Maximal Initial Learning Rates in Deep ReLU Networks

Gaurav Iyer [1 2] **Boris Hanin** [3] **David Rolnick** [1 2]

Abstract

Training a neural network requires choosing a suitable learning rate, which involves a trade-off between speed and effectiveness of convergence. While there has been considerable theoretical and empirical analysis of how large the learning rate can be, most prior work focuses only on late-stage training. In this work, we introduce the *maximal initial learning rate* η^* – the largest learning rate at which a randomly initialized neural network can successfully begin training and achieve (at least) a given threshold accuracy. Using a simple approach to estimate η^*, we observe that in constant-width fully-connected ReLU networks, η^* behaves differently from the maximum learning rate later in training. Specifically, we find that η^* is well predicted as a power of (depth × width), provided that (i) the width of the network is sufficiently large compared to the depth, and (ii) the input layer is trained at a relatively small learning rate. We further analyze the relationship between η^* and the sharpness λ_1 of the network at initialization, indicating they are closely though not inversely related. We formally prove bounds for λ_1 in terms of (depth × width) that align with our empirical results.

1. Introduction

The learning rate plays a crucial role in the training of deep neural networks. Unfortunately, tuning the learning rate is a tricky task – too large a learning rate can cause the training loss to diverge, while too small a learning rate can result in inefficient use of time and computational resources. The optimal choice of learning rate has been observed to depend non-trivially on many factors, including the data, architecture, optimizer, and initialization scheme. As a result, it can be difficult to theoretically analyze the relationship between the learning rate and other elements of the training framework, and computationally expensive to find the best learning rate in practice.

While there have been numerous works rigorously analyzing effective learning rates both from theoretical and empirical standpoints (Li et al., 2020; Smith et al., 2018), these works tend to focus on behavior during late-stage training, considering which learning rates provide optimal performance (Park et al., 2019), or lead to convergence guarantees (Wang et al., 2022). However, it is not clear that optimal learning rates early in training follow similar patterns to those near convergence, and indeed a variety of heuristics for learning rate scheduling suggest that an early change in learning rate can actually boost performance (Goyal et al., 2017).

In this paper, we consider both empirically and theoretically how large the learning rate can be in early training. Our main contributions are as follows:

- In §3.1, we introduce the *maximal initial learning rate* η^* – the largest learning rate at which a randomly initialized neural network can successfully begin training – and show how it can be computed.

- For fully-connected deep ReLU networks, we empirically identify a power law relating the maximal initial learning rate and the product of width and depth:

$$\mathbb{E}[\ln \eta^*] = -\alpha \ln(\text{depth} \times \text{width}) + \gamma_1,$$

which can be observed in Figure 1[1]. Notably, while prior theoretical work (Karakida et al., 2019) suggests that at the end of training $\alpha = 1$, we find that at the start of training $\alpha < 1$, allowing for larger initial learning rates.

- We show empirically that the sharpness λ_1 (i.e. the largest eigenvalue of the training loss Hessian) at initialization is a function of η^*, but does not necessarily reflect the inverse relationship observed in late-stage training. We illustrate this in Figure 4.

---

[1]School of Computer Science, McGill University, Montreal, Canada [2]Mila – Quebec AI Institute, Montreal, Canada [3]Dept. of Operations Research & Financial Engineering, Princeton University, Princeton, USA. Correspondence to: Gaurav Iyer <gaurav.iyer@mila.quebec>.

---

[1]Our empirical results suggest that α is indeed task-dependent. While our results in Figure 1 largely show $\alpha \approx 0.7$ for sufficiently wide networks, significant differences can be seen in Figure 11(a), where we obtain smaller α on Gaussian data.
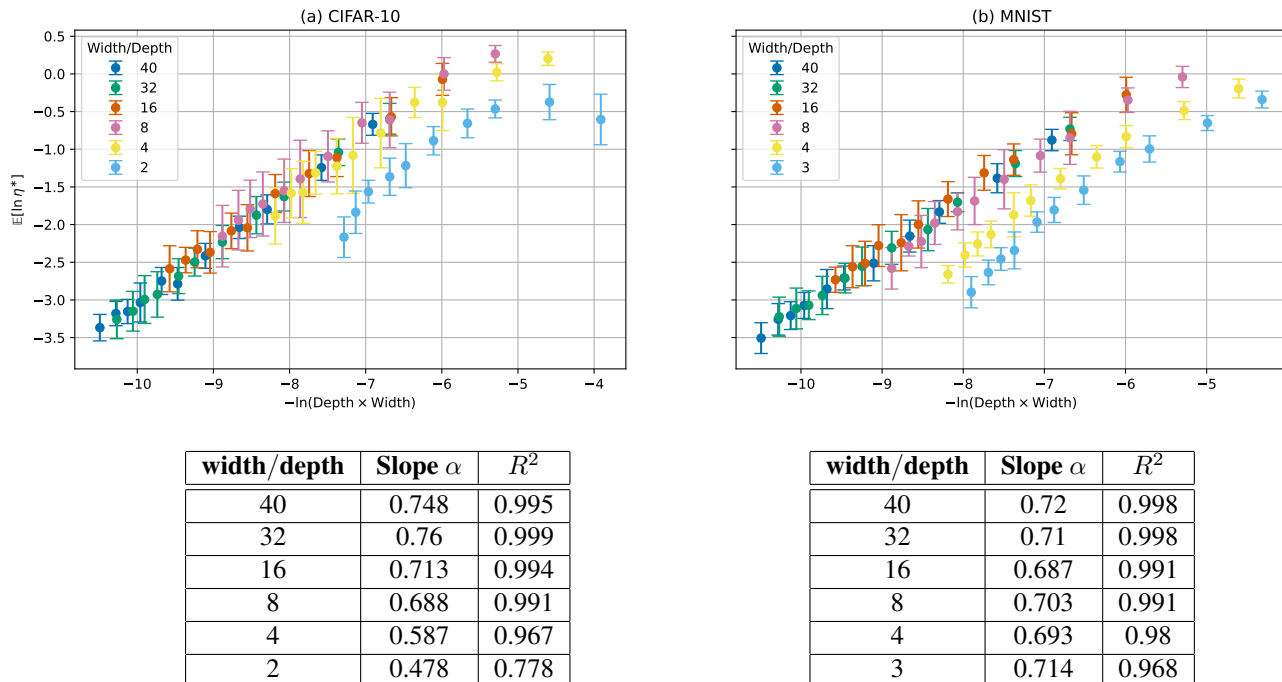
width/depth	Slope α	R^2
40	0.748	0.995
32	0.76	0.999
16	0.713	0.994
8	0.688	0.991
4	0.587	0.967
2	0.478	0.778

width/depth	Slope α	R^2
40	0.72	0.998
32	0.71	0.998
16	0.687	0.991
8	0.703	0.991
4	0.693	0.98
3	0.714	0.968

Figure 1: Relationship between maximal initial learning rate η^* and architecture for fully-connected networks with different values of width/depth. We use depths $\in \{5, 7, 10, 12, 15, 18, 20, 23, 25, 27, 30\}$, for 25 initializations per architecture. If width/depth is sufficiently large, the expected η^* displays a strong power relationship with $(\text{depth} \times \text{width})^{-1}$. Interestingly, all such η^* lie on the same line regardless of the exact width/depth value. As width/depth becomes smaller, networks deviate from the power relationship. The input layer of networks is trained at $\eta \cdot 10^{-2}$, where η is the learning rate for the rest of the network. We report slopes and coefficient of determination values for each width to emphasize the linear fit.

- In §5, we prove power law upper bounds on λ_1 at initialization as a function of $(\text{depth} \times \text{width})$ by studying the Frobenius norm of the Loss Hessian.

In the process, we also provide a counterexample to a claim about learning rates and sharpness made in Gilmer et al. (2022), as detailed in §4.3.

2. Related Work

Large learning rates are a topic of considerable interest (Smith & Topin, 2018; Li et al., 2019; Jastrzebski et al., 2018). For instance Lewkowycz et al. (2020), to which we compare our work more thoroughly below, investigate the benefits of large learning rates to generalization and predict maximum learning rates for some architectures in relation to the sharpness (see below). Through theoretical analysis of the Fisher information matrix and its statistics, Karakida et al. (2019) derive a relationship between the architecture of a network and the largest learning rate that would allow the network to converge to the global minimum when trained with SGD. They state that this learning rate should scale linearly with $(\text{depth} \times \text{width})$ for constant width fully-connected networks. Park et al. (2019) relate the optimal

learning rate (in the sense of optimal performance) to the effective width of neural networks by studying the normalized noise scale – a quantity derived from the learning rate, batch size, and training set size.

Cohen et al. (2021) introduce "progressive sharpening" and the "edge of stability" regime by investigating the evolution of sharpness over the course of training, which is extended to adaptive gradient methods in Cohen et al. (2022). More recently, there have also been several theoretical investigations (Ahn et al., 2022; Arora et al., 2022; Li et al., 2022) into this phenomenon. Gilmer et al. (2022) add to this line of work by looking at early training instabilities through the lens of sharpness and argue that seemingly different methods such as learning rate warmup and gradient clipping stabilize the learning process through the same mechanism – by reducing sharpness early in training. Jastrzebski et al. (2020) similarly look at the effect of hyperparameters used in the early stages of training and find that they determine properties of the entire training trajectory.

We present an average-case analysis of the maximal initial learning rate, and how it relates to the architecture and expected sharpness. Such average-case analyses can be useful in identifying gaps between theoretical possibilities and
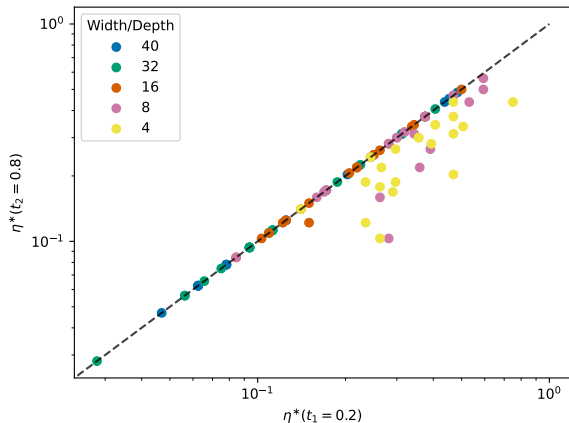
Figure 2: Behavior of the maximal initial learning rate η^* as network architecture and threshold accuracy t vary in constant-width fully-connected ReLU networks. (a) illustrates how η^* depends on t as a function of width/depth for networks trained on MNIST. We plot η^* at $t = 0.2$ and 0.8, on the x and y axes respectively, and for 5 initializations per architecture with depths $\in \{5, 10, 15, 20\}$ and width/depth $\in \{4, 8, 16, 32, 40\}$. There is no change in η^* for architectures with sufficiently large width/depth, and such points fall on the $x = y$ line. Otherwise, its value declines as t becomes larger. Essentially, the value of η^* is stable across a wide range of threshold accuracies t in constant width architectures with sufficiently large width/depth.

practical observations. Hanin et al. (2022); Hanin & Rolnick (2019a) and Hanin & Rolnick (2019b) provide average-case analyses for expected length distortion, the number of linear regions, and the number of activation regions in deep ReLU networks. For further examples of average-case analyses, see Shalev-Shwartz et al. (2017); Raghu et al. (2017).

We also point the reader to several prior articles which show that in wide fully-connected networks, it is the width-to-depth ratio, rather than depth or width separately that effectively controls the stability of optimization – this is the reason we separate architectures based on this ratio in our experimental results. Examples include work about fully-connected networks at initialization, which study the fluctuations of the forward pass (Hanin & Rolnick, 2018), the fluctuations in the backward pass (Hanin, 2018; Hanin & Nica, 2019a; Hanin, 2021), and the extent of feature learning early in training (Hanin & Nica, 2019b). Moreover, for large values of width/depth, many novel results of this kind were also obtained for networks *after training* in Roberts et al. (2021) (especially Chapter ∞).

We conclude the literature review by comparing our maximal initial learning rate to the maximal stable learning rate of Lewkowycz et al. (2020) in more depth. To start, note that these two notions of maximal learning rate are different. As we illustrate in Figure 7(b), our maximal initial learning rate can sometimes lead to instability late in training, suggesting that the maximal initial learning rate is likely larger than the maximal stable learning rate. A core proposal of Lewkowycz et al. (2020) is that the maximal stable learning rate has the form c_{act}/λ_1, where c_{act} is a constant and λ_1 is the largest eigenvalue of the NTK. For MSE loss and linear one-layer networks, Lewkowycz et al. (2020) suggest both theoretically and empirically that $c_{act} = 4$. For more general architectures and cross-entropy losses, however, Lewkowycz et al. (2020) obtains different values of c_{act}. Thus, in situations where the maximal initial and

maximal stable learning rates are comparable, our empirics can be viewed as capturing more of the full architecture dependence of c_{act}, suggesting that perhaps c_{act}/λ_1 scales like (depth $\times$ width)$^{-\alpha}$. Finally, as detailed in Appendix A of their work, experiments in Lewkowycz et al. (2020) sometimes use learning rate drops, cosine scheduling, data augmentation, batch normalization, and weak L_2 regularization in experiments. The experiments we undertake in this work do not make use of these, further complicating direct comparisons.

3. Preliminaries

3.1. The Maximal Initial Learning Rate

We define the *maximal initial learning rate* η^* to be the largest, constant learning rate at which a given network can achieve validation accuracy of at least t, where t is a given threshold accuracy.

The choice of learning rate is largely dictated by the data being used to train the network and its architecture. Since a theoretical formulation of the data itself is unrealistic, the learning rate needs to be empirically tuned for each problem statement, which is further affected by changes in the training setup. By introducing the maximal initial learning rate η^*, exploring its properties, and relating it to the architecture, we aim to make it easier to find large learning rates that work in practice.

Furthermore, several recent lines of work observe that the early phase of training can heavily impact training dynamics and performance at later stages (Frankle et al., 2020; Jastrzebski et al., 2020) – from this perspective, it is important to consider the behavior of η^* and how it changes as a function of architecture and training setup.

In Algorithm 1 we describe a simple method that can be used to approximate η^*. We take as input a network initialization,
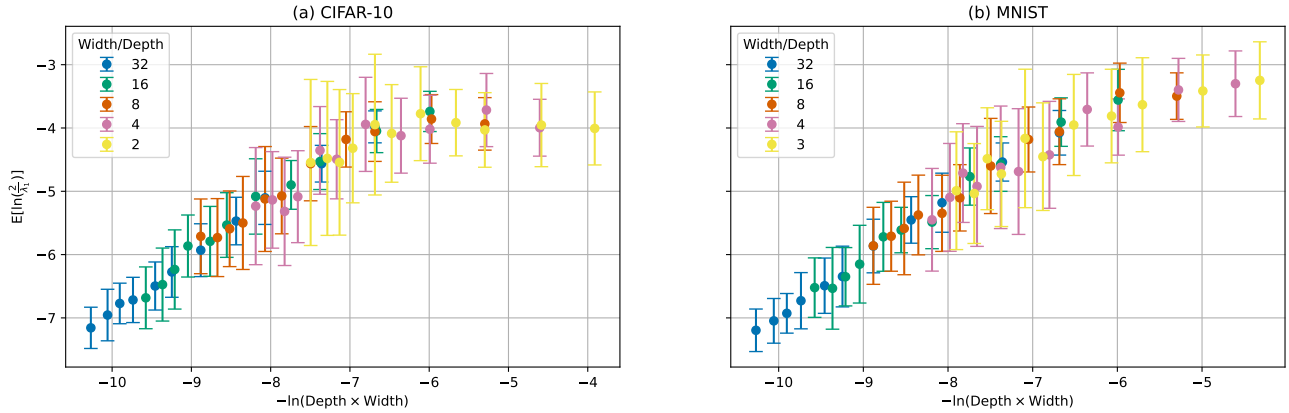
Figure 3: Plots of $2/\lambda_1$ at initialization as a function of (depth × width) for various fully-connected, Kaiming-initialized architectures, evaluated on CIFAR-10 (left) and MNIST (right). λ_1 is measured by considering the complete dataset as a single batch of data and is averaged over 25 initializations for each architecture, with error bars shown. See Figure 1 for architecture depths used. We observe that for sufficiently large (depth × width), the maximal initial learning rate η^* and $2/\lambda_1$ both show similar power relationships.

Algorithm 1 Maximal Initial Learning Rate η^*

Define threshold accuracy t, and upper and lower learning rates u and l respectively.
for a small number of search iterations s **do**
  Compute $m = \frac{u+l}{2}$ i.e. the midpoint of u and l
  for each epoch in a small number of epochs e **do**
    Train network at learning rate m
    Evaluate validation accuracy a
    If $a \geq t$, then break out of inner loop, and $l \leftarrow m$
  end for
  If $a < t$ after training for e epochs, then $u \leftarrow m$
end for
The last value of m satisfying $a \geq t$ is the desired η^*.

---

threshold accuracy t, and lower and upper learning rates l and u respectively. We perform a binary search on the continuous space of learning rates $\in (l, u)$ to identify η^*. More specifically, for each midpoint learning rate $m = \frac{u+l}{2}$, we train the network from initialization for a small number of epochs. If this trained network, at any point during training, achieves validation accuracy at least t, then we set $l = m$. Otherwise, we set $u = m$. The next search is performed in the interval (l, u). This is repeated for a small number of search iterations, and the final value of m that achieves validation accuracy at least t is output as η^*.

