Appendix B

Residual Learning

No Doc, not me, the other me! The one that's up on stage playing Johnny B. Goode*!*

Marty McFly [87].

In this final appendix, we'll analyze neural networks with residual connections, generally called **residual networks**. These networks were originally introduced in order to enable the training of deeper and deeper networks: traditionally deep networks suffer from the *exploding and vanishing gradient problem*, but even in networks where various tricks of the trade are used to ensure the propagation of forward and backward signals, overly deep networks are empirically found to have *higher training and test errors* than their shallower-network counterparts.

From the microscopic perspective, the increase of the generalization error is intuitive for a deeper model with more model parameters, but the increase in the training error is not: the additional parameters should naively enable *better* fits to the training data. At the very least, one might hope that the additional layers could – in principle – approximate the identity map and do *no worse*. Yet the empirical evidence mentioned above suggests that it's difficult for optimization algorithms to tune the hidden layers of a deep network to such a near-identity map. This is called **degradation** and in principle is a major limiting factor for developing larger scale deep-learning models.

From the macroscopic perspective of our effective theory, we can offer a *dual* explanation for this degradation problem. As a precursor to our explanation, first recall that, rather than using any heuristic approach to solve the exploding and vanishing gradient problem, in §9.4 we analytically solved its *exponential* manifestation by means of criticality and then solved its *polynomial* manifestation by means of the learning-rate equivalence principle. In §10.3 we further confirmed that the associated tunings of the initialization hyperparameters, $C_b^{(\ell)}$ and $C_W^{(\ell)}$, and of the training hyperparameters, $\lambda_b^{(\ell)}$ and $\lambda_W^{(\ell)}$, lead to the most robust generalization performance for MLPs.

Now, *even if* these hyperparameters are properly tuned, we would expect that overly deep networks will suffer horribly from instantiation-to-instantiation fluctuations, leading to the breakdown of criticality for any *particular* network. This problem was first

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

discussed in §3.4 for extremely deep (linear) networks, then more generally in §5.4, and finally in footnote 8 of Epilogue ε. Thus, combined with our discussion of the MLP's depth-to-width ratio $r \equiv L/n$ in §A.3, perhaps we can understand the training loss degradation problem in terms of the network leaving the regime of optimality and proceeding toward the regime of chaos.

If you'll please move the microscopic explanation back to the front of your mind, we can explain an ingenious solution to degradation by He *et al.* [88]. Rather than trying to find a better learning algorithm, we can instead modify the deep network architecture so that the hidden layers only have to learn a *residual function*: in place of a generic nonlinear $(\ell+1)$-th layer

$$z^{(\ell+1)} = \texttt{L}\left(z^{(\ell)}; \theta^{(\ell+1)}\right), \tag{B.1}$$

we design the layer as

$$z^{(\ell+1)} = \texttt{R}\left(z^{(\ell)}; \theta^{(\ell+1)}\right) + z^{(\ell)}, \tag{B.2}$$

such that the **residual block**, $\texttt{R}\left(z^{(\ell)}; \theta^{(\ell+1)}\right)$, is the residual of the function that we want our layer, (B.1), to implement. The basic structure of this generic residual layer is depicted in the left panel of Figure B.1 and will be further explained later on.

From a microscopic perspective, these residual connections make learning a near-identity map much easier. Indeed, it is far easier to set the residual block $\texttt{R}\left(z^{(\ell)}; \theta^{(\ell+1)}\right)$ to near-zero than it is to coax a generic nonlinear function $\texttt{L}\left(z^{(\ell)}; \theta^{(\ell+1)}\right)$ to approximate the identity. In particular, since standard building blocks of the neural network often have the property that they vanish when their parameters are set to zero – cf. (2.5) for the *MLP-layer block* – and since typical initialization distributions have zero mean – cf. (2.21) and (2.22) – residual networks make it fairly easy to find a solution with $\texttt{R}\left(z^{(\ell)}; \theta^{\star}\right) \approx 0$ such that the addition of the hidden layer doesn't necessarily degrade the performance of the network.

More generally, we hope that the preactivation will actually play two roles, one of *coarse-graining* the input signal according to representation group flow (§4.6) and the other of propagating an *undegraded* copy of the input signal. This is plausibly quite helpful as it allows us to train deeper models with more parameters, and it has indeed been empirically demonstrated that such deeper residual networks lead to significant performance gains on the test set.

One of the goals of this appendix is to provide a dual macroscopic explanation for why the residual connections let us train overly-deep networks. Our macroscopic explanation above for the origin of the degradation problem – combined with the empirical success of very deep residual networks – suggests that the inclusion of residual connections shifts the optimal aspect ratio $r^{\star}$ from its MLP value, (A.72), to higher values. This would extend the range of effectively-deep networks and thus explain why residual connections let us train deeper networks. To test this hypothesis, we will need to carry out our effective theory analysis for residual networks.

