{K_d^{(\ell)}}{2\pi}\,(a_+-a_-)^2\,\Big[\sin\left(\psi^{(\ell)}\right) - \psi^{(\ell)}\cos\left(\psi^{(\ell)}\right)\Big]\\
&+\left(\frac{a_+^2+a_-^2}{2}\right)K_d^{(\ell)}\cos\left(\psi^{(\ell)}\right).
\end{aligned}$$

The full nonperturbative recursion for the off-diagonal part of the kernel (5.148) thus evaluates to

$$
\begin{aligned}
&K_d^{(\ell+1)} \cos\Big(\psi^{(\ell+1)}\Big) \qquad (5.160)\\
&= C_b + C_W \left\{ \frac{K_d^{(\ell)}}{2\pi} (a_+ - a_-)^2 \Big[\sin\Big(\psi^{(\ell)}\Big) - \psi^{(\ell)} \cos\Big(\psi^{(\ell)}\Big)\Big] + \left(\frac{a_+^2 + a_-^2}{2}\right) K_d^{(\ell)} \cos\Big(\psi^{(\ell)}\Big) \right\}.
\end{aligned}
$$

One thing we notice here is that even though we evaluated the Gaussian expectation, we'll still have to deal with the fact that the recursion is highly nonlinear in $\psi^{(\ell+1)}$.

While you're here and we have your attention, let's record the result for one additional nonperturbative Gaussian expectation for the scale-invariant universality class: $\langle \sigma'(z_+)\,\sigma'(z_-)\rangle_{K^{(\ell)}}$. The integral here is much simpler to evaluate than the undifferentiated one above since in each quadrant the argument of the expectation, $\sigma'(z_+)\,\sigma'(z_-)$, is constant. Following otherwise the same set of steps as above, in this case we find

$$
\begin{aligned}
\langle \sigma'(z_+)\,\sigma'(z_-)\rangle_{K^{(\ell)}} &= \frac{(a_+^2 + a_-^2)}{4\pi} \left[\int_0^\infty dr\; r e^{-\frac{r^2}{2}}\right] \int_0^{2\pi} d\phi \Big|_{\cos(2\phi)+\cos\left(\psi^{(\ell)}\right)>0} \qquad (5.161)\\
&\quad + \frac{2a_+a_-}{4\pi} \left[\int_0^\infty dr\; r e^{-\frac{r^2}{2}}\right] \int_0^{2\pi} d\phi \Big|_{\cos(2\phi)+\cos\left(\psi^{(\ell)}\right)<0}\\
&= \left(\frac{a_+^2 + a_-^2}{2}\right) - \frac{\psi^{(\ell)}}{2\pi} (a_+ - a_-)^2 \,.
\end{aligned}
$$

We guess you guys aren't ready for that yet. But your future-selves are gonna love it.[30]

Criticality Analysis of the Polar Angle

Having evaluated the recursion, let's now tune to criticality and work out the correct large-ℓ asymptotic behavior of the polar angle $\psi^{(\ell)}$. Working at scale-invariant criticality, with $C_b = 0$ and $A_2 C_W = 1$, and where the diagonal kernel is constant as $K_d^{(\ell)} = K_d^\star$, the off-diagonal recursion (5.160) simplifies to a decoupled recursion for the polar angle,

$$
\cos\Big(\psi^{(\ell+1)}\Big) = \cos\Big(\psi^{(\ell)}\Big) + \rho \Big[\sin\Big(\psi^{(\ell)}\Big) - \psi^{(\ell)} \cos\Big(\psi^{(\ell)}\Big)\Big] \,. \qquad (5.162)
$$

Here, it was convenient to define a new constant,

$$
\rho \equiv \frac{1}{\pi} \frac{(a_+ - a_-)^2}{(a_+^2 + a_-^2)} \,, \qquad (5.163)
$$

that encapsulates all of the details of the specific scale-invariant activation function. Roughly, ρ is a dimensionless measure of the kinkiness of the activation function at the origin, equal to zero for the `linear` activation function and $1/\pi$ for the `ReLU`. We see right

[30] This result will turn out to be really useful in §10.3 when we investigate generalization error for the scale-invariant universality class at infinite width.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

away that the polar angle is exactly preserved for, and only for, $\rho = 0$. In particular, the preservation of the full two-input kernel matrix that we saw for the `linear` activation function in §3.2 doesn't extend to any other member of the universality class.

In order to determine the large-ℓ behavior of the polar angle $\psi^{(\ell)}$, we need a way to analyze the recursion (5.162). As we've been emphasizing, our main tool for analyzing such a nonlinear recursion is to find a fixed point and then linearize around it.[31] By inspection of the recursion, it's clear that $\psi^{\star} = 0$ is a fixed point. Thus, we should focus in on the small-angle regime: $\psi^{(\ell)} \ll 1$.