3.2. Experimental Setup

We primarily focus on constant width, fully-connected deep ReLU networks trained with SGD, that are initialized with the Kaiming normal initialization scheme. The batch size is set to 128 across all our experiments. We discuss reasons

for making this choice of initialization in subsection 4.3.

When using Algorithm 1, we set threshold accuracy t to be the accuracy that a linear classifier achieves on the given dataset, along with the number of training epochs $e = 10$. This ensures that networks trained at the maximal initial learning rate perform adequately while taking into account task difficulty. Namely, for MNIST and CIFAR-10 we use $t = 0.925$ and $t = 0.34$ respectively. Upper and lower learning rate limits u and l are set heuristically; we use $l = 0.0$ for all our experiments, and find that $s = 5$ search iterations are sufficient in practice to calculate η^*. For computing the sharpness λ_1, we use PyHessian (Yao et al., 2020).

4. Main Empirical Results

4.1. Maximal Initial Learning Rate and Architecture

We have described the maximal initial learning rate η^* and explored its behavior under various training setups and threshold accuracies t. We now consider how η^* depends on network architecture. From Figure 1, we see that if the ratio width/depth is sufficiently large and the input layer is trained at a sufficiently small learning rate, then the expected value of η^* is related to (depth × width) through a power law:

$$\mathbb{E}[\ln \eta^*] = -\alpha \ln(\text{depth} \times \text{width}) + \gamma_1$$

We also note that network architectures that deviate from the trend show much smaller η^*. In our experiments, these networks sometimes fail to beat the performance of a linear classifier, leading to no valid η^* being found at all. For experimental results on Fashion-MNIST, we point the reader
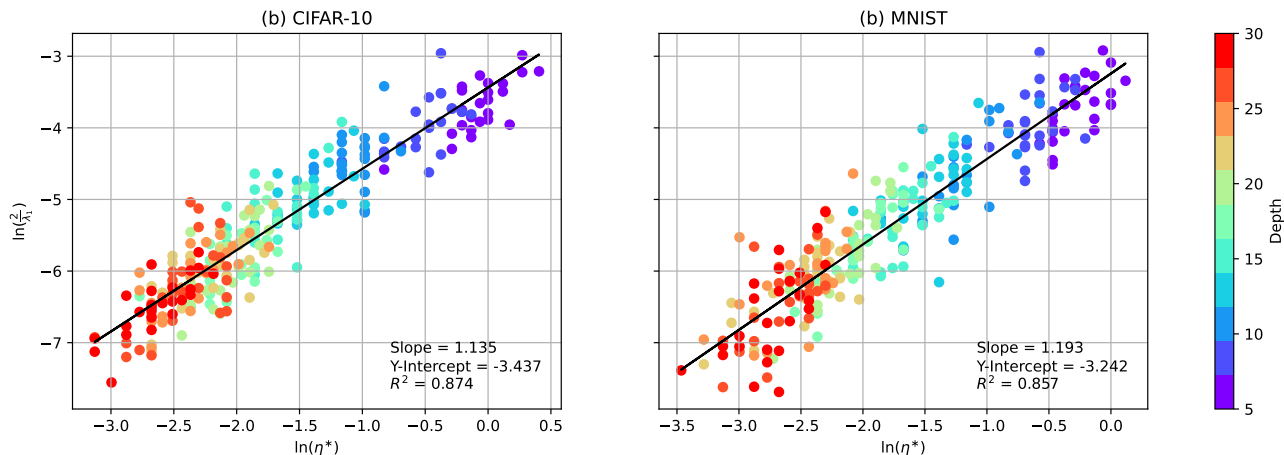
Figure 4: Plots of $2/\lambda_1$ against η^* for 25 initializations per architecture, evaluated on CIFAR-10 and MNIST. We use architectures with width/depth = 16, with the same depths as in Figure 1. We find that $\eta^* > 2/\lambda_1$ and that $\ln(2/\lambda_1) = \beta \ln \eta^* + \gamma_2$ for $\beta \neq 1$, strongly contrasting with patterns observed later in training in e.g. Cohen et al. (2021).

to Figure 10 (a) in the Appendix.

We use a small learning rate for the input layer weights because in virtually all principled initialization schemes the learning rate of a weight depends on the width of the previous layer, with larger widths corresponding to smaller learning rates (see e.g. Table 1 in Yang & Hu (2021)).

Thus, while our experiments utilize networks that have constant *hidden* layer widths, this suggests that the maximal initial learning rate for input layer weights may differ from the maximal learning rate appropriate for deeper layers, especially when the input dimension is large compared to the network width. In accordance with these expectations, we find that the behavior of η^* is essentially unchanged if the input layer weights are frozen at initialization, while at smaller values of (depth $\times$ width) the maximal initial learning rate η^* deviates from the power-law predictions we otherwise observe (see Appendix C).

In the context of the above experiments and results, we pose the following questions for consideration in future work:

- Through theoretical analysis of the Fisher information matrix, Karakida et al. (2019) have suggested that the largest learning rate ensuring global convergence of SGD should scale linearly with (depth $\times$ width):

$$\mathbb{E}[\ln \eta^*] \propto -\alpha \ln(\text{depth} \times \text{width}), \quad \alpha = 1$$

However, none of the setups considered here display this behavior. While there is no direct contradiction here, since global convergence is not guaranteed when training at η^*, how can one explain the discrepancy between these regimes?

- Is there a method for initializing the input layer that

allows it to train normally while preserving the strong trends observed in Figure 1? It is possible that answering this question could shed light on the conflict between different methods for initializing the input layer.

- What is the full functional relationship between (depth $\times$ width) and η^* at moderate to large values of width/depth?

4.2. Relationship to Expected Sharpness at Initialization

The *sharpness* λ_1 of a network is defined as the maximum eigenvalue of the training loss Hessian, and is often associated with the learning rate at which the network is trained. In particular, classical optimization tells us that the learning rate must be no larger than $2/\lambda_1$ to guarantee the convergence of SGD to the global minimum. However, this notion has lately been questioned in the context of deep neural networks (Cohen et al., 2021; Lewkowycz et al., 2020).

This motivates a need for a deeper understanding of sharpness and its connection to the learning rate and architecture in deep neural networks. To this end, we now explore how the expected sharpness of a network at initialization relates to architecture and maximal initial learning rate. In Figure 3, we consider the value of $2/\lambda_1$ at initialization as a function of (depth $\times$ width)$^{-1}$, finding a power law relationship as with η^*.

Note that the sharpness at initialization is the same regardless of whether the input layer is trained at the same or smaller learning rate as the rest of the network since no training is involved in computing sharpness at initialization.

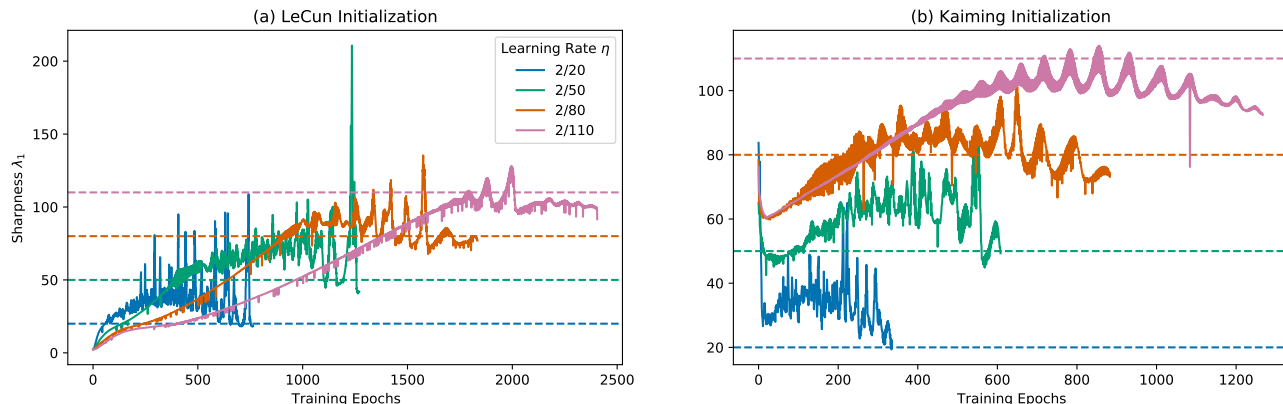Next, we directly compare the sharpness and the max-

Figure 5: Sharpness λ_1 as a function of training epochs in full-batch gradient descent, showcasing "progressive sharpening" – the tendency of λ_1 to continually rise until it reaches and hovers around the value of $2/\eta$ – for (a) LeCun and (b) Kaiming initialization. We replicate the experimental setup of Cohen et al. (2021), training the same network architecture until 99% training accuracy is reached, for learning rates $\eta \in \{2/20, 2/50, 2/80, 2/110\}$, on a 5k subset of CIFAR-10. Note that λ_1 at initialization is significantly larger for Kaiming-initialized networks than for LeCun-initialized networks (also see Figure 8 in the Appendix). The steep drop in sharpness at the beginning of training these networks may impact the value of η^* since the maximal initial learning rate depends on the state of the network in early-stage training as well as at initialization itself.

imal initial learning rate at initialization (see Figure 4, and Figure 10 (b) in the Appendix), using networks with width/depth $= 16$ and training the input layer at a small learning rate as in Figure 1. While the work of Cohen et al. (2021) suggests that $\eta \sim 2/\lambda_1$ as networks converge to global optima, we find that at initialization the data closely fit a different power law, with $\ln(2/\lambda_1) \sim \beta \ln \eta^*$ for $\beta \neq 1$. To study this comparison, we plot $2/\lambda_1$ as a function of η^*, using networks with width/depth $= 16$ and train the input layer at a small learning rate. This is done in order to compute η^* that preserves the correlations observed in previous experiments.

Since judging linear fits on a log-log plot can be difficult, we provide another version of this figure in the Appendix in Figure 12, averaging over initializations for each architecture. This suggests that the coefficient $\beta \neq 1$ estimated from Figure 4 is unlikely to be a product of noise.

It is also worth noting that for each point in the above plots, η^* is clearly greater than $2/\lambda_1$. Recall that by definition, the computed η^* ensures that a network initialization performs at least as well as a linear classifier on a given dataset, without diverging. This goes against the established wisdom of "$\eta \leq 2/\lambda_1$" for the convergence of SGD, hence supporting the recent lines of work that question this notion.

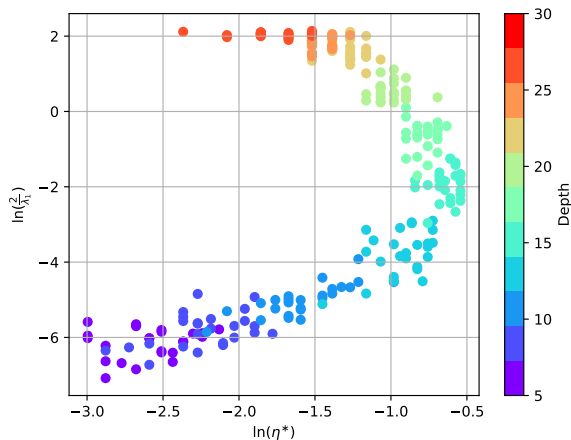### 4.3. Relationship to Edge of Stability

In the previous experiments, we considered the sharpness at initialization, but since the definition of the maximal initial learning rate involves the ability to train a network past initialization, it makes sense that it could also be influenced

by the value of the sharpness immediately following initialization. To gain further insight into this behavior, we revisit the concept of "progressive sharpening" identified in Cohen et al. (2021). This term refers to the tendency when training at learning rate η for the sharpness λ_1 to continually rise until it reaches and hovers around the value of $2/\eta$. In Figure 5, we replicate the experimental setup of Cohen et al. (2021). In particular, we do so for both the LeCun initialization used in Cohen et al. (2021), which initializes weights from a uniform distribution on $[-1/\text{fan-in}, 1/\text{fan-in}]$, and for Kaiming initialization, which initializes weights from a centered Gaussian with variance $2/\text{fan-in}$.

We use Kaiming initialization in this paper both because it is the more common initialization and also because in some sense it is the "correct" way to initialize deep ReLU networks. Namely, Hanin & Rolnick (2018) show that in ReLU networks, Kaiming initialization prevents the mean size of the activations from growing exponentially large or small as a function of the depth, which occurs in LeCun initialization. Hanin (2018) shows a similar benefit to Kaiming initialization in reducing the problem of gradient explosion (Bengio et al., 1994).

Figure 5 illustrates that λ_1 at initialization scales quite differently for the two initialization schemes. For LeCun-initialized networks, λ_1 exponentially vanishes as depth increases (an effect more visible in Figure 8 in the Appendix, and described formally in 2), while it increases modestly with width and depth for Kaiming-initialized networks.

We find that the "edge of stability" phenomenon occurs for both initialization schemes, but takes slightly differ-

Figure 6: Plot of $2/\lambda_1$ against η^* evaluated on CIFAR-10 for networks initialized with Neural Tangent Kernel (NTK) parametrization. Other experimental details are identical to those in Figure 4. We find that in NTK-initialized networks, $2/\lambda_1$ and η^* display a relationship very different from that observed in Kaiming-initialized networks. We also note key differences in empirical results between Kaiming and NTK-initialized networks: (a) $2/\lambda_1$ increases with depth (i.e. λ_1 decreases with depth), and (b) the relationship displayed between $2/\lambda_1$ and η^* is non-monotonic – as $2/\lambda_1$ (and network depth) increases, η^* first increases and then decreases.

ent forms. Namely, λ_1 at initialization is much larger for Kaiming-initialized networks and steeply drops off at the beginning of training, before rising (or dropping more slowly) until it hovers slightly above the $2/\eta$ line. This behavior is pertinent to our consideration of η^*, since the precipitous drop of λ_1 in very early training means that it is possible η^* is able to take on larger values than it could if the sharpness remained at its initial value. This is thus a possible explanation of the behavior observed in Figure 4.

The Neural Tangent Kernel (NTK) parametrization (Jacot et al., 2018; Sohl-Dickstein et al., 2020) is an initialization scheme popularly used to analyze networks in the infinite width limit. In the context of our experiments, it provides an alternate scaling of network weights with respect to layer width. In Figure 6 we plot $2/\lambda_1$ against η^* for NTK-initialized networks and find that this relationship is highly different from that observed in previous experimental results.

We conclude this section by recalling Gilmer et al. (2022), which claims that "When the learning rate only slightly exceeds $2/\lambda_1$, optimization is unstable until the parameters move to a region with smaller λ_1." This claim is contradicted by our Figure 5 (specifically the pink curve for Kaiming initialization and $\eta = 2/110$), from which we find that in fact the learning rate need not exceed $2/\lambda_1$ for the parameters to move to a region with smaller λ_1. To elaborate, note that the value of λ_1 (in the pink curve in Figure 5 (b)) at the beginning of training is much smaller than $2/\eta$ (where $\eta = 2/110$) at initialization. Even in this case, we observe that parameters move to a region with smaller λ_1, implying that the condition stated by Gilmer et al. (2022) is not necessary for this to occur.

5. Estimates for the Largest Eigenvalue of the Loss Hessian at Initialization

In this section, we present our main theoretical result, Theorem 1, which computes the average squared Frobenius norm of the loss Hessian at initialization. Before stating it exactly, we present informally a simple corollary that gives bounds for the largest eigenvalue of the loss Hessian at initialization.

Corollary 1 (Informal). Consider a randomly initialized fully connected ReLU network of constant width, and denote by λ_1 the largest eigenvalue of the Hessian of an empirical MSE loss. We have the following upper bound on the largest eigenvalue:

$$\mathbb{E}\left[|\lambda_1|\right] = O(\text{depth} \times \text{width})$$

and the following lower bound on the sharpness:

$$\mathbb{E}\left[\frac{2}{\text{sharpness}}\right] = \mathbb{E}\left[\frac{2}{|\lambda_1|}\right] = \Omega\left(\frac{1}{\text{depth} \times \text{width}}\right),$$

where the average in both estimates is over initialization.

The preceding estimates show that depth times width, the key parameter which we found determines the maximal initial learning rate and the sharpness, naturally appears when computing the Frobenius norm of the loss Hessian (see also Theorem 1) and hence can be used to obtain bounds on $|\lambda_1|$ and sharpness at initialization.

5.1. Formal Statements

In order to state this Corollary and Theorem 1 more precisely, we need some notation. We consider a ReLU network, which for an input $x \in \mathbb{R}^{n_0}$ outputs $z_1^{(L+1)}(x) \in \mathbb{R}$ via hidden layer pre-activations $z^{(\ell)}(x) \in \mathbb{R}^{n_\ell}$ as follows:

$$z_i^{(\ell+1)}(x) = \begin{cases} \sum_{j=1}^{n_\ell} W_{ij}^{(\ell)} \sigma\left(z_j^{(\ell)}(x)\right), & \ell \geq 1 \\ \sum_{j=1}^{n_\ell} W_{ij}^{(1)} x_j, & \ell = 0 \end{cases}$$

for $i = 1, \ldots, n_\ell$. Note that we have set the biases in the network to 0. Moreover, we will assume that the weights are independent Gaussians $W_{ij}^{(\ell)} \sim \mathcal{N}(0, 2/n_{\ell-1})$.