Despite the long prelude, this appendix will be relatively brief and only involve some simple calculations. These exercises will serve two purposes. First, given the

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

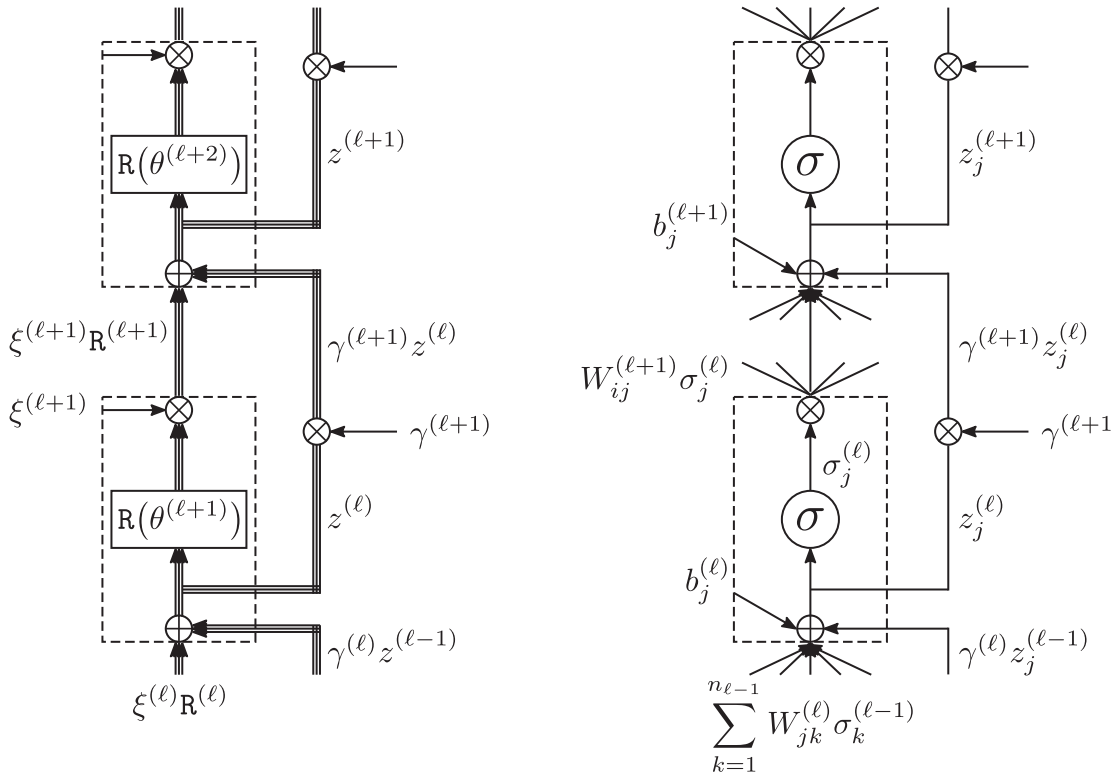

Figure B.1 **Left:** two residual blocks from adjacent layers for a very general *residual network*. This detailed structure depicts how each layer *(i)* adds a weighted block and the weighted preactivation to produce the next layer's preactivation, *(ii)* copies the preactivation to skip to the next layer, and *(iii)* generates the next layer's block. **Right:** two neurons from adjacent layers for a *residual MLP*. This detailed structure depicts how each layer *(i)* adds the bias, the weighted activation, and the weighted preactivation to produce the next layer's preactivation, *(ii)* copies the preactivation to skip to the next layer, *(iii)* generates the activation from the preactivation, and *(iv)* multiplies the activation by the next-layer weight.

overwhelming ubiquity of **residual connections** – sometimes called *skip connections* or *shortcuts* – in modern deep learning architectures, it is practically useful to explain how the critical initialization hyperparameters *shift* when residual connections are included. Second, this will showcase how our effective theory formalism can easily be applied to neural network architectures other than vanilla MLPs.

To those ends – and to the end of the book – in §B.1 we'll begin by briefly introducing the perhaps simplest residual network, a *multilayer perceptron with residual connections*. In §B.2, we'll study the infinite-width limit of this model, performing a criticality analysis in order to understand how the *residual hyperparameters* interplay with the critical initialization hyperparameters.

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

Then, in §B.3 we'll study the residual MLP at finite width. Using our auxiliary unsupervised learning objective from the last appendix, we'll see how the inclusion of residual connections can shift the *optimal aspect ratio* of a network $r^\star$ to large values. We will also learn that for networks with aspect ratios below a certain threshold, residual connections are *always* harmful according to this criterion. Altogether, this provides a new effective-theory macroscopic perspective on how residual connections solve the degradation problem described above, and further lets us understand the tradeoff of propagating signals through the residual block – in this case, a nonlinear MLP-layer block – against propagation through the identity block – skipping signals to deeper layers.

Finally, in §B.4 we'll give a hybrid theoretical-empirical recipe applying the analyses in the previous two sections to *general* residual networks with arbitrarily complicated residual blocks $\mathtt{R}\left(z^{(\ell)};\theta^{(\ell+1)}\right)$. We hope this may have broad application to the many deep learning architectures that implement residual connections.[1] After this, you will have finished all your residual learning from this book.

B.1 Residual Multilayer Perceptrons

A **multilayer perceptron with residual connections** can be defined by the forward equation

$$z_{i;\delta}^{(\ell+1)} = \xi^{(\ell+1)}\left[b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)}\sigma_{j;\delta}^{(\ell)}\right] + \gamma^{(\ell+1)} z_{i;\delta}^{(\ell)}. \tag{B.3}$$

Compared with our schematic description, (B.2), here we've picked just a standard MLP layer as our residual block, $\mathtt{R}\left(z^{(\ell)};\theta^{(\ell+1)}\right)$, and we've allowed for *scalar* **residual hyperparameters**, $\xi^{(\ell)}$ and $\gamma^{(\ell)}$, that control the relative magnitudes of the residual-block term vs. the identity term. Here we need to take

$$n_\ell = n_{\ell+1} \equiv n \tag{B.4}$$

so that we can add the weighted and biased activation back to the preactivation before the activation, and such residual connections are thus only included for the network's hidden layers.[2] A graph of two neurons in adjacent layers from a residual MLP is shown

[1] Residual networks were first described by He *et al.* [88]. Typically, when referring to a residual network as a *ResNet*, the base block is composed of convolutional layers, cf. (2.8), that are further augmented with a very popular heuristic for mitigating the exploding and vanishing gradient problem [89]. While the original domain of ResNets was computer vision tasks, they have now been applied to other domains; more broadly, *residual connections* are components in a wide variety of modern deep learning architectures, including importantly the transformer-based language models that have been revolutionizing natural language processing [14].

