Scale-Invariant Parameterizations

Karl Stratos

1 The Family of Parameterizations

Everett *et al.* (2024) consider the d-dimensional L-layer bigram language model:

$$h_0 = x \qquad h'_{L+2} = \mathrm{softmax}(h_{L+2}) - y \tag{1}$$
$$h_{l+1} = d^{-a_l} W_l h_l \quad \forall l = 0 \dots L+1 \qquad h'_l = d^{-a_l} W_l^\top h'_{l+1} \quad \forall l = L+1 \dots 1$$
$$W'_l = d^{-a_l} h'_{l+1} h_l^\top \quad \forall l = L+1 \dots 0$$

where $x, y \in \{0,1\}^V$ are one-hot vectors, $W_0 \in \mathbb{R}^{d \times V}$ and $W_{L+1} \in \mathbb{R}^{V \times d}$ are the embedding/readout layers, and $d^{-a_l} > 0$ is a "parameter multiplier". To understand the motivation behind this form, the reader is encouraged to first go over standard parameterization (SP, Appendix A) and muP (Appendix B).

1.1 First Step

We use the usual layerwise-variance for weight initialization as in SP (A.1) using the *a posteriori* power-law form:

$$\mathrm{Var}\left(W_l\right) = d^{-2b_l} \tag{2}$$

Then in the first forward and backward passes we have (Lemma D.5)

$$\mathrm{Var}\left(h_{l+1}\right) = d^{l - 2(\sum_{i=0}^{l} a_i + b_i)} \qquad \forall l = 0 \dots L+1 \tag{3}$$
$$\mathrm{Var}\left(h'_l\right) = d^{(L+1) - l - 2(\sum_{i=l}^{L+1} a_i + b_i)} \qquad \forall l = L+1 \dots 1 \tag{4}$$
$$\mathrm{Var}\left(W'_l\right) = d^{L + [[1 \leq l \leq L]] - 2((\sum_{i=0}^{L+1} a_i + b_i) - b_l)} \qquad \forall l = L+1 \dots 0 \tag{5}$$

whose square roots coincide with RMS in the infinite-width regime. We define **stability** as having constant activations $\mathrm{RMS}(h_l) = \Theta(1)$ for $l = 1 \dots L+1$ and bounded logits $\mathrm{RMS}(h_{L+2}) = O(1)$. The iterative nature of (3) implies the following unique conditions for stability in the first forward pass:

$$a_0 + b_0 = 0 \tag{6}$$
$$a_l + b_l = 1/2 \qquad \forall l = 1 \dots L \tag{7}$$
$$a_{L+1} + b_{L+1} \geq 1/2 \tag{8}$$

Under these conditions, $\sum_{i=l}^{L+1} a_i + b_i = (L+1-l)/2 + a_{L+1} + b_{L+1}$ and thus (4) and (5) imply

$$\mathrm{RMS}(h'_l) = \Theta(d^{-(a_{L+1} + b_{L+1})}) \qquad \forall l = L+1 \dots 1 \tag{9}$$
$$\mathrm{RMS}(W'_{L+1}) = \Theta(d^{-a_{L+1}}) \qquad \mathrm{RMS}(W'_l) = \Theta(d^{-(a_{L+1} + b_{L+1} + a_l)}) \qquad \forall l = L \dots 0 \tag{10}$$

For convenience, we write $\mathrm{RMS}(W'_l) = \Theta(d^{-g_l})$ where $g_l = a_{L+1} + [[l \leq L]](b_{L+1} + a_l)$.

1.2 Second Step

We assume *a posteriori* the learning rate has the power-law form

$$\eta_l = C d^{-c_l} \tag{11}$$

Parameterization	param. multiplier a_0	a_h	a_{L+1}	weight init. b_0	b_h	b_{L+1}	LR scale c_0	c_h	c_{L+1}
SP	0	0	0	0	1/2	1/2	0	1	1
NTK	0	1/2	1/2	0	0	0	0	1/2	1/2
muP	−1/2	0	1/2	1/2	1/2	1/2	1/2	1	1/2
MF	0	1/2	1	0	0	0	0	1/2	0

Table 1: Examples of scale-invariant parameterizations that ensure stability at initialization (6–8) and in subsequent steps (21–24), using momentumless Adam with full alignment.

for some constant $C > 0$. The change in weight $\Delta W_l = -\eta_l \mathbf{OPT}(W_l')$ depends on the optimizer, e.g.,

$$\begin{aligned}
&\text{(SGD)} && \Delta W_l = -Cd^{-c_l} W_l' && \Rightarrow && \Delta W_{l,i,j} = \Theta(d^{-(c_l+g_l)}) \\
&\text{(Adam)} && \Delta W_l = -Cd^{-c_l} \mathbf{sign}(W_l') && \Rightarrow && \Delta W_{l,i,j} = \Theta(d^{-c_l}) \qquad (12) \\
&\text{(Adafactor)} && \Delta W_l = -Cd^{-c_l} \mathrm{RMS}(W_l)\mathbf{sign}(W_l') && \Rightarrow && \Delta W_{l,i,j} = \Theta(d^{-(c_l+b_l)})
\end{aligned}$$

For simplicity we will assume (12) and parameterize the update scale as

$$\mathrm{RMS}(\Delta h_l) = O(d^{-r_l}) \tag{13}$$

for some $r_l \geq 0$. Maintaining stability requires $r_l \geq 0$ for all l (updates do not grow with width). Since $\Delta h_1 = d^{-a_0}\mathrm{col}(\Delta W_0)$ and thus $\Delta h_{1,i} = \Theta(d^{-(a_0+c_0)})$ under (12), we must first have

$$r_1 = a_0 + c_0 \geq 0 \tag{14}$$

For $l = 1 \dots L+1$, we have $\Delta h_{l+1} = d^{-a_l}(W_l \Delta h_l + \Delta W_l h_l + \Delta W_l \Delta h_l)$. To measure each term's alignment strength, we impose the *a posteriori* forms

$$\mathrm{RMS}(W_l \Delta h_l) = \Theta(d^{\omega_l} \times \mathrm{RMS}(W_l) \times \mathrm{RMS}(\Delta h_l)) \tag{15}$$
$$\mathrm{RMS}(\Delta W_l h_l) = \Theta(d^{\alpha_l} \times \mathrm{RMS}(\Delta W_l) \times \mathrm{RMS}(h_l)) \tag{16}$$
$$\mathrm{RMS}(\Delta W_l \Delta h_l) = \Theta(d^{u_l} \times \mathrm{RMS}(\Delta W_l) \times \mathrm{RMS}(\Delta h_l)) \tag{17}$$