Our goal is to understand the Hessian

$$H^{(L+1)} := \left(\partial_{W_{ij}^{(\ell)}} \partial_{W_{i'j'}^{(\ell')}} \mathcal{L} \right),$$

(here the indices summarize $1 \leq \ell, \ell' \leq L+1$, $1 \leq i, i' \leq n_\ell$, $1 \leq j, j' \leq n_{\ell-1}$) of the empirical MSE

$$\mathcal{L} = \frac{1}{2k} \sum_{i=1}^{k} \left(z_1^{(L+1)}(x_i) - y_i \right)^2$$

over k input-output pairs (x_i, y_i). Specifically, we compute the mean squared Frobenius norm of $H^{(L+1)}$ given by

$$\mathbb{E} \left[\sum_{\ell, \ell'=1}^{L+1} \sum_{i,i'=1}^{n_\ell} \sum_{j,j'=1}^{n_{\ell-1}} \left(\partial_{W_{ij}^{(\ell)}} \partial_{W_{i'j'}^{(\ell')}} \mathcal{L} \right)^2 \right],$$

where the average is over the Gaussian distribution of the weights. Our main result is

Theorem 1. Fix $n_0 \geq 1$ as well as $c, C > 0$ and a network input x satisfying $\|x\|^2 = n_0$. Then, there exists a constant C_1, depending only on c, C with the following property. For any $L \geq 1$ there exists a constant C_2, depending only on L, c, C such that if $cn \leq n_1, \ldots, n_L \leq Cn$, then

$$\left| \mathbb{E} \left[\left\| H^{(L+1)} \right\|_F^2 \right] - C_1 n^2 L^2 \right| \leq C_2 n.$$

The preceding Theorem gives the following upper bound on the largest eigenvalue of $H^{(L+1)}$ and its reciprocal:

Corollary 2 (Precise Statement of Corollary 1). With the notation of Theorem 1, denote by λ_1 the largest eigenvalue of the Hessian of an empirical MSE loss. There exists a constant $K > 0$ depending only on the constants n_0, C, c from Theorem 1 such that

$$\mathbb{E} \left[|\lambda_1| \right] \leq KnL \left(1 + O(n^{-1}) \right) \quad \text{and}$$

$$\mathbb{E} \left[2/\text{sharpness} \right] = \mathbb{E} \left[2/|\lambda_1| \right] \geq \frac{1}{KnL} \left(1 + O(n^{-1}) \right).$$

These results hold for Kaiming initialization. For LeCun initialization the same results hold, except K must be replaced by $2^{-L/2} K$.

Proof. Since the squared Frobenius norm of $H^{(L+1)}$ is the sum of squares of its eigenvalues, Theorem 1 yields

$$\mathbb{E} \left[|\lambda_1| \right] \leq \sqrt{\mathbb{E} \left[\lambda_1^2 \right]} \leq \sqrt{\mathbb{E} \left[\left\| H^{(L+1)} \right\|_F^2 \right]}$$
$$= C_1^{1/2} nL (1 + O(n^{-1})).$$

Further, since $x \mapsto 1/x$ is convex on $(0, \infty)$, we have

$$\mathbb{E} \left[\frac{2}{|\lambda_1|} \right] \geq \frac{2}{\mathbb{E} \left[|\lambda_1| \right]} \geq \frac{2}{C_1^{1/2} nL} (1 + O(n^{-1})).$$

Since depth L ReLU networks are homogeneous of degree L in their weights, The change from Kaiming to He initialization causes the network output (and hence its derivatives) to be re-scaled by a factor of $2^{-L/2}$. □

5.2. Proof Outline for Theorem 1

Our strategy for estimating the Frobenius norm of $H^{(L+1)}$ is to proceed recursively in L. To explain the main idea (full details in the Appendix) we need some notation. First, we will use μ, ν to denote generic variables indexing network weights. Next, for any $\ell = 1, \ldots, L$ and any expressions $f_k(z)$ depending on z and $\partial_\mu z$ we write

$$Y^{(\ell)}[f_1, \ldots, f_k] := \mathbb{E} \left[\sum_{\mu \leq \ell} \frac{1}{n_\ell^k} \sum_{j_1, \ldots, j_k=1}^{n_\ell} \prod_{i=1}^{k} f_i(z_{j_i}^{(\ell)}) \right].$$

Thus, for example,

$$Y^{(\ell)}[z\partial_\mu z] = \mathbb{E} \left[\sum_{\mu \leq \ell} \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} z_j^{(\ell)} \partial_\mu z_j^{(\ell)} \right].$$

In both cases, $\mu \leq \ell$ denotes the collection of weights in layers $1, \ldots, \ell$. Similarly, if the functions $f_k(z)$ depend in addition on $\partial_\nu z$ and $\partial_{\mu\nu} z$ then we we will write

$$Y^{(\ell)}[f_1, \ldots, f_k] := \mathbb{E} \left[\sum_{\mu, \nu \leq \ell} \frac{1}{n_\ell^k} \sum_{j_1, \ldots, j_k=1}^{n_\ell} \prod_{i=1}^{k} f_i(z_{j_i}^{(\ell)}) \right].$$

Thus, for example, $Y^{(\ell)} \left[(\partial_\mu z)^2, z\partial_{\mu\nu} z \right]$ equals

$$\mathbb{E} \left[\sum_{\mu \leq \ell} \frac{1}{n_\ell^2} \sum_{j_1, j_2=1}^{n_\ell} \left(\partial_\mu z_{j_1}^{(\ell)} \right)^2 z_{j_2}^{(\ell)} \partial_{\mu\nu} z_{j_2}^{(\ell)} \right].$$

The key steps in proving Theorem 1 are now as follows:

1. Integrate out the weights in layer $L + 1$ to rewrite $\mathbb{E} \left[\left\| H^{(L+1)} \right\|_F^2 \right]$ in terms of various $Y^{(L)}$'s. This is done in the Appendix in Lemma 2 and Corollary 4. Since the Hessian involves second derivatives, the $Y^{(L)}$ that appear depend on various combinations of $z, \partial_\mu z, \partial_\nu z, \partial_{\mu\nu} z$.

2. Obtain recursive expressions for the $Y^{(\ell+1)}$'s that depend only $z, \partial_\mu z$ in terms of the corresponding $Y^{(\ell)}$'s at the previous layer. This is done in Lemma 3. Each such recursion is derived by considering two cases.

First, the case where the parameter μ is a weight in layer $\ell + 1$. This gives expressions no longer containing any derivatives that depend only on moments of the norm of the vector of pre-activations $z^{(\ell)}$ at layer ℓ. Such moments are well-known. Second, the case where the parameter μ is a weight in layers $1, \ldots, \ell$. By explicitly integrating out the weights in layer $\ell + 1$, we obtain expressions involving various $Y^{(\ell)}$.

3. Solve the recursions for the $Y^{(\ell+1)}$'s that depend only on $z, \partial_\mu z$ to understand how they grow with depth and width. This is done in Corollary 5.

4. Obtain consistent recursive expressions for the $Y^{(\ell+1)}$'s that depend only $z, \partial_\mu z, \partial_\nu z, \partial_{\mu\nu} z$ in terms of $Y^{(\ell)}$'s that depend only on the same expressions. This is done in Lemma 4. The strategy is the same as in deriving Lemma 3, except one must consider three cases: μ, ν are both weights in layer $\ell + 1$, exactly one of μ, ν is a weight in layer $\ell + 1$, and neither of μ, ν are weights in layer $\ell + 1$.

5. Solve the recursions for the $Y^{(\ell+1)}$'s that depend only $z, \partial_\mu z, \partial_\nu z, \partial_{\mu\nu} z$ to understand how they grow with depth and width. This is done in Corollary 6.

6. Combine Corollaries 4, 5, and 6 to obtain estimates for the average of the squared Frobenius norm of $H^{(L+1)}$.

6. Conclusion

We have introduced the maximal initial learning rate along with a simple algorithm to compute it. We empirically show that the maximal initial learning rate is closely related to the architecture and sharpness λ_1 at initialization in Kaiming-initialized fully-connected ReLU networks through:

$$\mathbb{E}[\ln \eta^*] = -\alpha \ln(\text{depth} \times \text{width}) + \gamma_1,$$
$$\ln(2/\lambda_1) = \beta \ln \eta^* + \gamma_2, \quad \beta \neq 1$$

as long as the network's width/depth is sufficiently large and the input layer is trained at a relatively small learning rate. Further, we formally prove bounds for the sharpness in terms only of the value of (depth $\times$ width).

To close, we emphasize several limitations and directions for future work. First, our experiments were performed only for constant width fully-connected ReLU networks trained by vanilla SGD with a fixed batch size. It would be therefore be interesting to further understand the architecture dependence of the maximal initial learning rate on: batch size, the presence of non-constant hidden layer widths, non-ReLU activations, non-fully connected layers, and the presence of normalization (e.g. BatchNorm). These factors can significantly impact network behavior early in training, which would, in turn, limit the direct application of our results in practical settings. For a rather preliminary investigation of maximal learning rates in ResNets see Appendix G.

Second, there is a rich vein of prior work concerning the dependence of learning rate and details of the optimization protocol. It would therefore be of interest to understand how the maximal initial learning rate η^* varies with batch size (Goyal et al., 2017; Jastrzebski et al., 2018; Hoffer et al., 2017; Smith et al., 2017; 2021) as well as momentum coefficient, ℓ_2 regularization strength, and data augmentation scheme. Similarly, it could be useful to study the relationship between architecture and η^* when using adaptive optimizers such as Adam or Adagrad.

Further, both our experiments and theoretical analyses focused on optimization with a single fixed learning rate. In practice, however, learning rate protocols ranging from a simple learning rate drop after a fixed number of epochs to more intricate schemes such as warmup (Goyal et al., 2017) or cosine schedules can improve network performance. Developing a theory of maximal learning rates that is valid throughout training could be of significant value.

Finally, we do not have a theoretical explanation that would predict the somewhat exotic power-law exponents α, β in the dependence of $\ln \eta^*$ on (depth $\times$ width) and on $\ln(2/\lambda_1)$, and it would be interesting to understand their origin.

Acknowledgments

References

Ahn, K., Zhang, J., and Sra, S. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning (ICML)*, 2022.

Arora, S., Li, Z., and Panigrahi, A. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning (ICML)*, 2022.

Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 1994.

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

Cohen, J. M., Ghorbani, B., Krishnan, S., Agarwal, N., Medapati, S., Badura, M., Suo, D., Cardoze, D., Nado, Z., Dahl, G. E., and Gilmer, J. Adaptive gradient methods at the edge of stability. *Preprint arXiv:2207.14484*, 2022.

Frankle, J., Schwab, D. J., and Morcos, A. S. The early phase of neural network training. In *International Conference on Learning Representations (ICLR)*, 2020.

Gilmer, J., Ghorbani, B., Garg, A., Kudugunta, S., Neyshabur, B., Cardoze, D., Dahl, G. E., Nado, Z., and Firat, O. A loss curvature perspective on training instabilities of deep learning models. In *International Conference on Learning Representations (ICLR)*, 2022.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *Preprint arXiv:1706.02677*, 2017.

Hanin, B. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hanin, B. Correlation functions in random fully connected neural networks at finite width. *Preprint arXiv:2204.01058*, 2021.

Hanin, B. and Nica, M. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 2019a.

Hanin, B. and Nica, M. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations (ICLR)*, 2019b.

Hanin, B. and Rolnick, D. How to start training: The effect of initialization and architecture. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hanin, B. and Rolnick, D. Deep ReLU networks have surprisingly few activation patterns. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Hanin, B. and Rolnick, D. Complexity of linear regions in deep networks. In *International Conference on Machine Learning (ICML)*, 2019b.

Hanin, B., Jeong, R. S., and Rolnick, D. Deep ReLU networks preserve expected length. In *International Conference on Learning Representations (ICLR)*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD. *Preprint arXiv:1711.04623*, 2018.

Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., and Geras, K. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representation (ICLR)*, 2020.

Karakida, R., Akaho, S., and Amari, S.-i. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: the catapult mechanism. *Preprint arXiv:2003.02218*, 2020.

Li, Y., Wei, C., and Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Li, Z., Wang, Z., and Li, J. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. *Preprint arXiv:2207.12678*, 2022.

Park, D., Sohl-Dickstein, J., Le, Q., and Smith, S. The effect of network width on stochastic gradient descent and generalization: an empirical study. In *International Conference on Machine Learning (ICML)*, 2019.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the expressive power of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Roberts, D. A., Yaida, S., and Hanin, B. The principles of deep learning theory. *Preprint arXiv:2106.10165*, 2021.

Shalev-Shwartz, S., Shamir, O., and Shammah, S. Failures of gradient-based deep learning. In *International Conference on Machine Learning (ICML)*, 2017.

Smith, L. N. and Topin, N. Super-convergence: Very fast training of residual networks using large learning rates. *Preprint arXiv 1708.07120*, 2018.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

Smith, S. L., Kindermans, P.-J., and Le, Q. V. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations (ICLR)*, 2018.

Smith, S. L., Dherin, B., Barrett, D. G., and De, S. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

Sohl-Dickstein, J., Novak, R., Schoenholz, S. S., and Lee, J. On the infinite width limit of neural networks with a standard parameterization. *Preprint arXiv:2001.07301*, 2020.

Wang, Y., Chen, M., Zhao, T., and Tao, M. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations (ICLR)*, 2022.

Yang, G. and Hu, E. J. Tensor programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Yao, Z., Gholami, A., Keutzer, K., and Mahoney, M. W. Py-Hessian: Neural networks through the lens of the Hessian. In *IEEE International Conference on Big Data*, 2020.

A. Performance of Fully-Connected Networks Trained with Maximal Initial Learning Rate
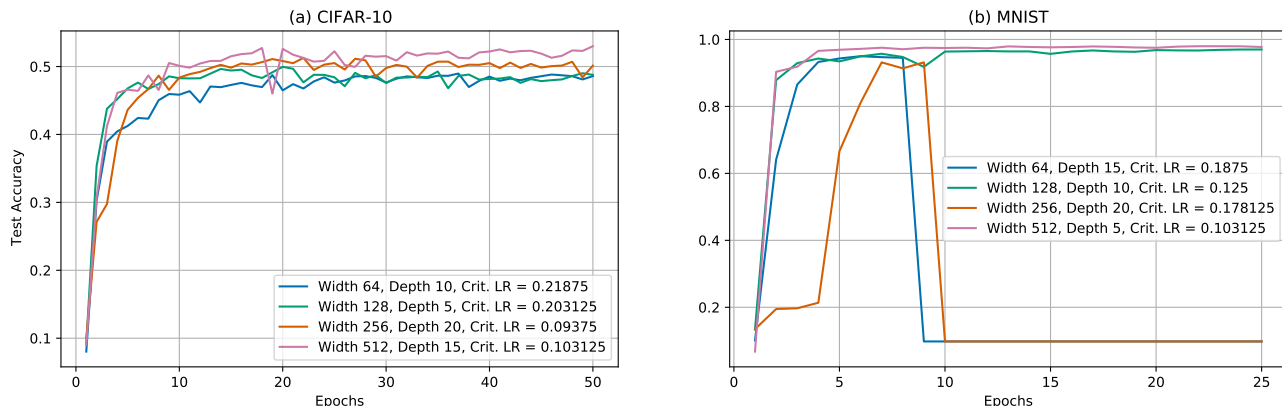

Figure 7: Performance of fully-connected networks when trained at the maximal initial learning rate η^*. (a) Validation performance of initializations with different fully-connected network architectures when trained at learning rate $\eta = \eta^*$ computed by Algorithm 1 with $t = 0.34$ on CIFAR-10, over 50 training epochs. All initializations achieve $\approx 50\%$ validation accuracy, which is reasonable for fully-connected networks; (b) The same for MNIST with $t = 0.925$, over 25 training epochs. Note that training networks with η^* guarantees that the network reaches the given threshold accuracy, but not long-term training stability. Refer to §3.2 for specifics of the experimental setup.

Based on Figure 7, we make the following observations about the maximal initial learning rate η^*:

1. Networks train reasonably well when trained at η^* – the fully connected networks we consider achieve $\approx 50\%$ validation accuracy on CIFAR-10. Although the computed η^* may not be small enough to achieve optimal performance when held constant, we believe they can serve well as large initial learning rates which can later be decayed for further improvement in performance.

2. However, it must be noted that training at $\eta = \eta^*$ only guarantees that threshold performance will be achieved, and not long-term training stability. This is particularly easy to notice on "easy" datasets, such as MNIST. Fortunately, this can be easily overcome by simply employing early stopping, or a learning rate decay scheme.

B. Sharpness at Initialization for Different Initialization Schemes

Figure 8 compares λ_1 at initialization between the Kaiming and LeCun initialization schemes, as a function of changing fully-connected network architecture. Based on this figure, we make the following observations:

- For Kaiming-initialized networks, λ_1 scales with both width and depth, and its variance primarily scales with the depth of the network architecture.