[2] More generally, if we wanted $n_{\ell+1} \neq n_\ell$, we could instead use an iteration equation such as

$$z_{i;\delta}^{(\ell+1)} = \sum_{j=1}^{n_{\ell+1}} \xi_{ij}^{(\ell+1)}\left[b_j^{(\ell+1)} + \sum_{k=1}^{n_\ell} W_{jk}^{(\ell+1)}\sigma_{k;\delta}^{(\ell)}\right] + \sum_{j=1}^{n_\ell} \gamma_{ij}^{(\ell+1)} z_{j;\delta}^{(\ell)}, \tag{B.5}$$

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

in the right panel of Figure B.1, and if you need to recall the overall global structure of an MLP, please refer back to Figure 2.1.

Although nice for the symmetry, one of the residual hyperparameters is redundant: you can show that adjusting $\xi^{(\ell)}$ has the same effect on the ensemble of residual MLPs as rescaling the initialization hyperparameters, $C_b^{(\ell)}$ and $C_W^{(\ell)}$, and the training hyperparameters, $\lambda_b^{(\ell)}$ and $\lambda_W^{(\ell)}$. As such, we'll henceforth set

$$\xi^{(\ell)} = 1 \tag{B.6}$$

without loss of generality, leaving us only one (potentially layer-dependent) residual hyperparameter: $\gamma^{(\ell)}$. Note importantly that in the limit $\gamma^{(\ell)} \to 0$, we recover a vanilla MLP without any residual connections. A customary choice for this hyperparameter is $\gamma^{(\ell)} = 1$, cf. (B.2), but let us now show that this is suboptimal because it *breaks* criticality.[3]

B.2 Residual Infinite Width: Criticality Analysis

To understand how to set the residual hyperparameter $\gamma^{(\ell)}$ and preserve criticality for the residual MLP, let's work out a recursion for the two-point correlator:

$$\mathbb{E}\left[z_{i_1;\delta_1}^{(\ell)} z_{i_2;\delta_2}^{(\ell)}\right] = \delta_{i_1 i_2} G_{\delta_1\delta_2}^{(\ell)}. \tag{B.7}$$

Using the forward equation (B.3), we find

$$G_{\delta_1\delta_2}^{(\ell+1)} = C_b + C_W \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\sigma_{j;\delta_1}^{(\ell)} \sigma_{j;\delta_2}^{(\ell)}\right] + \gamma^2 G_{\delta_1\delta_2}^{(\ell)}. \tag{B.8}$$

Here, mirroring all our previous discussions of criticality that are now too many to enumerate with references, throughout this section we'll set the bias variance, $C_b^{(\ell)}$, the rescaled weight variance, $C_W^{(\ell)}$, and the residual hyperparameter, $\gamma^{(\ell)}$, to be uniform across layers:

$$C_b^{(\ell)} = C_b\,, \qquad C_W^{(\ell)} = C_W\,, \qquad \gamma^{(\ell)} = \gamma. \tag{B.9}$$

Compared to the metric recursion for a vanilla MLP (4.72),

$$G_{\delta_1\delta_2}^{(\ell+1)} = C_b + C_W \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\sigma_{j;\delta_1}^{(\ell)} \sigma_{j;\delta_2}^{(\ell)}\right], \tag{B.10}$$

where the residual hyperparameters are now matrices: $\xi_{ij}^{(\ell+1)}$ is an $n_{\ell+1}$-by-$n_{\ell+1}$ matrix and $\gamma_{ij}^{(\ell+1)}$ is an $n_{\ell+1}$-by-n_ℓ matrix, and in general both matrices could vary from layer to layer.

[3] In particular, this choice is problematic without any additional heuristics, e.g., [89], for otherwise mitigating the exploding and vanishing gradient problem. The following analysis thus offers an explanation for why deep $\gamma^{(\ell)} = 1$ residual MLPs without such heuristics will always perform poorly.

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

the $(\ell+1)$-th-layer metric for the residual MLP is shifted by the (rescaled) identity term from the forward equation, leading to an additional contribution of the ℓ-th-layer metric as the (rescaled) final term of (B.8).

In particular, the infinite-width limit of the residual MLP leads to the following recursion for its kernel:

$$K^{(\ell+1)}_{\delta_1\delta_2} = C_b + C_W \left\langle \sigma_{\delta_1}\sigma_{\delta_2} \right\rangle_{K^{(\ell)}} + \gamma^2 K^{(\ell)}_{\delta_1\delta_2}. \tag{B.11}$$

Then, writing the kernel in our $\gamma^{[a]}$ basis (5.15), using our δ expansion, (5.22)–(5.24), we can follow the two-input perturbative analysis of §5.1 to derive component recursions

$$K^{(\ell+1)}_{00} = C_b + C_W\, g\Big(K^{(\ell)}_{00}\Big) + \gamma^2 K^{(\ell)}_{00}\,, \tag{B.12}$$

$$\delta K^{(\ell+1)}_{[1]} = \chi_{\parallel}\Big(K^{(\ell)}_{00}\Big)\, \delta K^{(\ell)}_{[1]}\,, \tag{B.13}$$

$$\delta\delta K^{(\ell+1)}_{[2]} = \chi_{\perp}\Big(K^{(\ell)}_{00}\Big)\, \delta\delta K^{(\ell)}_{[2]} + h\Big(K^{(\ell)}_{00}\Big)\, \Big(\delta K^{(\ell)}_{[1]}\Big)^2\,, \tag{B.14}$$