where $\omega_l, \alpha_l, u_l \in [0,1]$ are invariant to scale and thus capture only interaction.[1] For (16) and (17), we know retrospectively that $\alpha_l, u_l \in [1/2, 1]$ with $\alpha_l = u_l = 1$ under full alignment and $\alpha_l = u_l = 1/2$ under no alignment (see (30)). Thus a sufficient condition to ensure $r_{l+1} \geq 0$ for $l = 1 \dots L+1$ is

$$\begin{aligned}
d^{-a_l} \times (15) &= O(d^{-a_l} \times d^{\omega_l} \times d^{-b_l} \times d^{-r_l}) = O(1) && \Leftrightarrow && a_l + b_l + r_l - \omega_l \geq 0 \qquad (18) \\
d^{-a_l} \times (16) &= O(d^{-a_l} \times d^{\alpha_l} \times d^{-c_l} \times d^{0}) = O(1) && \Leftrightarrow && a_l + c_l - \alpha_l \geq 0 \qquad (19) \\
d^{-a_l} \times (17) &= O(d^{-a_l} \times d^{u_l} \times d^{-c_l} \times d^{-r_l}) = O(1) && \Leftrightarrow && a_l + c_l + r_l - u_l \geq 0 \qquad (20)
\end{aligned}$$

where (18) simplifies to $1/2 + r_l - \omega_l \geq 0$ for $l = 1 \dots L$ by (7). Assuming full alignment, and assuming $r_l \geq 0$ is maintained iteratively $l = 1 \dots L+1$, we can intersect the conditions (14) and (18–20) against (6–8) to have

$$\begin{aligned}
a_0 + c_0 &\geq 0 && && (21) \\
a_l + c_l &\geq 1 && \forall l = 1 \dots L+1 && (22) \\
\omega_l &\leq 1/2 && \forall l = 1 \dots L && (23) \\
a_{L+1} + b_{L+1} &\geq \max(1/2, \omega_{L+1}) && && (24)
\end{aligned}$$

Since ω_l is not configurable, assuming (23) is a clean sufficient assumption to achieve stability. However, the readout layer allows for some wiggle room. muP assumes the worst-case dependence $\omega_{L+1} = 1$ and uses $a_{L+1} = b_{L+1} = 1/2$ to satisfy (24). Everett *et al.* (2024) relax the assumption to $\omega_{L+1} = 1/2$ and demonstrate empirical scale invariance. Example parameterizations that satify these conditions are reproduced in Table 1.

[1] More formally we may write, e.g., $\omega_l = \lim_{d\to\infty} \log_d \frac{\mathrm{RMS}(W_l \Delta h_l)}{\mathrm{RMS}(W_l)\mathrm{RMS}(\Delta h_l)}$ in probability.

1.2.1 Equivalence classes

Pick any parameterization (a_l, b_l, c_l) satisfying (6–8) and (21–24). Pick any scalar $\theta_l \in \mathbb{R}$ and redefine

$$a_l \leftarrow a_l + \theta_l \qquad b_l \leftarrow b_l - \theta_l \qquad c_l \leftarrow c_l - \theta_l$$

It is clear that the conditions still hold. Thus one stable parameterization defines an infinite family of equivalent parameterizations. In particular, in Table 1 we see that SP $\equiv$ NTK and muP $\equiv$ MF.

1.3 Subsequent Steps

The above conditions maintain $\text{RMS}(h_l) = \Theta(1)$ for all l. The Adam update does not modify the asymptotic size of the weights and gradients. Thus assuming that the interaction scales $\omega_l, \alpha_l, u_l \in [0, 1]$ remain stable throughout training, stability is maintained inductively for a constant number of steps $T = \Theta(1)$.

2 Attention

Attention is used to extend the bigram language model (1) to n-grams. All inputs maintain independent MLP structures except in the attention layer parameterized by per-head weights $W_q, W_k, W_v \in \mathbb{R}^{d_H \times d}$ and $W_o \in \mathbb{R}^{d \times d_H}$. The score between a pair of activations $h, h_{\text{past}} \in \mathbb{R}^d$ is computed by

$$q = \underbrace{W_q}_{d_H \times d} \underbrace{h}_{d \times 1} \qquad k = \underbrace{W_k}_{d_H \times d} \underbrace{h_{\text{past}}}_{d \times 1} \qquad s = \frac{1}{\sqrt{d_H}} \sum_{i=1}^{d_H} q_i k_i$$

With stable initialization (6–8), the variance of both q and k is $\Theta(1)$.[2] Thus $\text{Var}(s) = (1/d_H) \sum_{i=1}^{d_H} \Theta(1)\Theta(1) = \Theta(1)$ (conditioning on h, h_{past}) thanks to the explicit scale factor proposed in the original transformer paper.[3] Given a sequence of past activations $X \in \mathbb{R}^{d \times n}$ and a distribution $p \in \mathbb{R}^n$ (computed using these scores), the per-head output is computed by

$$V = \underbrace{W_v}_{d_H \times d} \underbrace{X}_{d \times n} = [v_1 \dots v_n] \qquad o = \sum_{j=1}^{n} p_j v_j$$

where $o_i = \mathbf{E}[v_j]$ implies $\text{Var}(o_i) = \Theta(1)$. The final output combines H such heads $o^{(1)} \dots o^{(H)} \in \mathbb{R}^{d_H}$ by

$$o_{\text{final}} = \sum_{k=1}^{H} \underbrace{W_o^{(h)}}_{d \times d_H} \underbrace{o^{(h)}}_{d_H \times 1}$$

Since $\text{Var}(W_o) = \Theta(1/d)$, we have $\text{Var}(o_{\text{final},i}) = \Theta(1)$ assuming the number of heads growing in width $H = \Theta(d)$. This output $o_{\text{final}} \in \mathbb{R}^d$ is fed into the next MLP layer. Thus the whole network remains stable at initialization even with attention layers, and any scale-invariant parameterization that ensures the activation change stays constant (e.g., scaling the learning rates for attention weights properly) will maintain this stability.

References

Everett, K. E., Xiao, L., Wortsman, M., Alemi, A. A., Novak, R., Liu, P. J., Gur, I., Sohl-Dickstein, J., Kaelbling, L. P., Lee, J., and Pennington, J. (2024). Scaling exponents across parameterizations and optimizers. In *Forty-first International Conference on Machine Learning*.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Yang, G. and Hu, E. J. (2020). Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*.

[2] In fact, a popular practice now is to have an explicit RMSNorm applied to q and k ("QK-norm") which guarantees this stability.

[3] Typically $d_H = \Theta(1)$ is a fixed constant (e.g., we match $d = d_H H$ by only changing the number of heads H), so technically this explicit scaling is not necessary for the purpose of width invariance.