- LeCun-initialized networks show very little variation in $|\lambda_1|$ with a change in width, but considerably more with depth. Furthermore, $|\lambda_1|$ decreases exponentially as the depth gets larger, which is opposite to the trend we noticed for Kaiming initializations. It is also interesting to note that λ_1 may be positive or negative for LeCun-initialized networks, whereas Kaiming-initialized networks show only positive λ_1 values.

The above observations illustrate that the way in which λ_1 varies and scales with architecture is largely dependent on the initialization scheme employed. While we have the means to estimate the top eigenvalue(s) of the training Hessian, we do not yet *exactly* understand how it is impacted by network architecture, data, and initialization. This is partly because the complete Hessian is difficult to compute and theoretically analyze in large-scale settings. We believe that a better understanding of this quantity could help us understand the role that sharpness plays in the optimization of neural networks.
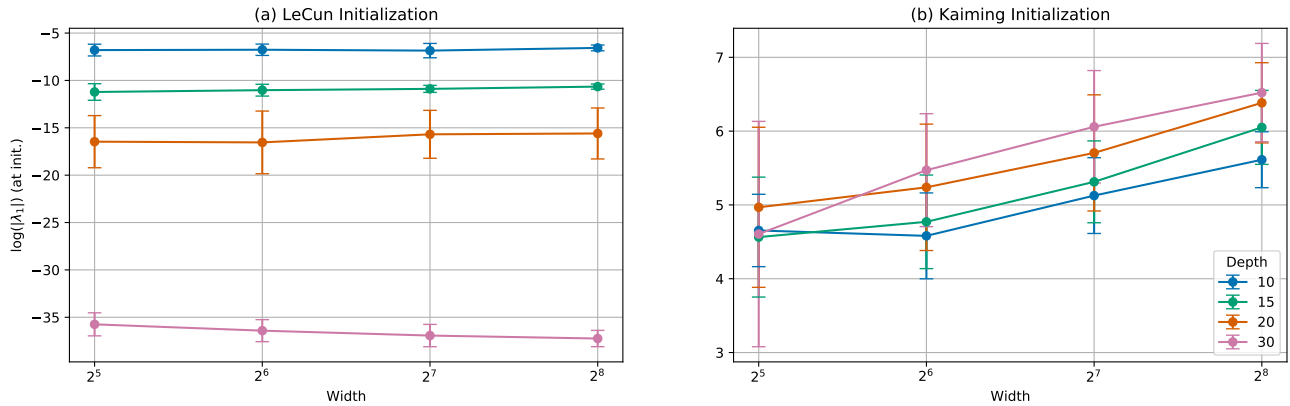
Figure 8: Visualization of sharpness λ_1 as a function of width and depth at initialization for (a) LeCun (LeCun et al., 2012) and (b) Kaiming (He et al., 2015) initialization schemes. We take absolute values and log scaling for λ_1 in (b) for the sake of clear representation - λ_1 values are extremely small in magnitude and can be positive or negative in sign. Width is given on the x-axis, and the different colors indicate different depths. There is a clear difference in how λ_1 scales for the considered initialization schemes – while it becomes larger as width and depth increase for Kaiming-initialized networks, it becomes smaller with depth for LeCun-initialized networks.

C. Maximal Initial Learning Rates with Standard Training Setup

When all layers of the network are trained at the same learning rate η^*, the trend observed in Figure 1 breaks, and the relationship becomes non-linear at small (depth $\times$ width) values, especially for networks with small width/depth. It is also worth noting that the linear relationship is preserved for much larger values of (depth $\times$ width).

This raises a few questions: What exactly is the influence of the input layer on η^*? What is the correct way to initialize it so we see a linear trend between η^* and (depth $\times$ width)?

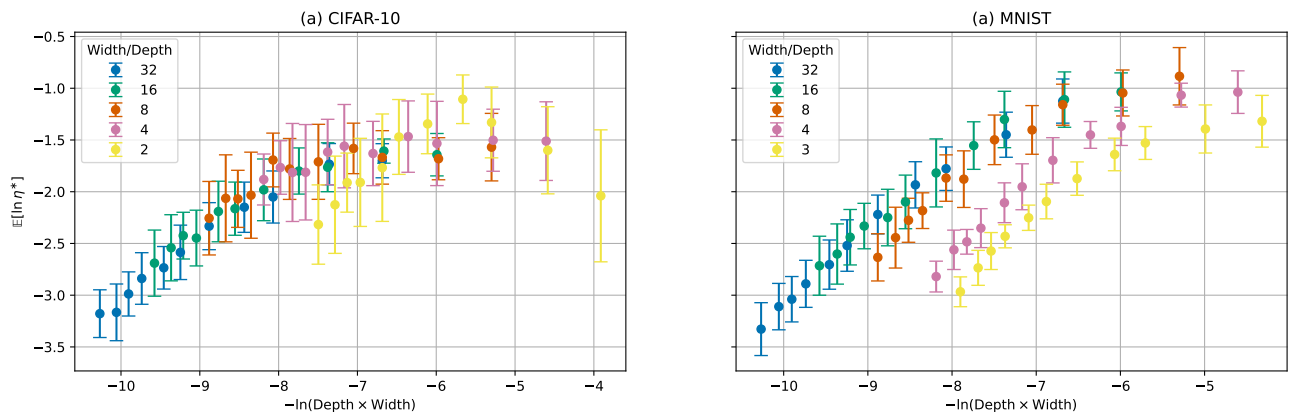Figure 9: Relationship between the maximal initial learning rate η^* and architecture for fully-connected networks with different width/depth values, trained on CIFAR-10 and MNIST. We use the same network architectures as in Figure 1. We observe a consistent power relationship for networks with relatively large widths and small depths. However, this soon becomes non-linear for other, relatively deeper architectures.
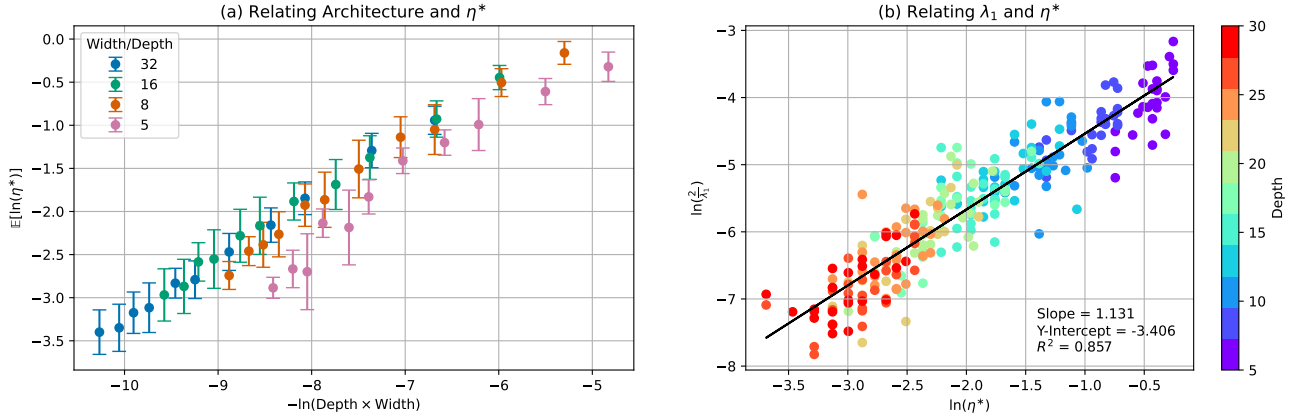
13

D. Results on Fashion-MNIST

Figure 10: Experimental results for Fashion-MNIST. We obtain a threshold accuracy of 0.84 for Fashion-MNIST. The experimental setup remains identical to those in previous experiments, and the results further confirm the empirical and theoretical results obtained in this work.
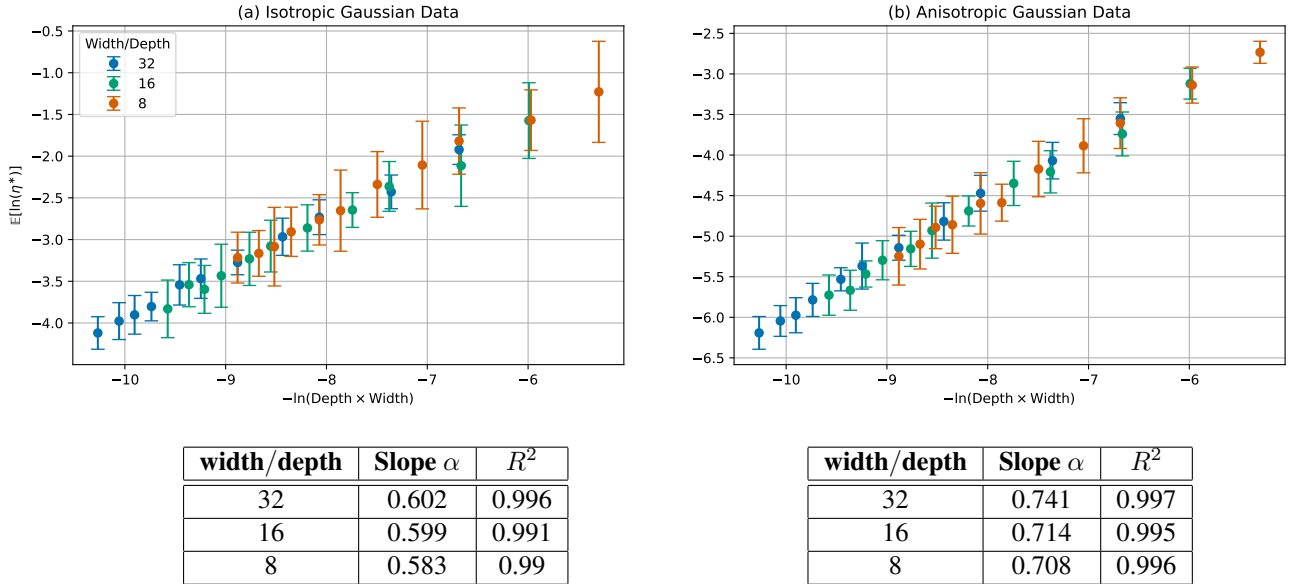
E. Results on Gaussian Data

width/depth	Slope α	R^2
32	0.602	0.996
16	0.599	0.991
8	0.583	0.99

width/depth	Slope α	R^2
32	0.741	0.997
16	0.714	0.995
8	0.708	0.996

Figure 11: Relationship between maximal initial learning rate η^* and architecture for (a) isotropic and (b) anisotropic Gaussian datasets. We use the same values for depth as in Figure 1. Data is sampled from two multivariate normal distributions (i.e. binary classification). The training set and validation set respectively consist of 9k and 1k samples from each distribution, leading to a total of 20k samples (with 18k samples in the training set, and 2k in the validation set). Each sample is 100-dimensional, and the means are sampled from a standard normal distribution. For the anisotropic Gaussian dataset, we sample 100-dimensional covariance matrices from a standard normal distribution as well. For the isotropic and anisotropic Gaussian datasets, we obtain threshold accuracies of 1.0 and 0.81 respectively.

We note that for the isotropic Gaussian data, the slope values are significantly smaller than those observed in other experiments. This emphasizes that even in the relatively simple case of fully connected ReLU networks, we do not have a theoretical explanation of the empirical scaling laws for the maximal initial learning rate as a function of architecture. Through these experiments, we hope to understand these simple situations before analyzing more complex cases.

14

F. Comparison of Averaged Sharpness λ_1 and Maximal Initial Learning Rate η^*
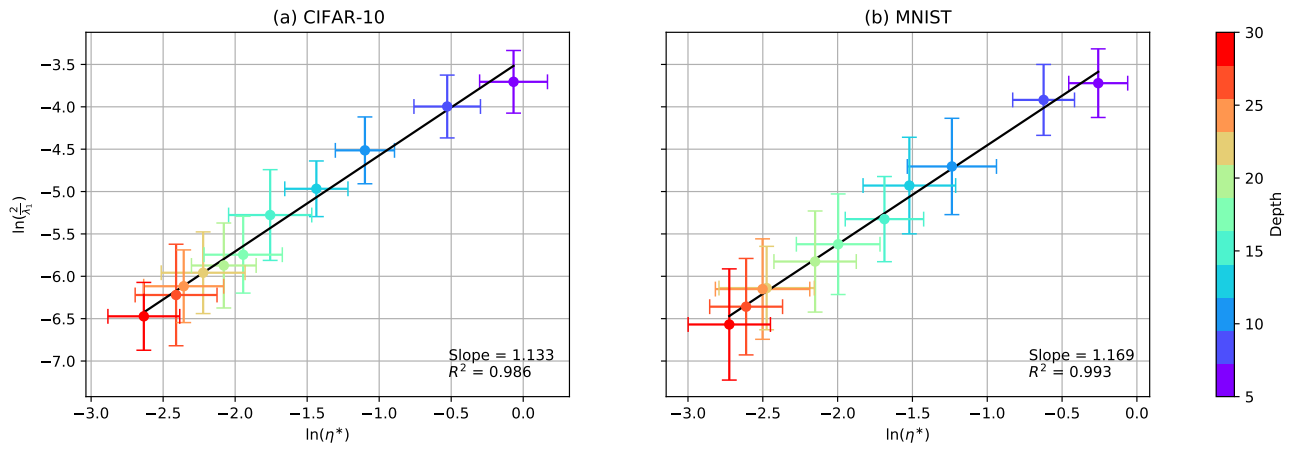
Figure 12: Correlation between $2/\lambda_1$ and η^*, averaged over 25 initializations per architecture. Each architecture has a sufficiently large width/depth = 16, to preserve the established power relationship.

G. Performance of ResNet-20 Networks Trained with Maximal Initial Learning Rate

Figure 13: Performance of ResNet-20 (He et al., 2016) networks with different learning rate setups. Each line in the figure is an average of 3 runs, along with error bars to indicate deviation in performance. A well-tuned, constant learning rate consistently beats η^*, but performance is competitive when using a scheduler is employed. Refer to MosaicML's Model Card for details of the learning rate scheduler setup.

H. Proof of Theorem 1

H.1. Setup and Preparatory Lemmas

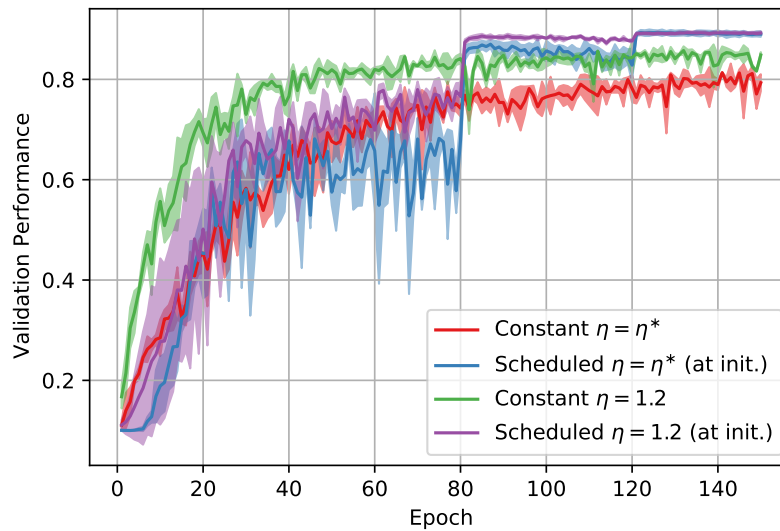Let us first recall the notation. We consider a ReLU network, which for an input $x \in \mathbb{R}^{n_0}$ computes $z_1^{(L+1)} \in \mathbb{R}$ via intermediate representations $z^{(\ell)} \in \mathbb{R}^{n_\ell}$

$$z_i^{(\ell+1)} = \begin{cases} \sum_{j=1}^{n_\ell} W_{ij}^{(\ell)} \sigma\left(z_j^{(\ell)}\right), & \ell \geq 1 \\ \sum_{j=1}^{n_\ell} W_{ij}^{(1)} x_j, & \ell = 0 \end{cases}, \qquad i = 1, \ldots, n_\ell.$$

Moreover, we will assume that the weights are independent Gaussians

$$W_{ij}^{(\ell)} \sim \mathcal{N}\left(0, \frac{2}{n_{\ell-1}}\right) \quad \text{independent.}$$

Instead of simply considering the loss Hessian as in the statement of Theorem 1, we will study a slightly more general effective Hessian

$$H_{\text{eff}} := \left(\widehat{\eta}_\mu \widehat{\eta}_\nu \partial_{\mu\nu} \left\{\frac{1}{2}\left(z_1^{(L+1)} - y\right)^2\right\}\right)_{\mu,\nu},$$

where μ, ν run over all network weights and for any weight $\mu = W_{ij}^{(\ell)}$ we write

$$\widehat{\eta}_{W_{ij}^{(\ell)}} = n_{\ell-1}^{-1/2} \eta^{(\ell)}$$

for the corresponding learning rates. We've introduced the rescaled learning rates $\eta^{(\ell)}$ for weights in layer ℓ for notational convenience in what follows. Our goal is to compute the mean of the Hilbert-Schmidt norm

$$\mathbb{E}\left[||H_{\text{eff}}||_{HS}^2\right] = \mathbb{E}\left[\sum_{\mu,\nu \leq L+1} \left(\widehat{\eta}_\mu \widehat{\eta}_\nu \partial_\mu \partial_\nu \left\{\frac{1}{2}\left(z_1^{(L+1)} - y\right)^2\right\}\right)^2\right], \tag{1}$$

where we remind the reader that for any ℓ the notation $\mu \leq \ell$ means the set of all weights in layers $1, \ldots, \ell$. In order to effectively evaluate equation 1, we need two preparatory results. The first is well-known and can be found in Theorem 3 of Hanin (2018) and Proposition 2 of Hanin & Nica (2019a)

Lemma 1. The indicator random variables $\mathbf{1}_{\left\{z_i^{(\ell)} > 0\right\}}$ are independent of any even function of the network weights and of each other. Their marginal distribution is Bernoulli $1/2$.