which are modified from their vanilla expressions (5.46)–(5.48).[4] Here, the *parallel susceptibility*,

$$\chi_{\parallel}(K) \equiv \gamma^2 + \frac{C_W}{2K^2}\Big\langle \sigma(z)\,\sigma(z)\,\Big(z^2 - K\Big)\Big\rangle_K\,, \tag{B.15}$$

and the *perpendicular susceptibility*,

$$\chi_{\perp}(K) \equiv \gamma^2 + C_W \left\langle \sigma'(z)\,\sigma'(z)\right\rangle_K\,, \tag{B.16}$$

are each shifted from their previous values, (5.50) and (5.51), by a constant γ^2 term, while the helper functions, $g(K)$ and $h(K)$, (5.49) and (5.52), are the same as before:

$$g(K) \equiv \left\langle \sigma(z)\,\sigma(z)\right\rangle_K\,, \tag{B.17}$$

$$h(K) \equiv \frac{1}{2}\frac{d}{dK}\chi_{\perp}(K). \tag{B.18}$$

As should already be clear from this simple analysis, inclusion of residual connections ($\gamma \neq 0$) will shift the critical initialization hyperparameters. Fixing the criticality conditions

$$\chi_{\parallel}(K^\star) = 1\,, \qquad \chi_{\perp}(K^\star) = 1, \tag{B.19}$$

we can find the new critical initializations for our universality classes: specifically, for the scale-invariant universality class, the critical initialization hyperparameters, (5.67), are modified as

$$C_b = 0\,, \qquad C_W = C_W(\gamma) \equiv \frac{1}{A_2}\left(1 - \gamma^2\right), \tag{B.20}$$

[4] Note that, as per this first recursion (B.12), a perturbation to the single-input kernel, $K^{(\ell)}_{00} = K^\star_{00} + \Delta K^{(\ell)}_{00}$, is governed by the *shifted* parallel susceptibility (B.15), cf. (5.9).

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

where $A_2 \equiv (a_+^2 + a_-^2)/2$, cf. (5.59); analogously, for the the $K^\star = 0$ universality class, the critical initialization hyperparameters (5.90) are modified as

$$C_b = 0\,, \qquad C_W = C_W(\gamma) \equiv \frac{1}{\sigma_1^2}\left(1 - \gamma^2\right), \tag{B.21}$$

where $\sigma_1 \equiv \sigma'(0)$.

In §2.3, we discussed that the *zero initialization*, $b_i^{(\ell)} = W_{ij}^{(\ell)} = 0$, fails to break the permutation symmetry among the n neurons in a hidden layer. In conjunction with this reasoning, we now see that the criticality conditions for either class, (B.20) or (B.21), are unsatisfiable for the customary choice $\gamma = 1$ of the residual hyperparameter.[5] More broadly, for each universality class, there is a *one-parameter family* of critical rescaled weight variances: for each $0 < \gamma^2 < 1$, there is an associated critical value of $C_W = C_W(\gamma)$. Thus, in order to directly mitigate the exploding and vanishing gradient problem, for a particular choice of $0 < \gamma^2 < 1$ and activation function, we must pick C_W according to the appropriate $C_W(\gamma)$.

B.3 Residual Finite Width: Optimal Aspect Ratio

From our infinite-width analysis, we just saw that residual networks have a one-parameter family of critical solutions: $C_W(\gamma)$. As per the kernel recursion (B.11), these solutions trade off the degree to which the identity branch vs. the MLP-layer branch contributes to the next layer's preactivations. In the strict infinite-width limit, criticality ensures that the kernel is preserved for any choice $C_W(\gamma)$ in the range $0 < \gamma^2 < 1$, and all fluctuations are suppressed; thus, we're in principle completely indifferent between any of these critical tunings.

However, as should now be a familiar point of discussion, networks in the infinite-width limit effectively have a depth-to-width ratio $r \equiv L/n \to 0$. This means that infinite-width analysis cannot really get at the question of **degradation**: why do extremely deep networks have *higher training errors* than otherwise equivalent shallower networks. To compare wide networks of different depths, we need to consider finite-width networks with different aspect ratios $r > 0$.[6]

[5] It is a different story when the residual connection skips more than one layer, but we'll have to save that tale for another time. In that narrative, you will still need to follow something like the recipe of §B.4 in order to tune to criticality.

[6] In particular, at finite width the *representation group flow* through the MLP-layer blocks leads to two competing finite-width effects: *(i)* the relevant and thus *growing* dNTK and ddNTKs lead to nontrivial representation learning during training, and *(ii)* the relevant and *growing* four-point vertex, NTK variance, and NTK–preactivation cross correlation lead to fluctuations in the ensemble from instantiation to instantiation. As the residual connections are supposed to mitigate this second harmful effect by xeroxing the undegraded input via the identity branch, we naturally expect that they will have a meaningful physical effect on this competition. In other words, we expect that residual networks of different residual hyperparameters γ and different aspect ratios r will lead to very different test-set generalization.

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

Our main tool of analysis in this section will be a computation of the *mutual information at initialization* between nonoverlapping representations in deep layers of a residual MLP. As per our auxiliary unsupervised learning criterion from the last appendix, cf. footnote 28 of §A.3, this mutual information gives a natural way to estimate the optimal aspect ratio $r^\star$ of a finite-width network. Given that residual networks without an exploding and vanishing gradient problem solve the degradation problem, a natural theoretical prediction for our residual MLPs is that the inclusion of residual connections, $\gamma > 0$, and then tuning to criticality as $C_W(\gamma)$ will together *shift their optimal aspect ratios to larger values*.