A Standard Parameterization (SP)

An L-layer transformer without attention and normalization is a bigram language model with weights $W_0 \dots W_{L+1}$ where $W_l \in \mathbb{R}^{d_{l+1} \times d_l}$. We view training as a function of the hidden widths $d_1 \dots d_{L+1}$, so we can omit elementwise nonlinearity (Appendix C). Given a bigram $x, y \in \{0,1\}^V$ as one-hot vectors, the forward and backward passes for the cross-entropy loss compute

$$\begin{aligned} h_0 &= x & h'_{L+2} &= \mathrm{softmax}(h_{L+2}) - y \\ h_{l+1} &= W_l h_l \quad \forall l = 0 \dots L+1 & h'_l &= W_l^\top h'_{l+1} \quad \forall l = L+1 \dots 1 \\ & & W'_l &= h'_{l+1} h_l^\top \quad \forall l = L+1 \dots 0 \end{aligned} \tag{25}$$

A.1 Initialization

We assume that the weights $W_{l,i,j}$ are sampled iid from a symmetric zero-mean distribution with variance $\sigma_l^2 > 0$. Let $\mathrm{Var}\,(X)$ denote the variance of a single entry of X when all entries have the same variance. Then in the first forward and backward passes (Lemma D.1)

$$\begin{aligned} \mathrm{Var}\,(h_1) &= \sigma_0^2 \\ \mathrm{Var}\,(h_l) &= \sigma_{l-1}^2 d_{l-1} \mathrm{Var}\,(h_{l-1}) & \forall l &= 2 \dots L+2 \\ \mathrm{Var}\,(h'_{L+1}) &= \sigma_{L+1}^2 \mathbf{E}[\|h'_{L+2}\|^2] \\ \mathrm{Var}\,(h'_l) &= \sigma_l^2 d_{l+1} \mathrm{Var}\,(h'_{l+1}) & \forall l &= L \dots 1 \\ \mathrm{Var}\,(W'_{L+1}) &= \mathrm{Var}\,(h_{L+1})\, \mathbf{E}[(h'_{L+2,i})^2] \\ \mathrm{Var}\,(W'_l) &= \mathrm{Var}\,(h'_{l+1})\, \mathrm{Var}\,(h_l) & \forall l &= L \dots 1 \\ \mathrm{Var}\,(W'_{0,i,j}) &= [[x_j = 1]] \mathrm{Var}\,(h'_1) \end{aligned}$$

The logit gradient $h'_{L+2} \in [-1,1]^V$ is width-invariant, so $\mathbf{E}[(h'_{L+2,i})^2] = \Theta(1)$ and $\mathbf{E}[\|h'_{L+2}\|^2] = \Theta(1)$.

A.1.1 Hidden layers

Using $\sigma_l^2 = 1/d_l$ for $l = 1 \dots L+1$ prevents exploding variance in activations, yielding

$$\mathrm{Var}\,(h_l) = \sigma_0^2 \qquad \forall l = 1 \dots L+2 \tag{26}$$

On the other hand, using $\sigma_l^2 = 1/d_{l+1}$ for $l = L \dots 1$ prevents exploding variance in gradients, yielding

$$\mathrm{Var}\,(h'_l) = \Theta(\sigma_{L+1}^2) \qquad \forall l = L \dots 1 \tag{27}$$

A popular tradeoff is to use the average width $\sigma_l^2 = 2/(d_l + d_{l+1})$ (Glorot and Bengio, 2010) for $l = 1 \dots L$, but in practice it does not matter since typically $d_1 \dots d_{L+1}$ grow proportionally (e.g., $d_{l+1} = c_l d_l$ where c_l is some constant factor like 4 or 1/4). Thus in the asymptotic regime, we assume $d = d_1 = \cdots d_{L+1}$ WLOG and use $\sigma_l^2 = 1/d$ for $l = 1 \dots L$.

A.1.2 Embedding and readout layers

Note that σ_{L+1}^2 triggers a tradeoff: using $\sigma_{L+1}^2 = 1/d$ stabilizes the logits h_{L+2} (26) but shrinks the activation gradients (27). The choices of σ_0^2 and σ_{L+1}^2 together control $\mathrm{Var}\,(W'_l)$. Table 2 lists elementwise variances under different choices of σ_0^2 and σ_{L+1}^2 (no RMSNorm). With tied embeddings $W_0 = W_{L+1}^\top$, the gradient will be accumulated and will not affect the asymptotic behavior, but we have no choice but to use $\sigma_0^2 = \sigma_{L+1}^2$.

A.1.3 Bonus: RMSNorm

In real transformers, we apply $X \mapsto \mathrm{RMSNorm}(X) = X/\mathrm{RMS}(X)$ between layers, making the activations unit-variance for any X in the forward pass. But the normalization layer also annihilates the component of the gradient parallel to X (i.e., we cannot learn from the magnitude, which was not used) and scales it by $1/\mathrm{RMS}(X)$ in the backward pass (Lemma D.2). For illustration, consider incorporating RMSNorm as $h_l = \mathrm{RMSNorm}(W_{l-1} h_{l-1})$ for all layers except $l = L+2$. It turns out that (Lemma D.3)

RMSNorm	σ_0^2	σ_{L+1}^2	$h_1 \dots h_{L+1}$	h_{L+2}	h'_{L+2}	$h'_{L+1} \dots h'_1$	W'_{L+1}	$W'_L \dots W'_1$	W'_0
	1	$1/d$	1	1	1	$1/d$	1	$1/d$	$1/d$
✓			1	1	1	$1/d$	1	$1/d$	$1/d$
	$1/d$	$1/d$	$1/d$	$1/d$	$1/d$	$1/d$	$1/d$	$1/d^2$	$1/d$
✓			1	1	1	$1/d$	1	$1/d$	1
	1	1	1	d	1	1	1	1	1
✓			1	d	1	1	1	1	1

Table 2: Elementwise variances (asymptotic in the hidden width d) under different choices of σ_0^2 and σ_{L+1}^2 at initialization. We use the first-order approximation $\mathrm{Var}\left(h'_{L+2}\right) \approx \Theta(\sigma_{L+1}^2 d_{L+1} \mathrm{Var}(h_{L+1}))$ when $h_{L+2} \approx 0$. When RMSNorm is ✓, we assume $h_l = \mathrm{RMSNorm}(W_{l-1}h_{l-1})$ for all layers except $l = L+2$. Most studies assume the first row for SP (25), which makes activations unit order without RMSNorm.

1. The RMS cancels the width propagation for activation gradients, so their variance is preserved for any σ_l^2.

2. Unfortunately, the weight gradients are still affected, so we should still use $\sigma_l^2 = 1/d$ for $l = 1 \dots L$.

The resulting variances shown in Table 2 (RMSNorm ✓).