The second result we need is a simple corollary of Lemma 1. To state it, we need some notation. For any $\ell = 1, \ldots, L$ and any expressions $f_k(z)$ depending on z and $\partial_\mu z$ we write

$$Y^{(\ell)}[f_1, \ldots, f_k] := \mathbb{E}\left[\sum_{\mu \leq \ell} (\widehat{\eta}_\mu)^2 \frac{1}{n_\ell^k} \sum_{j_1, \ldots, j_k = 1}^{n_\ell} f_1(z_{j_1}^{(\ell)}) \cdots f_k(z_{j_k}^{(\ell)})\right].$$

Thus, for example

$$Y^{(\ell)}[z\partial_\mu z] = \mathbb{E}\left[\sum_{\mu \leq \ell} (\widehat{\eta}_\mu)^2 \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} z_j^{(\ell)} \partial_\mu z_j^{(\ell)}\right].$$

Similarly, if the functions $f_k(z)$ depend in addition on $\partial_\nu z$ and $\partial_{\mu\nu} z$ then we we will write

$$Y^{(\ell)}[f_1, \ldots, f_k] := \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{1}{n_\ell^k} \sum_{j_1, \ldots, j_k = 1}^{n_\ell} f_1(z_{j_1}^{(\ell)}) \cdots f_k(z_{j_k}^{(\ell)})\right].$$

Thus, for example

$$Y^{(\ell)}\left[(\partial_\mu z)^2, z\partial_{\mu\nu} z\right] = \mathbb{E}\left[\sum_{\mu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{1}{n_\ell^2} \sum_{j_1, j_2 = 1}^{n_\ell} \left(\partial_\mu z_{j_1}^{(\ell)}\right)^2 z_{j_2}^{(\ell)} \partial_{\mu\nu} z_{j_2}^{(\ell)}\right].$$

We will use repeatedly the following Corollary of Lemma 1:

Corollary 3. Fix $k \geq 1$ and suppose that

$$f_j(z) = \sigma(z)^{a_j} \left(\partial_\mu \sigma(z)\right)^{b_j} \left(\partial_\nu \sigma(z)\right)^{c_j} \left(\partial_{\mu\nu} \sigma(z)\right)^{d_j}, \qquad j = 1, \ldots, k$$

with $a_j + b_j + c_j + d_j$ being even for every j. Write

$$\widehat{f}_j(z) := z^{a_j} \left(\partial_\mu z\right)^{b_j} \left(\partial_\nu z\right)^{c_j} \left(\partial_{\mu\nu} z\right)^{d_j}, \qquad j = 1, \ldots, k.$$

Then,

$$Y^{(\ell)}[f_1] = \frac{1}{2} Y^{(\ell)}\left[\widehat{f}_1\right], \qquad \widehat{f}_1(z) := z^a \left(\partial_\mu z\right)^b. \tag{2}$$

Further,

$$Y^{(\ell)}[f_1, f_2] = \frac{1}{4}\left[Y^{(\ell)}\left[\widehat{f}_1, \widehat{f}_2\right] + \frac{1}{n_\ell} Y^{(\ell)}\left[\widehat{f}_1 \cdot \widehat{f}_2\right]\right].$$

Proof. When $k = 1$, we have

$$Y^{(\ell)}[f_1] = \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \left(\sigma(z_j^{(\ell)})\right)^{a_j} \left(\partial_\mu \sigma(z_j^{(\ell)})\right)^{b_j} \left(\partial_\nu \sigma(z_j^{(\ell)})\right)^{c_j} \left(\partial_{\mu\nu} \sigma(z_j^{(\ell)})\right)^{d_j}\right]$$

$$= \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \left(z_j^{(\ell)}\right)^{a_j} \left(\partial_\mu z_j^{(\ell)}\right)^{b_j} \left(\partial_\nu z_j^{(\ell)}\right)^{c_j} \left(\partial_{\mu\nu} z_j^{(\ell)}\right)^{d_j} \mathbf{1}_{\left\{z_j^{(\ell)} \geq 0\right\}}\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \left(z_j^{(\ell)}\right)^{a_j} \left(\partial_\mu z_j^{(\ell)}\right)^{b_j} \left(\partial_\nu z_j^{(\ell)}\right)^{c_j} \left(\partial_{\mu\nu} z_j^{(\ell)}\right)^{d_j} \mathbf{1}_{\left\{z_j^{(\ell)} \geq 0\right\}} \,\Big|\, z^{(\ell-1)}\right]\right].$$

In the inner conditional expectation, the term $\left(z_j^{(\ell)}\right)^{a_j} \left(\partial_\mu z_j^{(\ell)}\right)^{b_j} \left(\partial_\nu z_j^{(\ell)}\right)^{c_j} \left(\partial_{\mu\nu} z_j^{(\ell)}\right)^{d_j} \mathbf{1}_{\left\{z_j^{(\ell)} \geq 0\right\}}$ is an even function of the weights in layer ℓ. Hence, by Lemma 1, it is independent of the indicator function. This yields equation 2. Similarly, we have

$$Y^{(\ell)}[f_1, f_2] = \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{1}{n_\ell^2} \sum_{k_1,k_2=1}^{n_\ell} \prod_{j=1}^{2} \left(\sigma(z_{k_j}^{(\ell)})\right)^{a_j} \left(\partial_\mu \sigma(z_{k_j}^{(\ell)})\right)^{b_j} \left(\partial_\nu \sigma(z_{k_j}^{(\ell)})\right)^{c_j} \left(\partial_{\mu\nu} \sigma(z_{k_j}^{(\ell)})\right)^{d_j}\right]$$

$$= \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{1}{n_\ell^2} \sum_{k_1,k_2=1}^{n_\ell} \prod_{j=1}^{2} \left(z_{k_j}^{(\ell)}\right)^{a_j} \left(\partial_\mu z_{k_j}^{(\ell)}\right)^{b_j} \left(\partial_\nu z_{k_j}^{(\ell)}\right)^{c_j} \left(\partial_{\mu\nu} z_{k_j}^{(\ell)}\right)^{d_j} \mathbf{1}_{\left\{z_{k_1}^{(\ell)} \geq 0\right\}} \mathbf{1}_{\left\{z_{k_2}^{(\ell)} \geq 0\right\}}\right]$$

$$= \frac{1}{n_\ell} \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(z_1^{(\ell)}\right)^{a_1+a_2} \left(\partial_\mu z_1^{(\ell)}\right)^{b_1+b_2} \left(\partial_\nu z_1^{(\ell)}\right)^{c_1+c_2} \left(\partial_{\mu\nu} z_1^{(\ell)}\right)^{d_1+d_2} \mathbf{1}_{\left\{z_{j_1}^{(\ell)} \geq 0\right\}}\right]$$

$$+ \left(1 - \frac{1}{n_\ell}\right) \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \prod_{j=1}^{2} \left(z_j^{(\ell)}\right)^{a_j} \left(\partial_\mu z_j^{(\ell)}\right)^{b_j} \left(\partial_\nu z_j^{(\ell)}\right)^{c_j} \left(\partial_{\mu\nu} z_j^{(\ell)}\right)^{d_j} \mathbf{1}_{\left\{z_j^{(\ell)} \geq 0\right\}}\right],$$

where the last equality follows by symmetry. Again conditioning on $z^{(\ell-1)}$ we thus find

$$Y^{(\ell)}[f_1, f_2] = \frac{1}{2n_\ell} \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(z_1^{(\ell)}\right)^{a_1+a_2} \left(\partial_\mu z_1^{(\ell)}\right)^{b_1+b_2} \left(\partial_\nu z_1^{(\ell)}\right)^{c_1+c_2} \left(\partial_{\mu\nu} z_1^{(\ell)}\right)^{d_1+d_2}\right]$$

$$+ \frac{1}{4}\left(1 - \frac{1}{n_\ell}\right) \mathbb{E}\left[\sum_{\mu,\nu \leq \ell} (\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \prod_{j=1}^{2} \left(z_j^{(\ell)}\right)^{a_j} \left(\partial_\mu z_j^{(\ell)}\right)^{b_j} \left(\partial_\nu z_j^{(\ell)}\right)^{c_j} \left(\partial_{\mu\nu} z_j^{(\ell)}\right)^{d_j}\right].$$

17

Running the above symmetry argument in reverse yields

$$Y^{(\ell)}[f_1, f_2] = \frac{1}{4}\left[Y^{(\ell)}\left[\widehat{f}_1, \widehat{f}_2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[\widehat{f}_1 \cdot \widehat{f}_2\right]\right],$$

as claimed. □

In what follows we will use Lemma 1 and Corollary 3 without mention.

H.2. Reducing $\mathbb{E}\left[||H_{eff}||_F^2\right]$ to $Y^{(\ell)}$'s

To make progress on evaluating the expression equation 1, let us first write

$$\partial_\mu\partial_\nu\left\{\frac{1}{2}\left(z_1^{(L+1)} - y\right)^2\right\} = \partial_\mu\left(\partial_\nu z_1^{(L+1)}\left(z_1^{(L+1)} - y\right)\right) = \partial_{\mu\nu}z_1^{(L+1)}\left(z_1^{(L+1)} - y\right) + \partial_\mu z_1^{(L+1)}\partial_\nu z_1^{(L+1)}.$$

Hence,

$$\left(\partial_\mu\partial_\nu\left\{\frac{1}{2}\left(z_1^{(L+1)} - y\right)^2\right\}\right)^2 = \left(\partial_{\mu\nu}z_1^{(L+1)}\left(z_1^{(L+1)} - y\right)\right)^2 + 2\partial_{\mu\nu}z_1^{(L+1)}\partial_\mu z_1^{(L+1)}\partial_\nu z_1^{(L+1)}\left(z_1^{(L+1)} - y\right)$$
$$+ \left(\partial_\mu z_1^{(L+1)}\partial_\nu z_1^{(L+1)}\right)^2.$$

Using that y has mean 0 and variance 1 as well as the fact that any term with an odd number of $z_1^{(L+1)}$'s has zero mean shows that

$$\mathbb{E}\left[\left(\partial_\mu\partial_\nu\left\{\frac{1}{2}\left(z_1^{(L+1)} - y\right)^2\right\}\right)^2\right] = \mathbb{E}\left[\left(\partial_{\mu\nu}z_1^{(L+1)}z_1^{(L+1)}\right)^2\right] + \mathbb{E}\left[\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right] \tag{3}$$
$$+ 2\mathbb{E}\left[\partial_{\mu\nu}z_1^{(L+1)}\partial_\mu z_1^{(L+1)}\partial_\nu z_1^{(L+1)}z_1^{(L+1)}\right] + \mathbb{E}\left[\left(\partial_\mu z_1^{(L+1)}\partial_\nu z_1^{(L+1)}\right)^2\right].$$

Our goal is to evaluate the sums of such terms over μ, ν recursively in L. We do this by first integrating out the weights in the last layer to reduce computing the expected squared Hilbert-Schmidt norm of the loss hessian to various Y's.

Lemma 2. We have

$$
\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right]
$$
$$
= \left(\eta^{(L+1)} \right)^2 \left[Y^{(L)} \left[(\partial_\mu z)^2, z^2 \right] + \frac{1}{n_L} Y^{(L)} \left[(z\partial_\mu z)^2 \right] \right] + Y^{(L)} \left[(\partial_{\mu\nu} z)^2, z^2 \right] + \frac{1}{n_L} Y^{(L)} \left[(z\partial_{\mu\nu} z)^2 \right]
$$
$$
+ 2Y^{(L)} \left[z\partial_{\mu\nu} z, z\partial_{\mu\nu} z \right] + \frac{2}{n_L} Y^{(L)} \left[(z\partial_{\mu\nu} z)^2 \right] \tag{4}
$$

$$
\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} \right)^2 \right]
$$
$$
= \left(\eta^{(L+1)} \right)^2 Y^{(L)} \left[(\partial_\mu z)^2 \right] + Y^{(L)} \left[(\partial_{\mu\nu} z)^2 \right] \tag{5}
$$

$$
\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \, \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)} \right]
$$
$$
= \left(\eta^{(L+1)} \right)^2 \left(Y^{(L)} \left[z\partial_\mu z, z\partial_\mu z \right] + \frac{1}{n_L} Y^{(L)} \left[(z\partial_\mu z)^2 \right] \right)
$$
$$
+ 2Y^{(L)} \left[\partial_{\mu\nu} z \partial_\mu z, z\partial_\nu z \right] + Y^{(L)} \left[z\partial_{\mu\nu} z, \partial_\mu z \partial_\nu z \right] + \frac{3}{n_L} Y^{(L)} \left[z\partial_{\mu\nu} z \partial_\mu z \partial_\nu z \right] \tag{6}
$$

$$
\mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} \right)^2 \right]
$$
$$
= \frac{1}{2} \left(\eta^{(L+1)} \right)^4 \left(\mathbb{E}\left[\left(\frac{1}{n_L} \left\| z^{(L)} \right\|_2^2 \right)^2 \right] + \frac{1}{n_L} \mathbb{E}\left[\frac{1}{n_L} \left\| z^{(L)} \right\|_4^4 \right] \right)
$$
$$
+ \left(\eta^{(L+1)} \right)^2 \left[Y^{(L)} \left[z^2, (\partial_\mu z)^2 \right] + \frac{1}{n_L} Y^{(L)} \left[(z\partial_\mu z)^2 \right] \right]
$$
$$
+ 2Y^{(L)} \left[\partial_\mu z \partial_\nu z, \partial_\mu z \partial_\nu z \right] + Y^{(L)} \left[(\partial_\mu z)^2, (\partial_\nu z)^2 \right] + \frac{3}{n_L} Y^{(L)} \left[(\partial_\mu z \partial_\nu z)^2 \right] \tag{7}
$$

Proof. We begin with deriving (equation 4). We have

$$
\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right] = \sum_{\mu,\nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right]
$$
$$
+ \sum_{\mu \leq L, \, \nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right]
$$
$$
+ \sum_{\mu \in L+1, \, \nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right]
$$
$$
+ \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right].
$$

Note first that if μ, ν are both weights in layer $L + 1$, then $\partial_{\mu\nu} z_1^{(L+1)} = 0$. Thus, the first sum vanishes. Next, the second

and third sums are equal. To evaluate them we proceed as follows:

$$
\sum_{\mu \leq L, \, \nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right] = \sum_{\mu \leq L} (\widehat{\eta}_\mu)^2 \, \mathbb{E}\left[\frac{\left(\eta^{(L+1)} \right)^2}{n_L} \sum_{j_1=1}^{n_L} \left(\partial_\mu \sigma(z_{j_1}^{(L)}) z_1^{(L+1)} \right)^2 \right]
$$

$$
= \sum_{\mu \leq L} (\widehat{\eta}_\mu)^2 \, \mathbb{E}\left[\frac{\left(\eta^{(L+1)} \right)^2}{n_L} \sum_{j_1=1}^{n_L} \left(\partial_\mu \sigma(z_{j_1}^{(L)}) \sum_{j_2=1}^{n_L} W_{1 j_2}^{(L+1)} \sigma(z_{j_2}^{(L)}) \right)^2 \right]
$$

$$
= \sum_{\mu \leq L} (\widehat{\eta}_\mu)^2 \, \mathbb{E}\left[\frac{2 \left(\eta^{(L+1)} \right)^2}{n_L^2} \sum_{j_1, j_2=1}^{n_L} \left(\partial_\mu \sigma(z_{j_1}^{(L)}) \sigma(z_{j_2}^{(L)}) \right)^2 \right]
$$

$$
= \sum_{\mu \leq L} (\widehat{\eta}_\mu)^2 \, 2 \left(\eta^{(L+1)} \right)^2 \mathbb{E}\left[\frac{1}{n_L} \left(\partial_\mu \sigma(z_1^{(L)}) \sigma(z_1^{(L)}) \right)^2 \right]
$$

$$
+ \sum_{\mu \leq L} (\widehat{\eta}_\mu)^2 \, 2 \left(\eta^{(L+1)} \right)^2 \mathbb{E}\left[\left(1 - \frac{1}{n_L} \right) \left(\partial_\mu \sigma(z_1^{(L)}) \sigma(z_2^{(L)}) \right)^2 \right]
$$

$$
= \sum_{\mu \leq L} (\widehat{\eta}_\mu)^2 \left(\eta^{(L+1)} \right)^2 \mathbb{E}\left[\frac{1}{n_L} \left(\partial_\mu z_1^{(L)} z_1^{(L)} \right)^2 \right]
$$

$$
+ \frac{1}{2} \sum_{\mu \leq L} (\widehat{\eta}_\mu)^2 \left(\eta^{(L+1)} \right)^2 \mathbb{E}\left[\left(1 - \frac{1}{n_L} \right) \left(\partial_\mu z_1^{(L)} z_2^{(L)} \right)^2 \right]
$$

$$
= \frac{\left(\eta^{(L+1)} \right)^2}{2} \left[Y^{(L)} \left[(\partial_\mu z)^2, z^2 \right] + \frac{1}{n_L} Y^{(L)} \left[(z \partial_\mu z)^2 \right] \right].
$$