To evaluate the optimal aspect ratio via our formula (A.73), we just need to compute the coefficient ν of normalized, output-layer, single-input four-point vertex for the residual MLP:

$$\frac{V^{(L)}}{n\left(G^{(L)}\right)^2} \equiv \nu r. \tag{B.22}$$

Note that we've already derived the recursion for the leading-order single-input metric, i.e., the kernel $K^{(L)} \equiv G^{\{0\}(L)}$, in (B.12). To compute the four-point vertex, first recall that the multi-input four-point vertex defines the four-point connected correlator (4.77) as

$$\begin{aligned}&\mathbb{E}\left[z^{(\ell)}_{i_1;\delta_1} z^{(\ell)}_{i_2;\delta_2} z^{(\ell)}_{i_3;\delta_3} z^{(\ell)}_{i_4;\delta_4}\right]\Big|_{\text{connected}} \\ &\quad = \frac{1}{n}\left[\delta_{i_1 i_2}\delta_{i_3 i_4} V^{(\ell)}_{(\delta_1\delta_2)(\delta_3\delta_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4} V^{(\ell)}_{(\delta_1\delta_3)(\delta_2\delta_4)} + \delta_{i_1 i_4}\delta_{i_2 i_3} V^{(\ell)}_{(\delta_1\delta_4)(\delta_2\delta_3)}\right].\end{aligned} \tag{B.23}$$

Using the forward equation (B.3) and following our analysis from §4.3, we see that the recursion for the residual-MLP's four-point vertex is shifted by two additional terms proportional to γ^2 and γ^4 as compared to the leading-order recursion for the vanilla MLP (4.90),

$$\begin{aligned}&V^{(\ell+1)}_{(\delta_1\delta_2)(\delta_3\delta_4)} \\ &\quad = C_W^2\left[\left\langle \sigma_{\delta_1}\sigma_{\delta_2}\sigma_{\delta_3}\sigma_{\delta_4}\right\rangle_{K^{(\ell)}} - \left\langle \sigma_{\delta_1}\sigma_{\delta_2}\right\rangle_{K^{(\ell)}}\left\langle \sigma_{\delta_3}\sigma_{\delta_4}\right\rangle_{K^{(\ell)}}\right] \\ &\qquad + \frac{C_W^2}{4}\sum_{\delta_5,\ldots,\delta_8\in\mathcal{D}} V^{(\delta_5\delta_6)(\delta_7\delta_8)}_{(\ell)} \left\langle \sigma_{\delta_1}\sigma_{\delta_2}\left(z_{\delta_5}z_{\delta_6} - K_{\delta_5\delta_6}\right)\right\rangle_{K^{(\ell)}}\left\langle \sigma_{\delta_3}\sigma_{\delta_4}\left(z_{\delta_7}z_{\delta_8} - K_{\delta_7\delta_8}\right)\right\rangle_{K^{(\ell)}} \\ &\qquad + C_W\gamma^2\Big[\left\langle \sigma_{\delta_1}\sigma_{\delta_2}\left(z_{\delta_3}z_{\delta_4} - K_{\delta_3\delta_4}\right)\right\rangle_{K^{(\ell)}} + \left\langle \sigma_{\delta_3}\sigma_{\delta_4}\left(z_{\delta_1}z_{\delta_2} - K_{\delta_1\delta_2}\right)\right\rangle_{K^{(\ell)}} \\ &\qquad\qquad + \frac{1}{2}\sum_{\delta_5,\ldots,\delta_8\in\mathcal{D}} V^{(\delta_5\delta_6)(\delta_7\delta_8)}_{(\ell)} \left\langle \sigma_{\delta_1}\sigma_{\delta_2}\left(z_{\delta_5}z_{\delta_6} - K_{\delta_5\delta_6}\right)\right\rangle_{K^{(\ell)}} K^{(\ell)}_{\delta_7\delta_3} K^{(\ell)}_{\delta_8\delta_4} \\ &\qquad\qquad + \frac{1}{2}\sum_{\delta_5,\ldots,\delta_8\in\mathcal{D}} V^{(\delta_5\delta_6)(\delta_7\delta_8)}_{(\ell)} \left\langle \sigma_{\delta_3}\sigma_{\delta_4}\left(z_{\delta_5}z_{\delta_6} - K_{\delta_5\delta_6}\right)\right\rangle_{K^{(\ell)}} K^{(\ell)}_{\delta_7\delta_1} K^{(\ell)}_{\delta_8\delta_2}\Big]\end{aligned} \tag{B.24}$$

$$\qquad + \gamma^4 V^{(\ell)}_{(\delta_1\delta_2)(\delta_3\delta_4)}, \tag{B.25}$$

where in parsing this expression, please remember that the raised indices of the four-point vertex are our shorthand for contraction with the ℓ-th-layer inverse kernel, cf. (4.83).

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

Here, in working out the middle terms proportional to γ^2, you might find the *intra*layer formula (4.64) to be useful.

Specializing now to a single input, we get

$$V^{(\ell+1)} = C_W^2 \left[\left\langle \sigma^4(z) \right\rangle_{K^{(\ell)}} - \left\langle \sigma^2(z) \right\rangle_{K^{(\ell)}}^2 \right] + \left(\chi_{\parallel}^{(\ell)}\right)^2 V^{(\ell)} + 4\gamma^2 \left(\chi_{\parallel}^{(\ell)} - \gamma^2\right) \left(K^{(\ell)}\right)^2 , \tag{B.26}$$

where for convenience we have abbreviated $\chi_{\parallel}^{(\ell)} \equiv \chi_{\parallel}\left(K^{(\ell)}\right)$ as we often do. Here, this parallel susceptibility is the shifted one appropriate for residual MLPs as defined in (B.15) with the γ^2 term; otherwise, we notice the addition of the γ-dependent final term compared with the single-input recursion for vanilla MLPs (5.109). At criticality with $\chi_{\parallel}^{(\ell)} = 1$, this final term will make a nontrivial difference to the deep asymptotic behavior of the four-point vertex. Now, let's solve this recursion at criticality for our two universality classes, with the usual initial condition $V^{(1)} = 0$ (cf. the title of §4.1).