A.2 Post-Initialization

We use RMS to measure per-element size more generally in training steps (e.g., it coincides with the square-root of Table 2 in the first step in the infinite-width regime). Maintaining RMS during training depends on

- The initial weight variance σ_l^2, which determines the initial RMS
- The choice of optimizer **OPT** and learning rate η_l, which determines the per-step weight update $\Delta W_l = -\eta_l O_l$ where $O_l = \mathbf{OPT}(W'_l)$ is a transformation of the gradient

Since the gradients depend on activations, maintaining the $\Theta(1)$ width-dependence of activations is key. The new activation after one training step is $h_{l+1}^{\mathrm{new}} = (W_l + \Delta W_l)(h_l + \Delta h_l)$, so we have

$$\begin{aligned}
\Delta h_1^{\mathrm{new}} &= -\eta_0 O_{0,:,i} && x_i = 1 \\
\Delta h_{l+1}^{\mathrm{new}} &= \underbrace{W_l \Delta h_l}_{①} + \underbrace{(-\eta_l O_l h_l^{\mathrm{new}})}_{②} && \forall l = 1 \dots L+1
\end{aligned} \tag{28}$$

The idea is we can choose η_l appropriately for the given **OPT** to make these elementwise $\Theta(1)$. At $l = 0$ we can ensure $\mathrm{RMS}(\Delta h_1^{\mathrm{new}}) = \Theta(1)$ by setting $\eta_0 = \Theta(1/O_0)$. Unfortunately in (28), ① is not controllable by the learning rate. To make analysis tractable, we enforce the following conditions.

Condition A.1. $||W_l||_2 = \Theta(1)$ for $l = 1 \dots L+1$ throughout training.

Condition A.2. $\mathrm{RMS}(W_l \Delta h_l) = \Theta(1)$ for $l = 1 \dots L+1$ throughout training.

Condition A.1 is relatively mild given that it holds at initialization.[4] Condition A.2, however, is not easily justifiable. Note that for $l = 1 \dots L$, Condition A.1 implies Condition A.2 since

$$\mathrm{RMS}(W_l \Delta h_l) = \frac{||W_l \Delta h_l||_2}{\sqrt{d_{l+1}}} \leq ||W_l||_2 \frac{||\Delta h_l||_2}{\sqrt{d}} = \Theta(1)\Theta(1) = \Theta(1) \qquad \forall l = 1 \dots L \tag{29}$$

where we inductively assume $\mathrm{RMS}(\Delta h_l) = \Theta(1)$ (i.e., $||\Delta h_l||_2 = \sqrt{d}$). This breaks at $l = L+1$ since $d_{l+1} = d_{L+2} = V = O(1)$ so that the bound becomes $\Theta(\sqrt{d})$. We will come back to this issue in muP (Appendix B) and assume both Condition A.1 and A.2 hold for SP.

[4] We invoke without proof the fact that an iid sub-Gaussian random matrix $B \in \mathbb{R}^{n \times m}$ with zero mean and variance $1/m$ satisfies $||B||_2 \to 1 + \sqrt{n/m}$ as $n, m \to \infty$, which is 2 for $l = 1 \dots L$ and 1 for $l = L+1$ in the case $B = W_l$ at initialization. We assume that subsequent updates are small enough to maintain $||W_l||_2 = \Theta(1)$.

A.2.1 Learning rates (LLN vs CLT)

2 has the entry $2_i = -\eta_l A_{l,i}$ where $A_{l,i} = \sum_{j=1}^d O_{l,i,j} h_{l,j}^{\text{new}}$ measures the update-activation alignment. Let $\mu_{l,i} = (1/d)\sum_{j=1}^d \mathbf{E}[O_{l,i,j} h_{l,j}^{\text{new}}]$ and assume $\|\text{Cov}((O_{l,i,j} h_{l,j}^{\text{new}})_{j=1}^d)\|_2 = \Theta(1)$ (the mean may still grow in d). Then

$$A_{l,i} = d\mu_{l,i} + O_p(\sqrt{d}) \tag{30}$$

where O_p is big-O in probability. So there are two cases:

- $\mu_{l,i} \neq 0 \Rightarrow A_{l,i} = \Theta(d)$: Set $\eta_l = \Theta(1/d)$ to make $2_i = \Theta(1)$.
- $\mu_{l,i} = 0 \Rightarrow A_{l,i} = \Theta(\sqrt{d})$: Set $\eta_l = \Theta(1/\sqrt{d})$ to make $2_i = \Theta(1)$.

These cases are so-called "LNN vs CLT" because (30) can be written as $A_{l,i}/d = \mu_{l,i} + O_p(1/\sqrt{d})$ which corresponds to the law of large numbers and $\bar{A}_{l,i}/\sqrt{d} = O_p(1)$ which corresponds to the central limit theorem. Note that alignment is not static; it seems inevitable that alignment will emerge during training given that weights and activations coevolve. But committing to one specific assumption allows us to prove concrete results like the following.

Example A.1. Assume $\sigma_0^2 = 1$ and $\sigma_l^2 = 1/d$ for $l = 1 \dots L+1$. Assume momentumless Adam for **OPT**. Assume Condition A.1 and A.2 hold. Set

$$\eta_0 = \Theta(1) \qquad \eta_l = \begin{cases} \Theta(1/d) & \text{if Adam is aligned} \\ \Theta(1/\sqrt{d}) & \text{if Adam is not aligned} \end{cases} \qquad \forall l = 1 \dots L+1$$

Then the initial RMS is maintained for all training steps (Lemma D.4).

B muP

muP (Yang and Hu, 2020) relaxes Condition A.2 for $l = L+1$ and instead assumes the full upper bound:

$$\text{RMS}(W_{L+1}\Delta h_{L+1}) = \Theta(\sqrt{d}) \tag{31}$$

One justification for (31) is that $W'_{L+1} = h'_{L+2} h_{L+1}^\top$ involves the logit gradient h'_{L+2} whose mean is never zero, so ΔW_{L+1} will accumulate rank-1 components $u h_{L+1}^\top$ causing W_{L+1} and Δh_{L+1} to be aligned. Since this component (1 in (28)) is not controllable by the learning rate, the only choice we have in order to make $\text{RMS}(\Delta h_{L+2}) = \Theta(1)$ is to *scale* the readout layer by $1/\sqrt{d}$. This changes the forward and backward passes as

$$h_{L+2} = (1/\sqrt{d}) W_{L+1} h_{L+1} \qquad \begin{aligned} h'_{L+1} &= (1/\sqrt{d}) W_{L+1}^\top h'_{L+2} \\ W'_{L+1} &= (1/\sqrt{d}) h'_{L+2} h_{L+1}^\top \end{aligned}$$

The gradients shrink by $\sqrt{d}$, but it does not matter for magnitude-invariant optimizers like Adam for training purposes. Nonetheless, muP also scales the embedding layer by $\sqrt{d}$ to have

$$h_1 = \sqrt{d} W_0 x \qquad W'_0 = \sqrt{d} h'_1 h_0^\top$$

while at the same time changing σ_0^2 from 1 to $1/d$ to preserve the forward pass. This has the effect of restoring the gradient scale for embeddings. Unlike SP, muP's parameter multipliers force different LR exponents: with Adam, take $\eta_0 = \eta_{L+1} = 1/\sqrt{d}$ and $\eta_h = 1/d$ if aligned, $\eta_h = 1/\sqrt{d}$ if not aligned. With these, the muP RMS scales in Table 3 are maintained for any fixed number of training steps, under the same interaction assumptions as in the general framework (Table 1).