Finally, the fourth sum is

$$
\sum_{\mu, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right] = \sum_{\mu, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\sum_{j_1, j_2=1}^{N_L} W_{1 j_1}^{(L+1)} W_{1 j_2}^{(L+1)} \sigma(z_{j_1}^{(L)}) \partial_{\mu\nu} \sigma(z_{j_2}^{(L)}) \right)^2 \right]
$$

$$
= \sum_{\mu, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L^2} \sum_{j_1, j_2=1}^{N_L} \left(\sigma(z_{j_1}^{(L)}) \partial_{\mu\nu} \sigma(z_{j_2}^{(L)}) \right)^2 \right]
$$

$$
+ 2 \sum_{\mu, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L^2} \sum_{j_1, j_2=1}^{N_L} \sigma(z_{j_1}^{(L)}) \partial_{\mu\nu} \sigma(z_{j_1}^{(L)}) \sigma(z_{j_2}^{(L)}) \partial_{\mu\nu} \sigma(z_{j_2}^{(L)}) \right].
$$

To proceed we evaluate the first term as follows:

$$
\sum_{\mu, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L^2} \sum_{j_1, j_2=1}^{N_L} \left(\sigma(z_{j_1}^{(L)}) \partial_{\mu\nu} \sigma(z_{j_2}^{(L)}) \right)^2 \right] = \sum_{\mu, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L} \left(\sigma(z_1^{(L)}) \partial_{\mu\nu} \sigma(z_1^{(L)}) \right)^2 \right]
$$

$$
+ \sum_{\mu, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \, 4 \left(1 - \frac{1}{n_L} \right) \left(\sigma(z_1^{(L)}) \partial_{\mu\nu} \sigma(z_2^{(L)}) \right)^2 \right]
$$

$$
= Y^{(L)} \left[(\partial_{\mu\nu} z)^2, z^2 \right] + \frac{1}{n_L} Y^{(L)} \left[(z \partial_{\mu\nu} z)^2 \right].
$$

Further, the second term is

$$\sum_{\mu,\nu\leq L} \mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2 \frac{4}{n_L^2}\sum_{j_1,j_2=1}^{N_L} \sigma(z_{j_1}^{(L)})\partial_{\mu\nu}\sigma(z_{j_1}^{(L)})\sigma(z_{j_2}^{(L)})\partial_{\mu\nu}\sigma(z_{j_2}^{(L)})\right]$$

$$= \sum_{\mu,\nu\leq L} \mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2 \frac{4}{n_L}\left(\sigma(z_1^{(L)})\partial_{\mu\nu}\sigma(z_1^{(L)})\right)^2\right]$$

$$+ \sum_{\mu,\nu\leq L}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\, 4\left(1-\frac{1}{n_L}\right)\sigma(z_1^{(L)})\partial_{\mu\nu}\sigma(z_1^{(L)})\sigma(z_2^{(L)})\partial_{\mu\nu}\sigma(z_2^{(L)})\right]$$

$$= Y^{(L)}\left[z\partial_{\mu\nu}z, z\partial_{\mu\nu}z\right] + \frac{1}{n_L}Y^{(L)}\left[(z\partial_{\mu\nu}z)^2\right]$$

Putting all this together yields

$$\sum_{\mu,\nu\leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}z_1^{(L+1)}\right)^2\right] = \left(\eta^{(L+1)}\right)^2\left[Y^{(L)}\left[(\partial_\mu z)^2, z^2\right] + \frac{1}{n_L}Y^{(L)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ Y^{(L)}\left[(\partial_{\mu\nu}z)^2, z^2\right] + \frac{1}{n_L}Y^{(L)}\left[(z\partial_{\mu\nu}z)^2\right]$$

$$+ 2Y^{(L)}\left[z\partial_{\mu\nu}z, z\partial_{\mu\nu}z\right] + \frac{2}{n_L}Y^{(L)}\left[(z\partial_{\mu\nu}z)^2\right],$$

which is precisely the statement of equation 4. Next, we establish equation 5 in a similar manner. We have

$$\sum_{\mu,\nu\leq L+1}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right] = \sum_{\mu,\nu\in L+1}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right]$$

$$+ \sum_{\mu\leq L,\nu\in L+1}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right]$$

$$+ \sum_{\mu\in L+1,\nu\leq L}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right]$$

$$+ \sum_{\mu,\nu\leq L}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right].$$

Again the first sum vanishes since $\partial_{\mu\nu}z_1^{(L+1)} = 0$ if μ,ν are weights in the final layer. Next, the second and third terms are equal and can be written as follows:

$$\sum_{\mu\leq L,\nu\in L+1}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right] = \left(\eta^{(L+1)}\right)^2\sum_{\mu\leq L}\frac{1}{n_L}\sum_{j=1}^{n_L}\mathbb{E}\left[(\widehat{\eta}_\mu)^2\left(\partial_\mu\sigma(z_j^{(L)})\right)^2\right]$$

$$= \frac{1}{2}\left(\eta^{(L+1)}\right)^2 Y^{(L)}\left[(\partial_\mu z)^2\right].$$

Finally, the fourth term in the sum can be rewritten in the following manner

$$\sum_{\mu,\nu\leq L}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\partial_{\mu\nu}z_1^{(L+1)}\right)^2\right] = \sum_{\mu,\nu\leq L}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\left(\sum_{j=1}^{n_L}W_{1j}^{(L+1)}\partial_{\mu\nu}\sigma(z_j^{(L)})\right)^2\right]$$

$$= \sum_{\mu,\nu\leq L}\mathbb{E}\left[(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{2}{n_L}\sum_{j=1}^{n_L}\left(\partial_{\mu\nu}\sigma(z_j^{(L)})\right)^2\right]$$

$$= Y^{(L)}\left[(\partial_{\mu\nu}z)^2\right].$$

21

Hence, altogether, we find

$$\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)}\right)^2\right] = \left(\eta^{(L+1)}\right)^2 Y^{(L)}\left[(\partial_\mu z)^2\right] + Y^{(L)}\left[(\partial_{\mu\nu} z)^2\right],$$

which is the statement of equation 5. Next, we establish equation 6. We have

$$\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)}\right]$$

$$= \sum_{\mu,\nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)}\right]$$

$$+ \sum_{\mu \leq L, \nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)}\right]$$

$$+ \sum_{\mu \in L+1, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)}\right]$$

$$+ \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)}\right].$$

The first sum vanishes since $\partial_{\mu\nu} z_1^{(L+1)} = 0$ when μ, ν are weights in the final layer. The second and third terms are again the same and equal

$$\sum_{\mu \leq L, \nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)}\right]$$

$$= \left(\eta^{(L+1)}\right)^2 \sum_{\mu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu)^2 \frac{1}{n_L} \sum_{j_1=1}^{n_L} \partial_\mu \sigma(z_{j_1}^{(L)}) \partial_\mu z_1^{(L+1)} \sigma(z_{j_1}^{(L)}) z_1^{(L+1)}\right]$$

$$= \left(\eta^{(L+1)}\right)^2 \sum_{\mu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu)^2 \frac{1}{n_L} \sum_{j_1=1}^{n_L} \partial_\mu \sigma(z_{j_1}^{(L)}) \sigma(z_{j_1}^{(L)}) \sum_{j_2=1}^{n_L} W_{1j_2}^{(L+1)} W_{1j_2}^{(L+1)} \partial_\mu \sigma(z_{j_2}^{(L)}) \sigma(z_{j_2}^{(L)})\right]$$

$$= \left(\eta^{(L+1)}\right)^2 \sum_{\mu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu)^2 \frac{2}{n_L^2} \sum_{j_1,j_2=1}^{n_L} \partial_\mu \sigma(z_{j_1}^{(L)}) \sigma(z_{j_1}^{(L)}) \partial_\mu \sigma(z_{j_2}^{(L)}) \sigma(z_{j_2}^{(L)})\right]$$

$$= \left(\eta^{(L+1)}\right)^2 \sum_{\mu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu)^2 \frac{2}{n_L} \left(\partial_\mu \sigma(z_1^{(L)}) \sigma(z_1^{(L)})\right)^2\right]$$

$$+ \left(\eta^{(L+1)}\right)^2 \sum_{\mu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu)^2 2\left(1 - \frac{1}{n_L}\right) \partial_\mu \sigma(z_1^{(L)}) \sigma(z_1^{(L)}) \partial_\mu \sigma(z_2^{(L)}) \sigma(z_2^{(L)})\right]$$

$$= \frac{1}{2}\left(\eta^{(L+1)}\right)^2 \left(Y^{(L)}\left[z\partial_\mu z, z\partial_\mu z\right] + \frac{1}{n_L} Y^{(L)}\left[(z\partial_\mu z)^2\right]\right).$$

Finally, the fourth term is

$$
\sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \, \partial_{\mu\nu} z_1^{(L+1)} \partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} z_1^{(L+1)} \right]
$$

$$
= \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \sum_{j_1,j_2,j_3,j_4=1}^{n_L} W_{1j_1}^{(L+1)} W_{1j_2}^{(L+1)} W_{1j_3}^{(L+1)} W_{1j_4}^{(L+1)} \partial_{\mu\nu}\sigma(z_{j_1}^{(L)}) \partial_\mu \sigma(z_{j_2}^{(L)}) \partial_\nu \sigma(z_{j_3}^{(L)}) \sigma(z_{j_4}^{(L)}) \right]
$$

$$
= 2 \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L^2} \sum_{j_1,j_2}^{n_L} \partial_{\mu\nu}\sigma(z_{j_1}^{(L)}) \partial_\mu \sigma(z_{j_1}^{(L)}) \partial_\nu \sigma(z_{j_2}^{(L)}) \sigma(z_{j_2}^{(L)}) \right]
$$

$$
+ \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L^2} \sum_{j_1,j_2}^{n_L} \partial_{\mu\nu}\sigma(z_{j_1}^{(L)}) \sigma(z_{j_1}^{(L)}) \partial_\mu \sigma(z_{j_2}^{(L)}) \partial_\nu \sigma(z_{j_2}^{(L)}) \right]
$$

$$
= 2 Y^{(L)} \left[\partial_{\mu\nu} z \partial_\mu z, z \partial_\nu z \right] + Y^{(L)} \left[z \partial_{\mu\nu} z, \partial_\mu z \partial_\nu z \right] + \frac{2}{n_L} Y^{(L)} \left[z \partial_{\mu\nu} z \partial_\mu z \partial_\nu z \right].
$$

Putting this all together yields

$$
\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_{\mu\nu} z_1^{(L+1)} z_1^{(L+1)} \right)^2 \right]
$$

$$
= \left(\eta^{(L+1)} \right)^2 \left(Y^{(L)} \left[z \partial_\mu z, z \partial_\mu z \right] + \frac{1}{n_L} Y^{(L)} \left[(z \partial_\mu z)^2 \right] \right)
$$

$$
+ 2 Y^{(L)} \left[\partial_{\mu\nu} z \partial_\mu z, z \partial_\nu z \right] + Y^{(L)} \left[z \partial_{\mu\nu} z, \partial_\mu z \partial_\nu z \right] + \frac{2}{n_L} Y^{(L)} \left[z \partial_{\mu\nu} z \partial_\mu z \partial_\nu z \right],
$$

which precisely the statement of equation 6. Finally, it remains to check equation 7. We have

$$
\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} \right)^2 \right] = \sum_{\mu,\nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} \right)^2 \right]
$$

$$
+ \sum_{\mu \leq L, \nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} \right)^2 \right]
$$

$$
+ \sum_{\mu \in L+1, \nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} \right)^2 \right]
$$

$$
+ \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} \right)^2 \right].
$$

The first term equals

$$
\sum_{\mu,\nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)} \right)^2 \right] = \left(\eta^{(L+1)} \right)^4 \mathbb{E}\left[\frac{1}{n_L^2} \sum_{j_1,j_2=1}^{n_L} \left(\sigma(z_{j_1}^{(L)}) \sigma(z_{j_2}^{(L)}) \right)^2 \right]
$$

$$
= \frac{1}{2} \left(\eta^{(L+1)} \right)^4 \left(\mathbb{E}\left[\left(\frac{1}{n_L} \left\| z^{(L)} \right\|_2^2 \right)^2 \right] + \frac{1}{n_L} \mathbb{E}\left[\frac{1}{n_L} \left\| z^{(L)} \right\|_4^4 \right] \right).
$$

The second and third terms are the same and equal

$$\sum_{\mu \leq L, \nu \in L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)}\right)^2\right]$$

$$= \left(\eta^{(L+1)}\right)^2 \sum_{\mu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu)^2 \frac{1}{n_L} \sum_{j_1,j_2,j_3=1}^{n_L} \left(\sigma(z_{j_1}^{(\ell)})\right)^2 W_{1,j_2}^{(L+1)} W_{1,j_3}^{(L+1)} \partial_\mu \sigma(z_{j_2}^{(L)}) \partial_\mu \sigma(z_{j_3}^{(L)})\right]$$

$$= \left(\eta^{(L+1)}\right)^2 \sum_{\mu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu)^2 \frac{2}{n_L^2} \sum_{j_1,j_2=1}^{n_L} \left(\sigma(z_{j_1}^{(\ell)}) \partial_\mu \sigma(z_{j_2}^{(L)})\right)^2\right]$$

$$= \left(\eta^{(L+1)}\right)^2 \left[Y^{(L)}\left[z^2, (\partial_\mu z)^2\right] + \frac{1}{n_L} Y^{(L)}\left[(z\partial_\mu z)^2\right]\right]$$

The fourth term equals

$$\sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)}\right)^2\right]$$

$$= \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \sum_{j_1,j_2,j_3,j_4=1}^{n_L} W_{1j_1}^{(L+1)} W_{1j_2}^{(L+1)} W_{1j_3}^{(L+1)} W_{1j_4}^{(L+1)} \partial_\mu \sigma(z_{j_1}^{(L)}) \partial_\nu \sigma(z_{j_2}^{(L+1)}) \partial_\mu \sigma(z_{j_3}^{(L)}) \partial_\nu \sigma(z_{j_4}^{(L+1)})\right]$$

$$= 2 \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L^2} \sum_{j_1,j_2=1}^{n_L} \partial_\mu \sigma(z_{j_1}^{(L)}) \partial_\nu \sigma(z_{j_1}^{(L+1)}) \partial_\mu \sigma(z_{j_2}^{(L)}) \partial_\nu \sigma(z_{j_2}^{(L+1)})\right]$$

$$+ \sum_{\mu,\nu \leq L} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \frac{4}{n_L^2} \sum_{j_1,j_2=1}^{n_L} \left(\partial_\mu \sigma(z_{j_1}^{(L)}) \partial_\nu \sigma(z_{j_2}^{(L+1)})\right)^2\right]$$

$$= 2Y^{(L)}\left[\partial_\mu z \partial_\nu z, \partial_\mu z \partial_\nu z\right] + Y^{(L)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + \frac{3}{n_L} Y^{(L)}\left[(\partial_\mu z \partial_\nu z)^2\right].$$

Putting this all together yields

$$\sum_{\mu,\nu \leq L+1} \mathbb{E}\left[(\widehat{\eta}_\mu \widehat{\eta}_\nu)^2 \left(\partial_\mu z_1^{(L+1)} \partial_\nu z_1^{(L+1)}\right)^2\right]$$

$$= \frac{1}{2} \left(\eta^{(L+1)}\right)^4 \left(\mathbb{E}\left[\left(\frac{1}{n_L} \left\|z^{(L)}\right\|_2^2\right)^2\right] + \frac{1}{n_L} \mathbb{E}\left[\frac{1}{n_L} \left\|z^{(L)}\right\|_4^4\right]\right)$$

$$+ 2 \left(\eta^{(L+1)}\right)^2 \left[Y^{(L)}\left[z^2, (\partial_\mu z)^2\right] + \frac{1}{n_L} Y^{(L)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ 2Y^{(L)}\left[\partial_\mu z \partial_\nu z, \partial_\mu z \partial_\nu z\right] + Y^{(L)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + \frac{3}{n_L} Y^{(L)}\left[(\partial_\mu z \partial_\nu z)^2\right],$$

which is precisely the statement of equation 7. $\square$

In particular, combining equation 1 and equation 3 with the result of the preceding Lemma yields the following result.