Scale-Invariant Universality Class

For the scale-invariant universality class, we tune the rescaled weight variance to criticality with (B.20) so that $\chi_{\parallel}^{(\ell)} = 1$. Then, the single-input kernel is exactly fixed, $K^{(\ell)} = K^\star$, and we also need to remember (5.113),

$$\left[\left\langle \sigma^4 \right\rangle_{K^\star} - \left\langle \sigma^2 \right\rangle_{K^\star}^2 \right] = (3A_4 - A_2^2) \left(K^\star\right)^2 , \tag{B.27}$$

with $A_2 \equiv (a_+^2 + a_-^2)/2$ and $A_4 \equiv (a_+^4 + a_-^4)/2$. Substituting all of this into (B.26), the recursion becomes

$$V^{(\ell+1)} = V^{(\ell)} + (1-\gamma^2) \left[(1-\gamma^2) \left(\frac{3A_4}{A_2^2} - 1 \right) + 4\gamma^2 \right] \left(K^\star\right)^2 , \tag{B.28}$$

which gives a simple additive solution of the form (B.22), with

$$\nu = \nu(\gamma) \equiv (1-\gamma^2) \left[(1-\gamma^2) \left(\frac{3A_4}{A_2^2} - 1 \right) + 4\gamma^2 \right] . \tag{B.29}$$

As a quick check, we see that ν reduces to our previous result for vanilla MLPs, (A.70), as $\gamma \to 0$. More importantly, $\nu(\gamma)$ is strictly positive in the allowed range $0 < \gamma^2 < 1$ and monotonically decreases to $\nu(1) = 0$ with increasing γ.[7]

$K^\star = 0$ Universality Class

For the $K^\star = 0$ universality class, we tune the rescaled weight variance to criticality with (B.21). In this case, the single-input asymptotic behavior of the kernel is modified.

[7] To see this, you'll need to use the fact that $A_4 \geq A_2^2$ for all scale-invariant activation functions.

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

Evaluating the residual-MLP kernel recursion (B.12) at criticality and recalling our expansion (5.83),

$$g(K) = \sigma_1^2 \left[K + a_1 K^2 + O\Big(K^3\Big) \right], \tag{B.30}$$

the single-input kernel recursion becomes

$$K^{(\ell+1)} = K^{(\ell)} + a_1(1-\gamma^2)\left(K^{(\ell)}\right)^2 + \cdots, \tag{B.31}$$

which has deep asymptotic behavior of

$$K^{(\ell)} = \left[\frac{1}{(-a_1)(1-\gamma^2)} \right] \frac{1}{\ell} + \cdots. \tag{B.32}$$

To evaluate the four-point vertex recursion, we need

$$\chi_{\parallel}(K) = 1 + (1-\gamma^2)\left[2a_1 K + O\Big(K^2\Big) \right], \tag{B.33}$$

$$\left[\Big\langle \sigma^4 \Big\rangle_{K^\star} - \Big\langle \sigma^2 \Big\rangle_{K^\star}^2 \right] = \sigma_1^4 \left[2K^2 + O\Big(K^3\Big) \right], \tag{B.34}$$

where the first expression is shifted from (5.121), as per (B.15) and (B.21), but the second expression is the same as before: (5.122). Plugging in these expressions and matching terms, we find

$$\nu = \nu(\gamma) \equiv \frac{2}{3}(1-\gamma^4). \tag{B.35}$$

This also reduces to our previous result for vanilla $K^\star = 0$ MLPs, (A.69), as $\gamma \to 0$, and it is also strictly positive in the allowed range $0 < \gamma^2 < 1$, monotonically decreasing to $\nu(1) = 0$ with increasing γ. Thus, $\nu(\gamma)$ for the $K^\star = 0$ universality class is qualitatively comparable to the scale-invariant universality class.

Physics of the Optimal Aspect Ratio

As we saw at the end of §A.3, the maximization of the $1/n^3$ mutual information (A.67) according to our unsupervised learning criterion led to a natural estimate of the optimal aspect ratio of a network (A.73). For residual MLPs, this gives

$$r^\star(\gamma) \equiv \left(\frac{4}{20 + 3n_L} \right) \frac{1}{\nu(\gamma)}, \tag{B.36}$$

with $\nu(\gamma)$ given by (B.29) for the scale-invariant universality class and (B.35) for the $K^\star = 0$ universality class. Thus, we see that the monotonic decrease of $\nu(\gamma)$ for either class leads to a monotonic increase of $r^\star(\gamma)$ as γ increases: beginning from vanilla MLPs at $\gamma = 0$, residual MLPs will prefer larger and larger aspect ratios. Thus, we have

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

realized our theoretical prediction at the beginning of the section, finding support in our effective theory formalism for the hypothesis that residual connections can solve the degradation problem.