C Omitting Elementwise Nonlinearity

Let $\phi_1 \dots \phi_{L+2}$ denote elementwise functions. The forward pass computes activations $h_1 \dots h_{L+2}$ from $h_0 = x$ by

$$\begin{aligned} u_l &= W_{l-1} h_{l-1} \in \mathbb{R}^{d_l} \\ h_l &= \phi_l(u_l) \in \mathbb{R}^{d_l} \end{aligned} \tag{32}$$

Model	σ_0^2	σ_{L+1}^2	$h_1 \dots h_{L+1}$	Δh_{L+2}	h'_{L+2}	$h'_{L+1} \dots h'_1$	W'_{L+1}	$W'_L \dots W'_1$	W'_0
SP (Condition A.2)	1	$1/d$	1	1	1	$1/\sqrt{d}$	1	$1/\sqrt{d}$	$1/\sqrt{d}$
SP (31)	1	$1/d$	1	$\sqrt{d}$	1	$1/\sqrt{d}$	1	$1/\sqrt{d}$	$1/\sqrt{d}$
SP+readout (31)	1	$1/d$	1	1	1	$1/d$	$1/\sqrt{d}$	$1/d$	$1/d$
SP+emb/readout (31)	$1/d$	$1/d$	1	1	1	$1/d$	$1/\sqrt{d}$	$1/d$	$1/\sqrt{d}$

Table 3: Asymptotic RMS that needs to be maintained under different models. The Δh_{L+2} column denotes the logit change per training step, which stays invariant with SP under Condition A.2 but grows as square-root width $\sqrt{d}$ when relaxed to (31). Scaling the readout layer by $1/\sqrt{d}$ fixes the logit issue but also shrinks the gradients by $\sqrt{d}$. Scaling the embedding layer by $\sqrt{d}$ and shrinking the variance accordingly preserves the forward pass while upscaling the embedding gradient (muP).

The gradient wrt. the logits is $h'_{L+2} = \mathrm{softmax}(h_{L+2}) - y \in [-1, 1]^V$. By the chain rule, the gradients wrt. $h_{L+1} \dots h_1$ and $W_{L+1} \dots W_0$ are computed as

$$\begin{aligned} u'_{l+1} &= \phi'_{l+1}(u_{l+1}) \odot h'_{l+1} \in \mathbb{R}^{d_{l+1}} \\ h'_l &= W_l^\top u'_{l+1} \in \mathbb{R}^{d_l} \\ W'_l &= u'_{l+1} h_l^\top \in \mathbb{R}^{d_{l+1} \times d_l} \end{aligned} \tag{33}$$

We assume that ϕ_l is Λ-Lipschitz $|\phi_l(a) - \phi_l(b)| \leq \Lambda\, |a - b|$ for some constant $\Lambda > 0$. Then $|\phi'_l| \leq \Lambda$ (this holds at kinks using sub-gradients), so when we view (32) and (33) as functions of the widths $d_1 \dots d_{L+2}$, we have

$$\begin{aligned} h_{l,i} &= \phi_l(u_{l,i}) = \phi_l(0) + O(u_{l,i}) \\ u'_{l+1,i} &= \phi'_{l+1}(u_{l+1,i}) \times h'_{l+1,i} = O(h'_{l+1,i}) \end{aligned}$$

All common activation functions are Lipschitz (ReLU/tanh/identity $\Lambda = 1$, sigmoid $\Lambda = 1/4$) and also usually satisfy $\phi_l(0) = 0$ so

$$\begin{aligned} h_l &= O(W_{l-1} h_{l-1}) \\ h'_l &= O(W_l^\top h'_{l+1}) \end{aligned}$$

(i.e., ϕ_l does not change the asymptotic behavior of the input in either the forward nor the backward pass).

D Lemmas

Lemma D.1. In the first forward and backward pass,

- *(Activations)*: $\mathbf{E}[h_l] = 0_{d_l}$ and $\mathrm{Cov}\,(h_l) = \sigma_{l-1}^2 \mathbf{E}[||h_{l-1}||^2] I_{d_l}$ for $l = 1 \dots L+2$.
- *(Logit gradient)*: $\mathbf{E}[h'_{L+2}] = (1/V) 1_V - y$. A first-order approximation of $\mathrm{Cov}\left(h'_{L+2}\right) = \mathrm{Cov}\,(\mathrm{softmax}(h_{L+2}))$ around $h_{L+2} = 0_V$ is $\sigma_{L+1}^2 \mathbf{E}[||h_{L+1}||^2]((1/V^2) I_V - (1/V^3) 1_V 1_V^\top)$.
- *(Activation gradients)*: $\mathbf{E}[h'_l] = 0_{d_l}$ and $\mathrm{Cov}\,(h'_l) = \sigma_l^2 \mathbf{E}[||h'_{l+1}||^2] I_{d_l}$ for $l = L+1 \dots 1$.
- *(Weight gradients)*: $\mathbf{E}[W'_l] = 0_{d_{l+1} \times d_l}$ and $\mathrm{Var}(W'_{l,i,j}) = \mathbf{E}[(h'_{l+1,i})^2] \mathbf{E}[h_{l,j}^2]$ for $l = L+1 \dots 0$ with zero correlation except within the columns of W'_{L+1}.

Proof. (Activations): $\mathbf{E}[h_l] = \mathbf{E}[W_{l-1} h_{l-1}] = \mathbf{E}[W_{l-1}] \mathbf{E}[h_{l-1}] = 0_{d_l}$ since $W_{l-1} \perp h_{l-1}$ at initialization and $\mathbf{E}[W_{l-1,i,j}] = 0$. Then $\mathrm{Cov}\,(h_l) = \mathbf{E}[h_l h_l^\top] = \mathbf{E}[W_{l-1} h_{l-1} h_{l-1}^\top W_{l-1}^\top]$ has $\sum_{k,t} \mathbf{E}[W_{l-1,i,k} W_{l-1,j,t}] \mathbf{E}[h_{l-1,k} h_{l-1,t}]$ as the (i,j)-th entry, which is zero unless $i = j$ since the rows of W_{l-1} are independent. The i-th diagonal entry is $\sum_k \mathbf{E}[W_{l-1,i,k}^2] \mathbf{E}[h_{l-1,k}^2] = \sigma_{l-1}^2 \mathbf{E}[||h_{l-1}||^2]$ (which is σ_0^2 at $l = 1$).