Taylor-expanding the trigonometric functions in the recursion (5.162) around a vanishing polar angle, the linearized recursion becomes

$$\psi^{(\ell+1)} = \psi^{(\ell)}\sqrt{1 - \frac{2\rho}{3}\psi^{(\ell)} + O(\psi^2)} = \psi^{(\ell)} - \frac{\rho}{3}\left(\psi^{(\ell)}\right)^2 + O\left(\psi^3\right). \tag{5.164}$$

To solve this recursion, we can use our scaling ansatz (5.93), which here reads

$$\psi^{(\ell)} = \left(\frac{1}{\ell}\right)^{p_\psi}\left[c_{0,0} + O\left(\frac{\log \ell}{\ell}\right)\right], \tag{5.165}$$

with the critical exponent p_ψ governing the decay of the polar angle. Plugging this ansatz into our recursion (5.164) and matching the terms on both sides of the equation, we find a solution:

$$\psi^{(\ell)} = \left(\frac{3}{\rho}\right)\frac{1}{\ell} + O\left(\frac{\log \ell}{\ell^2}\right). \tag{5.166}$$

From this we can read off the critical exponent, $p_\psi = 1$, which is universal excepting the degenerate `linear` limit of $\rho = 0$, for which we instead have $p_\psi = 0$.

In order to recast this result in the language of the rest of this chapter, let's project the two-input kernel (5.146) into the $\gamma^{[a]}_{\alpha\beta}$ representation using (5.20) and then insert (5.166):

$$K^{(\ell)}_{[0]} = K^{(\ell)}_{d}\left[\frac{1+\cos\left(\psi^{(\ell)}\right)}{2}\right] = K^{\star}_{d} + O\left(\frac{1}{\ell^2}\right), \tag{5.167}$$

$$K^{(\ell)}_{[2]} = K^{(\ell)}_{d}\left[\frac{1-\cos\left(\psi^{(\ell)}\right)}{2}\right] = K^{\star}_{d}\left(\frac{9}{4\rho^2}\right)\frac{1}{\ell^2} + O\left(\frac{\log \ell}{\ell^3}\right). \tag{5.168}$$

These solutions form the basis of what we claimed earlier in footnote 12. In particular, the perpendicular perturbation $K^{(\ell)}_{[2]}$ crosses over from being nearly constant for small

[31] Since we already nonperturbatively evaluated the Gaussian expectation (5.159) and fully took into account the lack of smoothness of the activation function – with the constant ρ (5.163) characterizing its kinkiness – at this point it's completely safe to employ a perturbative expansion.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press

depth $\ell \ll \ell_{\text{cross}}$ to power-law decaying $\sim 1/\ell^2$ for large depth $\ell \gg \ell_{\text{cross}}$.[32] This implies that the true critical exponent for scale-invariant perpendicular perturbations is $p_\perp = 2$.

Here, the crossover scale ℓ_{cross} is approximately given by

$$\ell_{\text{cross}} \sim \frac{3}{\rho\psi^{(\ell=1)}} \sim \frac{3}{2\rho}\sqrt{\frac{K_{[0]}^{(\ell=1)}}{K_{[2]}^{(\ell=1)}}}. \tag{5.169}$$

We get this by equating the small-depth constant answer, set by the first-layer condition, with the large-ℓ asymptotic answer given by (5.166); on the right-hand side of (5.169) we further wrote the polar angle $\psi^{(\ell=1)}$ in terms of the kernel components using (5.167) and (5.168). What we see is that the smaller this initial angle $\psi^{(\ell=1)}$ is – meaning that the closer the two inputs are to each other – the longer our original constant solution to the naive perpendicular recursion (5.65) is valid and the longer it takes for the power-law regime to kick in.

The discussion above explains why our δ expansion failed to see the crossover: in such an analysis, by construction, $K_{[2]}^{(\ell=1)}$ is infinitesimally small. This means that the crossover scale (5.169) is pushed to infinity, invisible to perturbation theory. Here is another way to see it. For small separation of two inputs, we can rewrite the angle as

$$\psi^{(\ell)} \approx 2\sqrt{\frac{\delta\delta K_{[2]}^{(\ell)}}{K_d^\star}} + \cdots, \tag{5.170}$$

and hence the angle recursion (5.164) can be recast – upon a squaring and a rearrangement of terms – as

$$\delta\delta K_{[2]}^{(\ell+1)} = \delta\delta K_{[2]}^{(\ell)} - \frac{4\rho}{3\sqrt{K_d^\star}}\left(\delta\delta K_{[2]}^{(\ell)}\right)^{\frac{3}{2}} + \cdots. \tag{5.171}$$

Unfortunately, it's impossible to generate such a noninteger power, 3/2, via a Taylor expansion. Given our ansatz for the perpendicular perturbation $K_{[2]}^{(\ell)}$ (5.24), this explains why the correction term was invisible before. (There is no such issue for smooth activation functions; our Taylor expansion and subsequent analysis can be completely trusted for the $K^\star = 0$ universality class.)

The overall lesson here is that we should be very careful whenever encountering singular Gaussian expectations. In the future when we need to consider multiple inputs for nonlinear scale-invariant activation functions, we'll make sure to recall the results here.

[32] For deep linear networks where $\rho = 0$, the solution (5.168) is degenerate and doesn't apply. However, from our discussion just before we know that for such networks the polar angle remains constant at any depth.

https://doi.org/10.1017/9781009023405.007 Published online by Cambridge University Press