Let us offer a further interpretation of this mechanism. In [90], it was suggested that the benefit of residual connections in extremely deep networks is that they let the global network architecture behave as an *ensemble* of many shallower networks: in this interpretation, each *particular* "shallow network" is given by following the path of a signal from the input to the output of the single residual network. What was discovered is that gradients are dominated by paths skipping over most of the network and only passing through a small fraction of the hidden-layer residual blocks. This actually gives an interesting way to look at our result (B.36): if our maximization of mutual information (A.67) is weighing the helpful effect of depth as an inductive bias for neural association (§6.4.1) against the harmful effect of depth due to growing fluctuations (§5.4), then such ensembling would be a natural way to suppress the fluctuations while preserving the neural association. Thus, using residual connections to extend the depth of *trainable* networks should, at least to a point, also lead to better test set performance.

Note finally that we can actually interpret (B.36) in another way: rather than estimating the optimal aspect ratio $r^\star(\gamma)$ for a fixed residual hyperparameter γ, *instead* we can think of it as estimating the optimal residual hyperparameter $\gamma^\star(r)$ for a fixed aspect aspect ratio r. Taking this latter perspective, we learn something interesting. On the one hand, for a very shallow network with $r \ll 1$, our criterion (B.36) is unsatisfiable since $\nu(\gamma)$ monotonically decreases from its maximal value at $\gamma = 0$. Such networks are shallower than optimal according to our original criterion (A.73) for vanilla MLPs, $r < r^\star(\gamma = 0)$, and their mutual information will be greatest *without* the residual connections $\gamma^\star(r) = 0$. On the other hand, for very deep networks with $r > r^\star(\gamma = 0)$, the optimal residual hyperparameter $\gamma^\star(r)$ monotonically asymptotes to 1 with growing r.[8] Altogether, this suggests that we should only turn on the residual connections for networks with aspect ratios r that are greater than the threshold $r^\star(\gamma = 0)$ set by the optimal aspect ratio of vanilla MLPs.

Incidentally, if you are worried about the validity of perturbation theory for large r, note that the real physical expansion parameter is given by the combination

$$\frac{V^{(L)}}{n\left(G^{(L)}\right)^2} = \nu\Big(\gamma = \gamma^\star(r)\Big) \times r, \tag{B.37}$$

which stays finite even as the ratio r asymptotes to infinity. In this sense, we can use $\gamma^\star(r)$ to arbitrarily extend the regime of effectively-deep networks that can be described by our effective theory.

[8] When setting $\gamma^\star$ it's important to always remember to also adjust the rescaled weight variance $C_W(\gamma^\star)$ according to (B.20) or (B.21) in order to maintain criticality. In the limit of $r \gg r^\star(\gamma = 0)$, the optimal residual hyperparameter asymptotes to one as $1 - [\gamma^\star(r)]^2 \sim 1/r$, and the critical initialization gets smaller as $C_W\Big(\gamma = \gamma^\star(r)\Big) \sim 1/r$.

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

B.4 Residual Building Blocks

In this final section, we will explain a hybrid theoretical-empirical method for tuning a very general residual network to criticality. This method may be implemented practically in order to tune the hyperparameters of residual networks beyond the multilayer perceptron architecture. We hope this discussion provides a *blueprint* for how a few simple measurements on small networks can then be scaled up to efficiently design much larger models according to our effective theory approach.

A *general* residual network can be defined by replacing a simple MLP-layer block (B.3) by a generic nonlinear **residual block**:

$$b_i^{(\ell+1)} + \sum_{j=1}^{n} W_{ij}^{(\ell+1)} \sigma_{j;\delta}^{(\ell)} \;\rightarrow\; \mathtt{R}_i\left(z_\delta^{(\ell)}; \theta^{(\ell+1)}\right) \equiv \mathtt{R}_{i;\delta}^{(\ell+1)}. \tag{B.38}$$

Here, the ℓ-th-layer residual block $\mathtt{R}_{i;\delta}^{(\ell)}$ is shaped by model parameters $\theta^{(\ell)}$, and we should pick $\mathtt{R}_{i;\delta}^{(\ell)}$ to be a square matrix in its neural indices, $n_\ell = n_{\ell+1} \equiv n$, so that we can add the output of the residual block back to the preactivation. This leads to the following forward equation:

$$z_{i;\delta}^{(\ell+1)} = \xi^{(\ell+1)} \mathtt{R}_i\left(z_\delta^{(\ell)}; \theta^{(\ell+1)}\right) + \gamma^{(\ell+1)} z_{i;\delta}^{(\ell)}, \tag{B.39}$$

where we've restored the second residual hyperparameter, $\xi^{(\ell)}$, in order to let us scale the overall magnitude of the residual-block term. This iteration equation (B.39) is sufficiently generic to schematically capture many popular deep learning architectures, including the computer vision workhorse architecture, the *residual convolutional network* or **ResNet**, and the multi-headed self-attention-seeking language-modeling **transformer**. A graph of two residual blocks in adjacent layers from a general residual network described by (B.39) is shown in the left panel of Figure B.1.

In practice, certain heuristics and ad hoc methods are used in these general architectures to try and mitigate the exploding and vanishing gradient problem. However, as we know from §9.4, *criticality* is a much more motivated solution. In §B.2, we saw that for residual MLPs, the residual hyperparameters can be used to control criticality for the network. Now, let's see how we can implement a form of criticality for our general residual network (B.39).