Corollary 4. We have,

$$
\mathbb{E}\left[||H_{\text{eff}}||_{HS}^2\right]
$$
$$
= \left(\eta^{(L+1)}\right)^2 \left[Y^{(L)}\left[(\partial_\mu z)^2, z^2\right] + \frac{1}{n_L}Y^{(L)}\left[(z\partial_\mu z)^2\right]\right] + Y^{(L)}\left[(\partial_{\mu\nu}z)^2, z^2\right] + \frac{1}{n_L}Y^{(L)}\left[(z\partial_{\mu\nu}z)^2\right]
$$
$$
+ 2Y^{(L)}\left[z\partial_{\mu\nu}z, z\partial_{\mu\nu}z\right] + \frac{2}{n_L}Y^{(L)}\left[(z\partial_{\mu\nu}z)^2\right]
$$
$$
+ \left(\eta^{(L+1)}\right)^2 Y^{(L)}\left[(\partial_\mu z)^2\right] + Y^{(L)}\left[(\partial_{\mu\nu}z)^2\right]
$$
$$
+ 2\left(\eta^{(L+1)}\right)^2 \left(Y^{(L)}\left[z\partial_\mu z, z\partial_\mu z\right] + \frac{1}{n_L}Y^{(L)}\left[(z\partial_\mu z)^2\right]\right)
$$
$$
+ 4Y^{(L)}\left[\partial_{\mu\nu}z\partial_\mu z, z\partial_\nu z\right] + 2Y^{(L)}\left[z\partial_{\mu\nu}z, \partial_\mu z\partial_\nu z\right] + \frac{6}{n_L}Y^{(L)}\left[z\partial_{\mu\nu}z\partial_\mu z\partial_\nu z\right]
$$
$$
+ \frac{1}{2}\left(\eta^{(L+1)}\right)^4 \left(\mathbb{E}\left[\left(\frac{1}{n_L}\left\|z^{(L)}\right\|_2^2\right)^2\right] + \frac{1}{n_L}\mathbb{E}\left[\frac{1}{n_L}\left\|z^{(L)}\right\|_4^4\right]\right)
$$
$$
+ \left(\eta^{(L+1)}\right)^2 \left[Y^{(L)}\left[z^2, (\partial_\mu z)^2\right] + \frac{1}{n_L}Y^{(L)}\left[(z\partial_\mu z)^2\right]\right]
$$
$$
+ 2Y^{(L)}\left[\partial_\mu z\partial_\nu z, \partial_\mu z\partial_\nu z\right] + Y^{(L)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + \frac{3}{n_L}Y^{(L)}\left[(\partial_\mu z\partial_\nu z)^2\right].
$$

H.3. Self-Consistent Recursions for $Y^{(\ell)}$'s

Our next task is to develop and then solve self-consistent recursions for those Y's that contain only one μ.

Lemma 3. We have

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2\right] = \frac{1}{2}\left(\eta^{(\ell+1)}\right)^2 \frac{1}{n_0}\left\|x\right\|_2^2 + Y^{(\ell)}\left[(\partial_\mu z)^2\right] \tag{8}
$$

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2, z^2\right] = \left(\eta^{(\ell+1)}\right)^2 \mathbb{E}\left[\left(\frac{1}{n_\ell}\|\sigma^{(\ell)}\|^2\right)^2\right] \tag{9}
$$
$$
+ Y^{(\ell)}\left[z^2, (\partial_\mu z)^2\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[z\partial_\mu, z\partial_\mu\right] + \frac{1}{n_\ell}\left(1 + \frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[(z\partial_\mu z)^2\right]
$$

$$
Y^{(\ell+1)}\left[z\partial_\mu z, z\partial_\mu z\right] = 2\frac{\left(\eta^{(\ell+1)}\right)^2}{n_{\ell+1}}\mathbb{E}\left[\left(\frac{1}{n_\ell}\|\sigma^{(\ell)}\|^2\right)^2\right] \tag{10}
$$
$$
+ \left(1 + \frac{1}{n_{\ell+1}}\right)Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + \frac{1}{n_\ell}\left(1 + \frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[(z\partial_\mu z)^2\right] + \frac{1}{n_{\ell+1}}Y^{(\ell)}\left[z^2, \partial_\mu z^2\right]
$$

$$
Y^{(\ell+1)}\left[(z\partial_\mu z)^2\right] = 2\left(\eta^{(\ell+1)}\right)^2 \mathbb{E}\left[\left(\frac{1}{n_\ell}\|\sigma^{(\ell)}\|^2\right)^2\right] \tag{11}
$$
$$
+ 2Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + Y^{(\ell)}\left[z^2, \partial_\mu z^2\right] + \frac{3}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right].
$$

Proof. We start with equation 8. We have

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2\right] = \mathbb{E}\left[\sum_{\mu\in\ell+1}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(\partial_\mu z_j^{(\ell+1)}\right)^2\right]
$$
$$
+ \mathbb{E}\left[\sum_{\mu\le\ell}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(\partial_\mu z_j^{(\ell+1)}\right)^2\right].
$$

The first term is

$$\mathbb{E}\left[\sum_{\mu\in\ell+1}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(\partial_\mu z_j^{(\ell+1)}\right)^2\right] = \left(\eta^{(\ell+1)}\right)^2\mathbb{E}\left[\frac{1}{n_\ell}\left|\left|\sigma^{(\ell)}\right|\right|^2\right] = \frac{||x||^2}{2n_0}\left(\eta^{(\ell+1)}\right)^2.$$

Next, for the second term, we have

$$\mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(\partial_\mu z_j^{(\ell+1)}\right)^2\right] = \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\sum_{k_1,k_2=1}^{n_L}W_{jk_1}^{(\ell+1)}W_{jk_2}^{(\ell+1)}\partial_\mu\sigma(z_{k_1}^{(\ell)})\sigma(z_{k_2}^{(\ell)})\right]$$

$$= \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{2}{n_\ell}\sum_{k=1}^{n_L}\left(\partial_\mu\sigma(z_k^{(\ell)})\right)^2\right]$$

$$= Y^{(\ell)}\left[(\partial_\mu z)^2\right].$$

Combining the preceding expressions yields

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2\right] = \frac{||x||^2}{2n_0}\left(\eta^{(\ell+1)}\right)^2 + Y^{(\ell)}\left[(\partial_\mu z)^2\right],$$

which is precisely equation 8. Next, let us derive equation 9. We have

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2, z^2\right] = \mathbb{E}\left[\sum_{\mu\in\ell+1}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(z_j^{(\ell+1)}\partial_\mu z_j^{(\ell+1)}\right)^2\right]$$

$$+ \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(z_j^{(\ell+1)}\partial_\mu z_j^{(\ell+1)}\right)^2\right].$$

The first term is

$$\mathbb{E}\left[\sum_{\mu\in\ell+1}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(z_j^{(\ell+1)}\partial_\mu z_j^{(\ell+1)}\right)^2\right] = \left(\eta^{(\ell+1)}\right)^2\mathbb{E}\left[\frac{1}{n_\ell}\sum_{j_1=1}^{n_\ell}\sigma(z_{j_1}^{(\ell)})^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(z_j^{(\ell+1)}\right)^2\right]$$

$$= \left(\eta^{(\ell+1)}\right)^2\mathbb{E}\left[\frac{1}{n_\ell^2}\sum_{j_2=1}^{n_\ell}\sigma(z_{j_1}^{(\ell)})^2\sigma(z_{j_2}^{(\ell)})^2\right]$$

$$= \left(\eta^{(\ell+1)}\right)^2\mathbb{E}\left[\left(\frac{1}{n_\ell}||\sigma^{(\ell)}||^2\right)^2\right].$$

In contrast, using Wick's theorem, the second term is

$$\mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(z_j^{(\ell+1)}\partial_\mu z_j^{(\ell+1)}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\sum_{j_1,j_2,j_3,j_4=1}^{n_\ell}W_{jj_1}^{(\ell+1)}W_{jj_2}^{(\ell+1)}W_{jj_3}^{(\ell+1)}W_{jj_4}^{(\ell+1)}\sigma(z_{j_1}^{(\ell)})\sigma(z_{j_2}^{(\ell)})\partial_\mu\sigma(z_{j_3}^{(\ell)})\partial_\mu\sigma(z_{j_4}^{(\ell)})\right]$$

$$= \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{4}{n_\ell^2}\sum_{j_1,j_2}^{n_\ell}\left(\sigma(z_{j_1}^{(\ell)})\partial_\mu\sigma(z_{j_2}^{(\ell)})\right)^2 + 2\sigma(z_{j_1}^{(\ell)})\partial_\mu\sigma(z_{j_1}^{(\ell)})\sigma(z_{j_2}^{(\ell)})\partial_\mu\sigma(z_{j_2}^{(\ell)})\right]$$

$$= Y^{(\ell)}\left[z^2,(\partial_\mu z)^2\right] + 2Y^{(\ell)}\left[z\partial_\mu, z\partial_\mu\right] + \frac{3}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right].$$

Hence, altogether, we find

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2, z^2\right] = \left(\eta^{(\ell+1)}\right)^2 \mathbb{E}\left[\left(\frac{1}{n_\ell}||\sigma^{(\ell)}||^2\right)^2\right] + Y^{(\ell)}\left[z^2, (\partial_\mu z)^2\right] + 2Y^{(\ell)}\left[z\partial_\mu, z\partial_\mu\right] + \frac{3}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right],$$

which is precisely equation 9. Next, we derive equation 10. We have

$$Y^{(\ell+1)}\left[z\partial_\mu z, z\partial_\mu z\right] = \mathbb{E}\left[\sum_{\mu\in\ell+1}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}} z_{j_1}^{(\ell+1)} z_{j_2}^{(\ell+1)} \partial_\mu z_{j_1}^{(\ell+1)} \partial_\mu z_{j_2}^{(\ell+1)}\right]$$

$$= \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}} z_{j_1}^{(\ell+1)} z_{j_2}^{(\ell+1)} \partial_\mu z_{j_1}^{(\ell+1)} \partial_\mu z_{j_2}^{(\ell+1)}\right].$$

The first term equals

$$\mathbb{E}\left[\sum_{\mu\in\ell+1}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}} z_{j_1}^{(\ell+1)} z_{j_2}^{(\ell+1)} \partial_\mu z_{j_1}^{(\ell+1)} \partial_\mu z_{j_2}^{(\ell+1)}\right] = \frac{\left(\eta^{(\ell+1)}\right)^2}{n_{\ell+1}}\mathbb{E}\left[\frac{1}{n_\ell}\sum_{j_1=1}^{n_\ell}\sigma(z_{j_1}^{(\ell)^2})\frac{1}{n_{\ell+1}}\sum_{j_2=1}^{n_{\ell+1}}\left(z_{j_2}^{(\ell+1)}\right)^2\right]$$

$$= 2\frac{\left(\eta^{(\ell+1)}\right)^2}{n_{\ell+1}}\mathbb{E}\left[\left(\frac{1}{n_\ell}||\sigma^{(\ell)}||^2\right)^2\right].$$

The second term is

$$\mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}} z_{j_1}^{(\ell+1)} z_{j_2}^{(\ell+1)} \partial_\mu z_{j_1}^{(\ell+1)} \partial_\mu z_{j_2}^{(\ell+1)}\right]$$

$$= \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}}\sum_{k_1,k_2,k_3,k_4=1}^{n_\ell} W_{j_1k_1}^{(\ell+1)}W_{j_1k_2}^{(\ell+1)}W_{j_2k_3}^{(\ell+1)}W_{j_2k_4}^{(\ell+1)}\sigma(z_{k_1}^{(\ell)})\partial_\mu\sigma(z_{k_2}^{(\ell)})\sigma(z_{k_3}^{(\ell)})\partial_\mu\sigma(z_{k_4}^{(\ell)})\right]$$

$$= \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2 \frac{4}{n_\ell^2}\sum_{k_1,k_2=1}^{n_\ell}\sigma(z_{k_1}^{(\ell)})\partial_\mu\sigma(z_{k_1}^{(\ell)})\sigma(z_{k_2}^{(\ell)})\partial_\mu\sigma(z_{k_2}^{(\ell)}) + \frac{2}{n_{\ell+1}}\left(\sigma(z_{k_1}^{(\ell)})\partial_\mu\sigma(z_{k_2}^{(\ell)})\right)^2\right]$$

$$= Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + \frac{1}{n_\ell}\left(1+\frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[(z\partial_\mu z)^2\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[z^2, \partial_\mu z^2\right].$$

Putting this together yields

$$Y^{(\ell+1)}\left[z\partial_\mu z, z\partial_\mu z\right] = 2\frac{\left(\eta^{(\ell+1)}\right)^2}{n_{\ell+1}}\mathbb{E}\left[\left(\frac{1}{n_\ell}||\sigma^{(\ell)}||^2\right)^2\right]$$

$$= Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + \frac{1}{n_\ell}\left(1+\frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[(z\partial_\mu z)^2\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[z^2, \partial_\mu z^2\right],$$

which is precisely equation 10. Finally, we derive equation 11. We have

$$Y^{(\ell+1)}\left[(z\partial_\mu z)^2\right] = \mathbb{E}\left[\sum_{\mu\in\ell+1}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(z_j^{(\ell+1)}\partial_\mu z_j^{(\ell+1)}\right)^2\right]$$

$$+ \mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}}\sum_{j=1}^{n_{\ell+1}}\left(z_j^{(\ell+1)}\partial_\mu z_j^{(\ell+1)}\right)^2\right].$$

The first term is

$$
\mathbb{E}\left[\sum_{\mu \in \ell+1} (\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}} \sum_{j=1}^{n_{\ell+1}} \left(z_j^{(\ell+1)} \partial_\mu z_j^{(\ell+1)}\right)^2\right] = \left(\eta^{(\ell+1)}\right)^2 \mathbb{E}\left[\frac{1}{n_\ell} \sum_{j_1=1}^{n_\ell} \sigma(z_{j_1}^{(\ell)}) \frac{1}{n_{\ell+1}} \sum_{j=1}^{n_{\ell+1}} \left(z_j^{(\ell+1)}\right)^2\right]
$$

$$
= 2\left(\eta^{(\ell+1)}\right)^2 \mathbb{E}\left[\left(\frac{1}{n_\ell} ||\sigma^{(\ell)}||^2\right)^2\right].
$$

The second term is

$$
\mathbb{E}\left[\sum_{\mu \le \ell} (\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}} \sum_{j=1}^{n_{\ell+1}} \left(z_j^{(\ell+1)} \partial_\mu z_j^{(\ell+1)}\right)^2\right]
$$

$$
= \mathbb{E}\left[\sum_{\mu \le \ell} (\widehat{\eta}_\mu)^2 \frac{1}{n_{\ell+1}} \sum_{j=1}^{n_{\ell+1}} \sum_{k_1,k_2,k_3,k_4=1}^{n_\ell} W_{jk_1}^{(\ell+1)} W_{jk_2}^{(\ell+1)} W_{jk_3}^{(\ell+1)} W_{jk_4}^{(\ell+1)} \sigma(z_{k_1}^{(\ell)}) \sigma(z_{k_2}^{(\ell)}) \partial_\mu \sigma(z_{k_3}^{(\ell)}) \partial_\mu \sigma(z_{k_4}^{(\ell)})\right]
$$

$$
= \mathbb{E}\left[\sum_{\mu \le \ell} (\widehat{\eta}_\mu)^2 \frac{4}{n_\ell} \sum_{k_1,k_2=1}^{n_\ell} \left(\sigma(z_{k_1}^{(\ell)}) \partial_\mu \sigma(z_{k_2}^{(\ell)})\right)^2 + 2\sigma(z_{k_1}^{(\ell)}) \partial_\mu \sigma(z_{k_1}^{(\ell)}) \sigma(z_{k_2}^{(\ell)}) \partial_\mu \sigma(z_{k_2}^{(\ell)})\right]
$$

$$
= 2Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + Y^{(\ell)}\left[z^2, \partial_\mu z^2\right] + \frac{3}{n_\ell} Y^{(\ell)}\left[(z\partial_\mu z)^2\right].
$$

So all together this yields

$$
Y^{(\ell+1)}\left[(z\partial_\mu z)^2\right] = 2\left(\eta^{(\ell+1)}\right)^2 \mathbb{E}\left[\left(\frac{1}{n_\ell} ||\sigma^{(\ell)}||^2\right)^2\right]
$$

$$
+ 2Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + Y^{(\ell)}\left[z^2, \partial_\mu z^2\right] + \frac{3}{n_\ell} Y^{(\ell)}\left[(z\partial_\mu z)^2\right],
$$

which is precisely equation 11. □

Inspecting the recursions in Lemma 3 immediately shows that, for $\ell = 1, \ldots, L$, we have that as $n \to \infty$

$$
Y^{(\ell)}\left[(\partial_\mu z)^2\right], Y^{(\ell)}\left[(\partial_\mu z)^2, z^2\right], Y^{(\ell)}\left[(z\partial_\mu z)^2\right] = O(1), \qquad Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] = O(n^{-1}).
$$

Thus, we obtain

Corollary 5. for $\ell = 1, \ldots, L$, we have that

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2\right] = \frac{1}{2n_0} ||x||_2^2 \sum_{\ell'=0}^{\ell} \left(\eta^{(\ell'+1)}\right)^2
$$

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2, z^2\right] = \frac{1}{2} \sum_{\ell'=0}^{\ell} \left(\eta^{(\ell'+1)}\right)^2 \mathbb{E}\left[\left(\frac{1}{n_{\ell'}} ||z^{(\ell')}||_2^2\right)^2\right] + O(n^{-1}).
$$

In particular, specializing to the case where $\eta^{(\ell)} = \eta$ is independent of ℓ, we obtain

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2\right] = \frac{\ell+1}{2n_0} ||x||_2^2 \eta^2
$$

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2, z^2\right] = \frac{\eta^2}{2} \left(\frac{||x||_2^2}{n_0}\right)^2 \sum_{\ell'=0}^{\ell} \prod_{\ell''=1}^{\ell'} \left(1 + \frac{2}{n_{\ell''}}\right) + O(n^{-1}).
$$

A simple consequence of this corollary is that, dropping terms on the order of $O(n^{-1})$, $O(\ell^{-1})$ and assuming that $||x||^2 = n_0$, gives

$$
Y^{(\ell+1)}\left[(\partial_\mu z)^2\right], Y^{(\ell+1)}\left[(\partial_\mu z)^2, z^2\right] = \frac{1}{2} \ell \eta^2.
$$

Our next step is to obtain obtain and solve recursions for Y's appearing in Lemma 2 that involve sums over two network weights μ and ν. The recursions are as follows.