(Logit gradient): Let $p = \mathrm{softmax}(h_{L+2})$. Conditioned on any h_{L+1}, the coordinates of $h_{L+2} = W_{L+1} h_{L+1} \in \mathbb{R}^V$

are iid (since the rows of W_{L+1} are iid), in particular exchangeable. This implies $\mathbf{E}[p_i] = 1/V$.[5] Thus $\mathbf{E}[h'_{L+2}] = \mathbf{E}[p] - y = (1/V)1_V - y$. Since $\mathrm{Cov}\left(h'_{L+2}\right) = \mathrm{Cov}\left(p\right)$ and the covariance of random variables bounded in $[0, 1]$ cannot exceed $1/4$, each entry is accordingly bounded. Let $J := \nabla_h \mathrm{softmax}(h)|_{h=0_V} = (1/V)I_V - (1/V^2)1_V 1_V^\top$ denote the Jacobian of softmax at 0_V. Then the first-order approximation of softmax around 0_V evaluated at h_{L+2} is $\hat{p} = (1/V)1_V + Jh_{L+2}$. Then $\mathrm{Cov}\left(h'_{L+2}\right) = \mathrm{Cov}\left(p\right) \approx \mathrm{Cov}\left(\hat{p}\right) = J\mathrm{Cov}\left(h_{L+2}\right) J^\top = \sigma^2_{L+1}\mathbf{E}[||h_{L+1}||^2]JJ^\top$ where $JJ^\top = (1/V^2)I_V - (1/V^3)1_V 1_V^\top$.

(Activation gradients): Let $\tilde{h}_{l+1}$ denote an iid copy of h_{l+1} sampled by independently re-drawing $\widetilde{W}_0 \dots \widetilde{W}_{L+1}$ and re-computing forward/backward ("ghost"). Clearly $\tilde{h}_{l+1}$ and h_{l+1} are equal in distribution but $\tilde{h}_{l+1} \perp W_l$, thus $\mathbf{E}[h'_l] = \mathbf{E}[W_l^\top h'_{l+1}] = \mathbf{E}[W_l^\top \tilde{h}'_{l+1}] = \mathbf{E}[W_l]^\top \mathbf{E}[\tilde{h}'_{l+1}] = 0_{d_l}$. The covariance is then $\mathrm{Cov}(h'_{l,i}, h'_{l,j}) = \mathbf{E}[h'_{l,i}h'_{l,j}] = \sum_{k,t}\mathbf{E}[W_{l,k,i}W_{l,t,j}h'_{l+1,k}h'_{l+1,t}] = \sum_{k,t}\mathbf{E}[W_{l,k,i}W_{l,t,j}\tilde{h}'_{l+1,k}\tilde{h}'_{l+1,t}] = \sum_{k,t}\mathbf{E}[W_{l,k,i}W_{l,t,j}]\mathbf{E}[\tilde{h}'_{l+1,k}\tilde{h}'_{l+1,t}]$. This is zero if $i \neq j$ and $\sigma_l^2\mathbf{E}[||h'_{l+1}||^2]$ otherwise.

(Weight gradients): We also have $\tilde{h}_{l+1} \perp h_l$ by construction, thus $\mathbf{E}[W'_l] = \mathbf{E}[h'_{l+1}h_l^\top] = \mathbf{E}[\tilde{h}'_{l+1}h_l^\top] = \mathbf{E}[h'_{l+1}]\mathbf{E}[h_l]^\top$. But $\mathbf{E}[h_l] = 0_{d_l}$ if $l \geq 1$ and $\mathbf{E}[h'_1] = 0_{d_1}$ from above, so $\mathbf{E}[W'_l] = 0_{d_{l+1}\times d_l}$ for all $l = 0 \dots L+1$. Then $\mathrm{Cov}(W'_{l,i,j}, W'_{l,k,t}) = \mathbf{E}[W'_{l,i,j}W'_{l,k,t}] = \mathbf{E}[h'_{l+1,i}h'_{l+1,k}h_{l,j}h_{l,t}] = \mathbf{E}[\tilde{h}'_{l+1,i}\tilde{h}'_{l+1,k}h_{l,j}h_{l,t}] = \mathbf{E}[h'_{l+1,i}h'_{l+1,k}]\mathbf{E}[h_{l,j}h_{l,t}]$. This is zero if $j \neq t$ (since $\mathbf{E}[h_{l,j}h_{l,t}] = 0$), or $l \in \{0 \dots L\}$ and $i \neq k$ (since $\mathbf{E}[h'_{l+1,i}h'_{l+1,k}] = 0$). □

Lemma D.2. Let

$$\mathrm{RMS}(u) := \sqrt{\frac{1}{d}\sum_{i=1}^{d} u_i^2} = \frac{||u||}{\sqrt{d}} \qquad\qquad v = \mathrm{RMSNorm}(u) := \frac{u}{\mathrm{RMS}(u)} = \sqrt{d}\bar{u}$$

where $\bar{u} = u/\,||u||$ (we omit epsilon and fix gating to 1 for simplicity). Then

- $v = \mathrm{RMSNorm}(cu)$ for all $c > 0$ with $\mathrm{RMS}(v) = 1$ and $||v|| = \sqrt{d}$.
- Let $g_{\mathrm{in}} = \frac{\partial \mathcal{L}}{\partial v}$ denote the incoming gradient and $g_{\mathrm{out}} = \frac{\partial \mathcal{L}}{\partial u}$ the outgoing gradient. Then

  $$g_{\mathrm{out}} = \frac{g_{\mathrm{in}}^\perp}{\mathrm{RMS}(u)}$$

  where $g_{\mathrm{in}}^\perp$ is the component of g_{in} perpendicular to u.

Proof. The first statement is obvious. The second statement follows from the Jacobian:

$$\nabla \mathrm{RMSNorm}(u) = \frac{1}{\mathrm{RMS}(u)}(I - \bar{u}\bar{u}^\top)$$

□

Lemma D.3. Let $h_0 = x \in \{0, 1\}^V$ and define the forward pass

$$u_l = W_{l-1}h_{l-1} \qquad\qquad h_l = \mathrm{RMSNorm}(u_l) \qquad\qquad \forall l = 1 \dots L+1$$
$$h_{L+2} = W_{L+1}h_{L+1}$$

Then for all $\sigma_0^2 \dots \sigma_{L+1}^2 > 0$ with $\sigma_{L+1}^2 = \Omega(1/d)$, in the infinite-width regime:

- $\mathrm{Var}\left(h_l\right) = 1$ for $l = 1 \dots L+1$ and $\mathrm{Var}\left(h_{L+2}\right) = \Omega(1)$.
- $\mathbf{E}[\|h'_{L+2}\|^2] = \Theta(1)$, $\mathrm{Var}\left(W'_{L+1}\right) = \Theta(1)$, and $\mathrm{Var}\left(h'_{L+1}\right) = \Theta(\sigma_{L+1}^2)$.