Broadly speaking, we now need to solve two problems of different natures and difficulties:

- First, we need to ensure that signals can easily propagate through the residual block, especially for blocks that are *deep* in some sense; for instance, if a block individually consists of $L_{\mathtt{R}}$ MLP layers, i.e., *an MLP-$L_{\mathtt{R}}$-layer block*, it will then have an internal version of the exploding and vanishing gradient problem. Theorists should make every effort to critically analyze such blocks of practical interest, but if we have to treat the residual block $\mathtt{R}_{i;\delta}^{(\ell)}$ as a black box for one reason or another,

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

then this will require some *engineering*: you need to measure the two-point block–block correlator at the output of a block,

$$\mathbb{E}\left[\mathtt{R}_{i_1;\delta_1}^{(\ell+1)}\mathtt{R}_{i_2;\delta_2}^{(\ell+1)}\right], \tag{B.40}$$

and then compare it with the two-point correlator of the input to the block,

$$\mathbb{E}\left[z_{i_1;\delta_1}^{(\ell)}z_{i_2;\delta_2}^{(\ell)}\right]. \tag{B.41}$$

To be brief and concrete here, we'll focus on their diagonal components with $i_1 = i_2 \equiv i$ and $\delta_1 = \delta_2 \equiv \delta$. In particular, let us take an average over neural indices $i = 1, \ldots, n$ as well as over the sample indices $\delta \in \mathcal{D}$, so that we can compare two scalar quantities,

$$\overline{G}_{\mathtt{RR}}^{(\ell+1)} \equiv \frac{1}{|\mathcal{D}|n}\sum_{i=1}^{n}\sum_{\delta\in\mathcal{D}}\mathbb{E}\left[\mathtt{R}_{i;\delta}^{(\ell+1)}\mathtt{R}_{i;\delta}^{(\ell+1)}\right], \qquad \overline{G}_{zz}^{(\ell)} \equiv \frac{1}{|\mathcal{D}|n}\sum_{i=1}^{n}\sum_{\delta\in\mathcal{D}}\mathbb{E}\left[z_{i;\delta}^{(\ell)}z_{i;\delta}^{(\ell)}\right], \tag{B.42}$$

rather than two matrices, (B.40) and (B.41).[9] After setting up these measurements, we need to adjust the *initialization hyperparameters* for the block such that these quantities have similar magnitudes. For instance, if internally the block is parameterized by many iterative layers – like our MLP-$L_{\mathtt{R}}$-layer block – then we need to make sure that the difference in magnitudes is no worse than a *polynomial* in this depth hyperparameter $L_{\mathtt{R}}$; importantly, we want to ensure that there's no exponential growth or decay in $\overline{G}_{\mathtt{RR}}^{(\ell+1)}/\overline{G}_{zz}^{(\ell)}$.[10] Note that while you're setting up measurements, it will also be helpful to measure the cross correlator

$$\overline{G}_{\mathtt{R}z}^{(\ell+0.5)} \equiv \frac{1}{|\mathcal{D}|n}\sum_{i=1}^{n}\sum_{\delta\in\mathcal{D}}\mathbb{E}\left[\mathtt{R}_{i;\delta}^{(\ell+1)}z_{i;\delta}^{(\ell)}\right], \tag{B.43}$$

which will be needed in the next bullet point.

- Now, using the forward equation (B.39), we can write a recursion for the diagonal component of the two-point correlator as

$$\overline{G}_{zz}^{(\ell+1)} = \gamma^2\,\overline{G}_{zz}^{(\ell)} + 2\gamma\xi\,\overline{G}_{\mathtt{R}z}^{(\ell+0.5)} + \xi^2\,\overline{G}_{\mathtt{RR}}^{(\ell+1)}, \tag{B.44}$$

[9] More generally, we should also account for off-diagonal components by averaging over pairs of inputs to estimate the appropriate analogs of $K_{[2]}^{(\ell)}$, cf. (5.19). We'd then also want to ensure that the analog of the recursion for this component, e.g., (B.14), is preserved. As our discussion in these bullets is somewhat schematic, we will leave these details to the PyTorch documentation.

[10] One reasonable heuristic solution to this problem, used by the *transformer* architecture, could be *layer normalization* [91]. However, more ideally, the block is not treated as a black box, and instead we use something about the structure of the block itself to find the criticality conditions.

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press

where we've suppressed the layer indices in the residual hyperparameters to declutter the expression.[11] To preserve this component of the two-point correlator of preactivations, we set the right-hand side of this equation equal to ℓ-th-layer correlator. Rearranging, we thus want

$$\left(1-\gamma^2\right)\overline{G}^{(\ell)}_{zz} = \xi^2\,\overline{G}^{(\ell+1)}_{\texttt{RR}} + 2\gamma\xi\,\overline{G}^{(\ell+0.5)}_{\texttt{R}z}. \tag{B.45}$$

Since we've supposedly measured all these quantities, this equation should give simple *analytical* solutions for the residual hyperparameters.[12]

Overall, this hybrid approach realizes one of our goals of using experimentally measurable observables as inputs to an effective theory analysis. We hope that similar ways of thinking will lead to powerful ways of designing and tuning deep-learning models in the future.

[11] Note that unlike the MLP-single-layer block we discussed in §B.2, we cannot in general simply scale away ξ by a rescaling of C_W: for instance, if we have a deep MLP-$L_{\texttt{R}}$-layer block with $L_{\texttt{R}} \gg 1$ built with `ReLU` activations, then it is probably easier to set $C_W = 2$ and adjust ξ at the end. Thus, in general we should think of the initialization hyperparameters of the block as being fixed *first* – to ensure criticality of the block internally – and *then* for each γ, we tune $\xi(\gamma)$ according to (B.45) to ensure criticality of the whole network.

[12] Accordingly, each critical solution (γ, ξ) to the equation (B.45) will then yield architectures with different optimal aspect ratios $r^\star(\gamma, \xi)$. The optimal aspect ratio for a particular (γ, ξ) can be analogously estimated for general residual networks with a hybrid approach by combining a theoretical analysis as we did in §B.3 with measurements of an appropriate combination of four-point connected correlators.

https://doi.org/10.1017/9781009023405.017 Published online by Cambridge University Press