Lemma 4. We have

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] = \left(\eta^{(\ell+1)}\right)^4 \mathbb{E}\left[\left(\frac{1}{n_\ell}\|\sigma(z^{(\ell)})\|_2^2\right)^2\right] + \left(\eta^{(\ell+1)}\right)^2\left[Y^{(\ell)}\left[(\partial_\mu z)^2, z^2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ Y^{(\ell)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[\partial_\mu z\partial_\nu z, \partial_\mu z\partial_\nu z\right] + \frac{1}{n_\ell}\left(1 + \frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[(\partial_\mu z\partial_\nu z)^2\right].$$

$$Y^{(\ell+1)}\left[\partial_\mu z\partial_\nu z, \partial_\mu z\partial_\nu z\right] = \left(\eta^{(\ell+1)}\right)^4 \mathbb{E}\left[\left(\frac{1}{n_\ell}\|\sigma^{(\ell)}\|_2^2\right)^2\right] + \frac{\left(\eta^{(\ell+1)}\right)^2}{n_{\ell+1}}\left[Y^{(\ell)}\left[(\partial_\mu z)^2, z^2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ Y^{(\ell)}\left[\partial_\mu z\partial_\nu z, \partial_\mu z\partial_\nu z\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + \left(1 + \frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[(\partial_\mu z\partial_\nu z)^2\right]$$

$$Y^{(\ell+1)}\left[z\partial_{\mu\nu}z, \partial_\mu z\partial_\nu z\right] = \frac{\left(\eta^{(\ell+1)}\right)^2}{n_{\ell+1}}\left[Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ Y^{(\ell)}\left[z\partial_{\mu\nu}z, \partial_\mu z\partial_\nu z\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[\partial_{\mu\nu}z\partial_\mu z, z\partial_\nu z\right] + \left(1 + \frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[z\partial_\mu z\partial_\nu z\partial_{\mu\nu}z\right]$$

$$Y^{(\ell+1)}\left[z\partial_\mu z, \partial_\mu z\partial_{\mu\nu}z\right] = Y^{(\ell)}\left[z\partial_\mu z, \partial_\mu z\partial_{\mu\nu}z\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[z\partial_{\mu\nu}z, (z\partial_\mu z)^2\right] + \left(1 + \frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[z\partial_\mu\partial_\nu\partial_{\mu\nu}z\right]$$

$$Y^{(\ell+1)}\left[(\partial_{\mu\nu}z)^2, z^2\right] = \left(\eta^{(\ell+1)}\right)^2\left[Y^{(\ell)}\left[(\partial_\mu z)^2, z^2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ Y^{(\ell)}\left[(\partial_{\mu\nu}z)^2, z^2\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[z\partial_{\mu\nu}z, z\partial_{\mu\nu}z\right] + \frac{1}{n_\ell}\left(1 + \frac{2}{n_{\ell+1}}\right)Y^{(\ell)}\left[(z\partial_{\mu\nu}z)^2\right]$$

$$Y^{(\ell+1)}\left[(z\partial_{\mu\nu}z)^2\right] = 2\left(\eta^{(\ell+1)}\right)^2\left[Y^{(\ell)}\left[z^2, (\partial_\mu z)^2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ Y^{(\ell)}\left[z^2, (\partial_{\mu\nu}z)^2\right] + 2Y^{(\ell)}\left[z\partial_{\mu\nu}z, z\partial_{\mu\nu}z\right] + \frac{2}{n_\ell}Y^{(\ell)}\left[(z\partial_{\mu\nu}z)^2\right]$$

$$Y^{(\ell+1)}\left[(\partial_\mu z\partial_\nu z)^2\right] = \left(\eta^{(\ell+1)}\right)^4 \mathbb{E}\left[\left(\frac{1}{n_\ell}\|\sigma^{(\ell)}\|^2\right)^2\right] + 2\left(\eta^{(\ell+1)}\right)^2\left[Y^{(\ell)}\left[z^2, (\partial_\mu z)^2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right]$$

$$+ Y^{(\ell)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + 2Y^{(\ell)}\left[\partial_\mu z\partial_\nu z, \partial_\mu z\partial_\nu z\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(\partial_\mu z\partial_\nu z)^2\right]$$

$$Y^{(\ell+1)}\left[z\partial_{\mu\nu}z, z\partial_{\mu\nu}z\right] = \frac{\left(\eta^{(\ell+1)}\right)^2}{n_{\ell+1}}\left(Y^{(\ell)}\left[z^2, (\partial_\mu z)^2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right) + \left(1 + \frac{1}{n_{\ell+1}}\right)Y^{(\ell)}\left[z\partial_{\mu\nu}z, z\partial_{\mu\nu}z\right]$$

$$+ \frac{1}{n_{\ell+1}}Y^{(\ell)}\left[z^2, (\partial_\mu z)^2\right] + \frac{1}{n_\ell}\left(1 + \frac{1}{n_{\ell+1}}\right)Y^{(\ell)}\left[(z\partial_{\mu\nu}z)^2\right]$$

$$Y^{(\ell+1)}\left[(\partial_{\mu\nu}z)^2\right] = \left(\eta^{(\ell+1)}\right)^2 Y^{(\ell)}\left[(\partial_\mu z)^2\right] + Y^{(\ell)}\left[(\partial_{\mu\nu}z)^2\right]$$

$$Y^{(\ell+1)}\left[z\partial_{\mu\nu}z\partial_\mu z\partial_\nu z\right] = \left(\eta^{(\ell+1)}\right)^2\left(Y^{(\ell)}\left[z\partial_\mu z, z\partial_\mu z\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right)$$

$$+ Y^{(\ell)}\left[z\partial_{\mu\nu}z, \partial_\mu z\partial_\nu z\right] + 2Y^{(\ell)}\left[z\partial_\mu z, \partial_\nu z\partial_{\mu\nu}z\right] + \frac{3}{n_\ell}Y^{(\ell)}\left[z\partial_{\mu\nu z}\partial_\mu z\partial_\nu z\right]$$

Proof. The proof of Lemma 4 is very similar to that of Lemma 3, so we will only give the details for $Y^{(\ell+1)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right]$.

We have

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] = \mathbb{E}\left[\sum_{\mu,\nu\in\ell+1}(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}}\left(\partial_\mu z_{j_1}^{(\ell+1)}\partial_\nu z_{j_2}^{(\ell+1)}\right)^2\right]$$

$$+ 2\mathbb{E}\left[\sum_{\mu\leq\ell,\nu\in\ell+1}(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}}\left(\partial_\mu z_{j_1}^{(\ell+1)}\partial_\nu z_{j_2}^{(\ell+1)}\right)^2\right]$$

$$+ \mathbb{E}\left[\sum_{\mu,\nu\leq\ell}(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}}\left(\partial_\mu z_{j_1}^{(\ell+1)}\partial_\nu z_{j_2}^{(\ell+1)}\right)^2\right].$$

The first term equals

$$\left(\eta^{(\ell+1)}\right)^4\mathbb{E}\left[\left(\frac{1}{n_\ell}\left\|\sigma(z^{(\ell)})\right\|^2\right)^2\right].$$

The second term is

$$2\mathbb{E}\left[\sum_{\mu\leq\ell,\nu\in\ell+1}(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}}\left(\partial_\mu z_{j_1}^{(\ell+1)}\partial_\nu z_{j_1}^{(\ell+1)}\right)^2\right]$$

$$= 2\left(\eta^{(\ell+1)}\right)^2\mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{1}{n_\ell}\sum_{j_1=1}^{n_\ell}\left(\sigma(z_{j_1}^{(\ell)})\right)^2\frac{1}{n_{\ell+1}}\sum_{j_2=1}^{n_{\ell+1}}\left(\partial_\mu z_{j_2}^{(\ell+1)}\right)^2\right]$$

$$= 2\left(\eta^{(\ell+1)}\right)^2\mathbb{E}\left[\sum_{\mu\leq\ell}(\widehat{\eta}_\mu)^2\frac{1}{n_\ell}\sum_{j_1=1}^{n_\ell^2}\left(\sigma(z_{j_1}^{(\ell)})\right)^2\left(\partial_\mu\sigma(z_{j_2}^{(\ell)})\right)^2\right]$$

$$= 2\left(\eta^{(\ell+1)}\right)^2\left[Y^{(\ell)}\left[z^2,(\partial_\mu z)^2\right] + \frac{1}{n_\ell}Y^{(\ell)}\left[(z\partial_\mu z)^2\right]\right].$$

Finally, the third term equals

$$\mathbb{E}\left[\sum_{\mu,\nu\leq\ell}(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}}\left(\partial_\mu z_{j_1}^{(\ell+1)}\partial_\nu z_{j_2}^{(\ell+1)}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{\mu,\nu\leq\ell}(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{1}{n_{\ell+1}^2}\sum_{j_1,j_2=1}^{n_{\ell+1}}\sum_{k_1,k_2,k_3,k_4=1}^{n_\ell}W_{j_1k_1}^{(\ell+1)}W_{j_1k_2}^{(\ell+1)}W_{j_2k_3}^{(\ell+1)}W_{j_2k_4}^{(\ell+1)}\partial_\mu\sigma(z_{k_1})^{(\ell)}\partial_\mu\sigma(z_{k_2})^{(\ell)}\partial_\nu\sigma(z_{k_3})^{(\ell)}\partial_\nu\sigma(z_{k_4})^{(\ell)}\right]$$

$$= \mathbb{E}\left[\sum_{\mu,\nu\leq\ell}(\widehat{\eta}_\mu\widehat{\eta}_\nu)^2\frac{4}{n_\ell^2}\sum_{j_1,j_2=1}^{n_\ell}\left(\partial_\mu\sigma(z_{j_1})^{(\ell)}\partial_\mu\sigma(z_{j_2})^{(\ell)}\right)^2 + \frac{2}{n_{\ell+1}}\partial_\mu\sigma(z_{j_1})^{(\ell)}\partial_\nu\sigma(z_{j_1})^{(\ell)}\partial_\mu\sigma(z_{j_2})^{(\ell)}\partial_\nu\sigma(z_{j_2})^{(\ell)}\right]$$

$$= Y^{(\ell)}\left[(\partial_\mu z)^2,(\partial_\nu z)^2\right] + \frac{2}{n_{\ell+1}}Y^{(\ell)}\left[\partial_\mu z\partial_\nu z,\partial_\mu z\partial_\nu z\right] + \frac{3}{n_\ell n_{\ell+1}}Y^{(\ell)}\left[(\partial_\mu z\partial_\nu z)^2\right].$$

Combining the preceding expressions completes the derivation of the recursion for $Y^{(\ell)}\left[(\partial_\mu z)^2,(\partial_\nu z)^2\right]$. $\qquad\square$

Inspecting these recursions immediately shows the following

Corollary 6. For $\ell = 1,\ldots,L$, we have that as $n\to\infty$

$$Y^{(\ell)}\left[\partial_\mu z\partial_\nu z,\partial_\mu z\partial_\nu z\right],\,Y^{(\ell)}\left[z\partial_\mu z,\partial_\mu z\partial_{\mu\nu}z\right],\,Y^{(\ell)}\left[z\partial_{\mu\nu}z,\partial_\mu z\partial_\nu z\right],\,Y^{(\ell)}\left[z\partial_{\mu\nu z},z\partial_{\mu\nu z}\right] = O(n^{-1}),$$

while all the other Y's are order 1. Moreover,

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2,(\partial_\nu z)^2\right] = \left(\eta^{(\ell+1)}\right)^4\mathbb{E}\left[\left(\frac{1}{2n_\ell}||z^{(\ell)}||_2^2\right)^2\right] + \left(\eta^{(\ell+1)}\right)^2Y^{(\ell)}\left[(\partial_\mu z)^2,z^2\right]$$

$$+ Y^{(\ell)}\left[(\partial_\mu z)^2,(\partial_\nu z)^2\right] + O(n^{-1})$$

30

Hence, assuming that $||x||^2 = n_0$, we find

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] = \frac{1}{4}\eta^4 + \frac{1}{2}\eta^2\ell + Y^{(\ell)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + O(n^{-1}) = \frac{\eta^2\ell^2}{4} + O(\ell^{-1}) + O(n^{-1}).$$

Similarly,

$$Y^{(\ell+1)}\left[(\partial_{\mu\nu}z)^2\right] = \ell\eta^4 + Y^{(\ell)}\left[(\partial_{\mu\nu}z)^2\right] + O(n^{-1}) = C_3\ell^2\eta^4 + O(n^{-1}).$$

And also,

$$Y^{(\ell+1)}\left[z^2, (\partial_{\mu\nu}z)^2\right] = \ell\eta^4 + Y^{(\ell)}\left[(\partial_{\mu\nu}z)^2\right] + O(n^{-1}) = C_4\ell^2\eta^4 + O(n^{-1}).$$

H.4. Completion of the Proof of Theorem 1

We are now in a position to complete the proof of Theorem 1. By combining the results at the end of the previous section with Corollary 4, we find

$$\mathbb{E}\left[||H_{\text{eff}}||^2_{HS}\right] = 2\left(\eta^{(\ell+1)}\right)^4\mathbb{E}\left[\frac{1}{n_\ell}\left|\left|\sigma^{(L)}\right|\right|^2\right]^2$$
$$+ \left(\eta^{(L+1)}\right)^2\left(2Y^{(L)}\left[z^2, (\partial_\mu z)^2\right] + Y^{(L)}\left[(\partial_\mu z)^2\right]\right)$$
$$+ Y^{(L)}\left[z^2, (\partial_{\mu\nu}z)^2\right] + Y^{(L)}\left[(\partial_{\mu\nu}z)^2\right] + Y^{(L)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + O(n^{-1})$$
$$= C_1\eta^4\left(\frac{1}{n_0}||x||^2\right)^2 + C_2\eta^4 L$$
$$+ Y^{(L)}\left[z^2, (\partial_{\mu\nu}z)^2\right] + Y^{(L)}\left[(\partial_{\mu\nu}z)^2\right] + Y^{(L)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + O(n^{-1})$$

for some universal constants C_1, C_2. Moreover, we also find

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] = \left(\eta^{(\ell+1)}\right)^4\mathbb{E}\left[\left(\frac{1}{2n_\ell}||z^{(\ell)}||^2_2\right)^2\right] + \left(\eta^{(\ell+1)}\right)^2 Y^{(\ell)}\left[(\partial_\mu z)^2, z^2\right]$$
$$+ Y^{(\ell)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + O(n^{-1}).$$

Specializing to the case where $\eta^{(\ell)} = \eta$ is independent of ℓ, dropping terms on the order of $O(n^{-1}), O(\ell^{-1})$, and assuming that $||x||^2 = n_0$, we find

$$Y^{(\ell+1)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] = \frac{1}{4}\eta^4 + \frac{1}{2}\eta^2\ell + Y^{(\ell)}\left[(\partial_\mu z)^2, (\partial_\nu z)^2\right] + O(n^{-1}) = \frac{\eta^2\ell^2}{4} + O(\ell^{-1}) + O(n^{-1}).$$

Similarly,

$$Y^{(\ell+1)}\left[(\partial_{\mu\nu}z)^2\right] = \ell\eta^4 + Y^{(\ell)}\left[(\partial_{\mu\nu}z)^2\right] + O(n^{-1}) = C_3\ell^2\eta^4 + O(n^{-1}).$$

And also,

$$Y^{(\ell+1)}\left[z^2, (\partial_{\mu\nu}z)^2\right] = \ell\eta^4 + Y^{(\ell)}\left[(\partial_{\mu\nu}z)^2\right] + O(n^{-1}) = C_4\ell^2\eta^4 + O(n^{-1}).$$

So all together, we obtain

$$\mathbb{E}\left[||H_{\text{eff}}||^2_{HS}\right] = \frac{\eta^2 L^2}{4}\left(1 + O(n^{-1}) + O(L^{-1})\right).$$

In the setting of Theorem 1 we have

$$\mathbb{E}\left[\left|\left|H^{(L+1)}\right|\right|^2_{HS}\right] = \mathbb{E}\left[||H_{\text{eff}}||^2_{HS}\right]n^2\left(1 + O(n^{-1})\right),$$

up to universal constants, which yields the stated estimate.