[5] More formally, $h_{L+2} = Ph_{L+2}$ for any permutation matrix $P \in \{0, 1\}$. Given any i, j, we can pick any P such that $P_{i,j} = 1$ and have

$$\mathbf{E}[p_i] = \mathbf{E}\left[\frac{\exp((Ph_{L+2})_i)}{\sum_{k=1}^V \exp((Ph_{L+2})_k)}\right] = \mathbf{E}\left[\frac{\exp(h_{L+2,j})}{\sum_{k=1}^V \exp(h_{L+2,k})}\right] = \mathbf{E}[p_j]$$

Thus $\mathbf{E}[p_i] = \pi$ for some constant $\pi > 0$ for $i = 1 \dots V$. Since $\mathbf{E}[\sum_{i=1}^V p_i] = \sum_{i=1}^V \mathbf{E}[p_i] = V\pi = 1$, we must have $\pi = 1/V$. Note that this bypasses the argument that $\mathbf{E}[\mathrm{softmax}(h_{L+2})] = \mathrm{softmax}(\mathbf{E}[h_{L+2}])$ (not true in general) and Jensen's inequality (exact only for constants and linear functions).

- $\mathrm{Var}\left(h'_l\right) = \Theta(\sigma^2_{L+1})$ and $\mathrm{Var}\left(W'_l\right) = \Theta(\frac{\sigma^2_{L+1}}{\sigma^2_l d})$ for $l = L \dots 1$.
- $\mathrm{Var}\left(W'_0\right) = \Theta(\frac{\sigma^2_{L+1}}{\sigma^2_0})$.

Proof. The forward pass is obvious. The backward pass for the cross-entropy loss is

$$\begin{aligned} h'_{L+2} &= \mathrm{softmax}(h_{L+2}) - y & & \\ h'_{L+1} &= W_{L+1}^\top h'_{L+2} & W'_{L+1} &= h'_{L+2} h_{L+1}^\top & \\ h'_l &= W_l^\top u'_{l+1} & W'_l &= u'_{l+1} h_l^\top & \forall l = L \dots 0 \end{aligned}$$

where $u'_l = \frac{\partial \mathcal{L}}{\partial u_l}$ for $l = 1 \dots L+1$ is given by (Lemma D.2)

$$u'_l = \frac{h''_l}{\mathrm{RMS}(u_l)} \qquad\qquad h''_l := \left(I_d - \bar{u}_l \bar{u}_l^\top\right) h'_l$$

At initialization $\mathrm{Var}\left(h''_l\right) = \Theta(\mathrm{Var}\left(h'_l\right))$.[6] Critically, since $u_l = W_{l-1} h_{l-1}$ has identically distributed entries with zero mean for $l = L+1 \dots 1$ at initialization, we may treat the RMS as constant variance in the infinite-width regime:

$$\mathrm{RMS}(u_l)^2 = \begin{cases} \mathrm{Var}\left(u_1\right) = \mathrm{Var}\left(W_0 x\right) = \sigma_0^2 & \text{if } l = 1 \\ \mathrm{Var}\left(u_l\right) = \mathrm{Var}\left(W_{l-1} h_{l-1}\right) = \sigma^2_{l-1} d \mathrm{Var}\left(h_{l-1}\right) = \sigma^2_{l-1} d & \text{if } l \geq 2 \end{cases}$$

This implies for $l = L \dots 1$:

$$\mathrm{Var}\left(h'_l\right) = \mathrm{Var}\left(W_l^\top u'_{l+1}\right) = \mathrm{Var}\left(W_l^\top \frac{h''_{l+1}}{\mathrm{RMS}(u_{l+1})}\right) = \frac{\mathrm{Var}\left(W_l^\top h''_{l+1}\right)}{\mathrm{RMS}(u_{l+1})^2} = \frac{\sigma_l^2 d \mathrm{Var}\left(h''_{l+1}\right)}{\sigma_l^2 d} = \mathrm{Var}\left(h''_{l+1}\right)$$

thus $\mathrm{Var}\left(h'_l\right) = \mathrm{Var}\left(h'_{L+1}\right) = \Theta(\sigma^2_{L+1})$. Likewise for $l = L \dots 1$:

$$\mathrm{Var}\left(W'_l\right) = \mathrm{Var}\left(u'_{l+1} h_l^\top\right) = \frac{\mathrm{Var}\left(h''_{l+1} h_l^\top\right)}{\mathrm{RMS}(u_{l+1})^2} = \frac{\mathrm{Var}\left(h''_{l+1}\right) \mathrm{Var}\left(h_l\right)}{\sigma_l^2 d} = \frac{\mathrm{Var}\left(h''_{l+1}\right)}{\sigma_l^2 d} = \Theta\left(\frac{\sigma^2_{L+1}}{\sigma_l^2 d}\right)$$

Finally, for the relevant column of W'_0, the variance is

$$\mathrm{Var}\left(W'_0\right) = \mathrm{Var}\left(u'_1\right) = \frac{\mathrm{Var}\left(h''_1\right)}{\mathrm{RMS}(u_1)^2} = \Theta\left(\frac{\sigma^2_{L+1}}{\sigma_0^2}\right)$$

□

Lemma D.4. Assume $\sigma_0^2 = 1$ and $\sigma_l^2 = 1/d$ for $l = 1 \dots L+1$. Assume momentumless Adam for **OPT**. Assume Condition A.1 and A.2 hold. Set

$$\eta_0 = \Theta(1) \qquad\qquad \eta_l = \begin{cases} \Theta(1/d) & \text{if Adam is aligned} \\ \Theta(1/\sqrt{d}) & \text{otherwise} \end{cases} \qquad \forall l = 1 \dots L+1 \tag{34}$$

Then the following invariants hold at all training steps:

$$\begin{aligned} \mathrm{RMS}(W_0) &= \Theta(1) & & (35) \\ \mathrm{RMS}(W_l) &= \Theta(1/\sqrt{d}) & \forall l = 1 \dots L+1 \quad & (36) \\ \mathrm{RMS}(h_l) &= \Theta(1) & \forall l = 1 \dots L+2 \quad & (37) \\ \mathrm{RMS}(h'_{L+2}) &= \Theta(1) & & (38) \\ \mathrm{RMS}(h'_l) &= \Theta(1/\sqrt{d}) & \forall l = L+1 \dots 1 \quad & (39) \\ \mathrm{RMS}(W'_{L+1}) &= \Theta(1) & & (40) \\ \mathrm{RMS}(W'_l) &= \Theta(1/\sqrt{d}) & \forall l = L \dots 0 \quad & (41) \end{aligned}$$

[6] $||h''_l||^2 = ||h'_l||^2 - (\bar{u}_l^\top h'_l)^2 \Rightarrow \mathbf{E}[||h''_l||^2] = \mathbf{E}[||h'_l||^2] - \mathbf{E}[(\bar{u}_l^\top h'_l)^2] = (d-1)\mathrm{Var}(h'_l) \Rightarrow \mathrm{Var}\left(h''_l\right) = (1 - 1/d)\mathrm{Var}(h'_l)$.

Proof. Since RMS coincides with standard deviation for variables with zero-mean iid elements (exact in the infinite-width regime, w.h.p. in general), the base case (i.e., the initial forward/backward pass) is immediate from the given initialization by taking the square-root of the first row of Table 2.

Assume (36–41) hold and consider a new forward/backward pass. Adam specifies $\Delta W_{l,i,j} = -\eta_l \mathbf{sign}(W'_{l,i,j}) = \Theta(\eta_l)$. We have $\Delta W_{0,i,j} = \Theta(1)$ and thus $W_{0,i,j} + \Delta W_{0,i,j} = \Theta(1) + \Theta(1) = \Theta(1)$ per element, so (35) is maintained. For $l = 1 \dots L+1$, we have $\Delta W_{l,i,j} = \Theta(\eta_l)$ where η_l is $\Theta(1/d)$ or $\Theta(1/\sqrt{d})$. In either case, $W_{l,i,j} + \Delta W_{l,i,j} = \Theta(1/\sqrt{d}) + \Theta(\eta_l) = \Theta(1/\sqrt{d})$ per element (since $1/\sqrt{d} \geq 1/d$), so (36) is maintained.

Likewise for the activations, it is sufficient to show Δh_l is of the same order as h_l per element (i.e., $\Theta(1)$). At $l = 1$ we have $\Delta h_1 = \Delta W_0 x = \mathrm{col}(\Delta W_0)$ where $\Delta W_{0,i,j} = \Theta(1)$, so we are done. For $l = 1 \dots L+1$, assume that $\Delta h_{l,i} = \Theta(1)$ (equivalently $||\Delta h_l||_2 = \Theta(\sqrt{d})$) and consider

$$\Delta h_{l+1} = \underbrace{W_l \Delta h_l}_{u} + \underbrace{\Delta W_l h_l^{\text{new}}}_{v}$$

For the first term, we have $\mathrm{RMS}(u) = \Theta(1)$ from Condition A.2. For the second term, we have

$$v_i = -\eta_l A_{l,i} = -\eta_l (d\mu_{l,i} + O_p(\sqrt{d})) = \begin{cases} \Theta(\eta_l d) & \text{if } \mu_{l,i} \neq 0 \\ \Theta(\eta_l \sqrt{d}) & \text{otherwise} \end{cases}$$

where O_p is big-O in probability. By our choice of the learning rate (34), this is $\Theta(1)$ always. Thus (37) is maintained.

For the activation gradients, (38) is trivial since $h'_{L+2} \in [-1,1]^V$. For $l = L+1 \dots 1$, since $h'_l = W_l^\top h'_{l+1}$ we have

$$\mathrm{RMS}(h'_l) \leq \frac{||W_l||_2 \, ||h'_{l+1}||_2}{\sqrt{d}} = \Theta(1)\Theta(1/\sqrt{d}) = \Theta(1/\sqrt{d})$$

which uses Condition A.1 and $\|h'_{l+1}\|_2 = \Theta(1)$ inductively ($\|h'_{L+2}\|_2 = \Theta(1)$ since V is constant). Thus $h'_{l,i} = \Theta(1/\sqrt{d})$ and (39) is maintained.

For the weight gradients $W'_l = h'_{l+1} h_l^\top$, we make similar arguments. At $l = L+1$ we have $\|W'_{L+1}\|_F \leq \|h'_{L+2}\|_2 \|h_{L+1}\|_2 = \Theta(1)\Theta(\sqrt{d}) = \Theta(\sqrt{d})$ and thus $\mathrm{RMS}(W'_{L+1}) = \Theta(\sqrt{d}/\sqrt{d}) = \Theta(1)$. Note that $\mathrm{RMS}(W'_{L+1}) = \Theta(\|W'_{L+1}\|_F/\sqrt{d})$ again because V is constant. For $l = L \dots 0$ we have $\|W'_l\|_F \leq \|h'_{l+1}\|_2 \|h_l\|_2 = \Theta(1)\Theta(\sqrt{d}) = \Theta(\sqrt{d})$ and thus $\mathrm{RMS}(W'_l) = \Theta(\sqrt{d}/d) = \Theta(1/\sqrt{d})$. So (40) and (41) are maintained.

□

Lemma D.5. Under (1) and (2), (3–5) hold.

Proof. For the forward pass, the base case is

$$\mathrm{Var}\left(h_{1,i}\right) = \mathrm{Var}\left(d^{-a_0}\mathrm{col}_i(W_0)\right) = d^{-2a_0}\mathrm{Var}\left(W_0\right) = d^{-2(a_0+b_0)}$$

For $l = 1 \dots L+1$, using the fact that W_l and h_l are independent at initialization,

$$\begin{aligned}\mathrm{Var}\left(h_{l+1,i}\right) = \mathrm{Var}\left(d^{-a_l}\sum_{j=1}^{d} W_{l,i,j} h_{l,j}\right) = d^{-2a_l}\sum_{j=1}^{d}\mathrm{Var}\left(W_{l,i,j}\right)\mathrm{Var}\left(h_{l,j}\right) &= d^{1-2(a_l+b_l)}\mathrm{Var}\left(h_{l,j}\right) \\ &= d^{1-2(a_l+b_l)} d^{(l-1)-2(\sum_{k=0}^{l-1} a_k+b_k)} \\ &= d^{l-2(\sum_{k=0}^{l} a_k+b_k)}\end{aligned}$$

For the backward pass, since $V = \Theta(1)$ and $h'_{L+2,j} \in [-1,1]$, the base case is

$$\mathrm{Var}\left(h'_{L+1,i}\right) = \mathrm{Var}\left(d^{-a_{L+1}}\sum_{j=1}^{V} W_{L+1,j,i} h'_{L+2,j}\right) = d^{-2(a_{L+1}+b_{L+1})}\mathrm{Var}\left(\sum_{j=1}^{V} h'_{L+2,j}\right) = \Theta(d^{-2(a_{L+1}+b_{L+1})})$$

For $l = L \dots 1$,

$$\begin{aligned}
\mathrm{Var}\left(h'_{l,i}\right) &= \mathrm{Var}\left(d^{-a_l}\sum_{j=1}^{d} W_{l,j,i} h'_{l+1,j}\right) \\
&= \mathrm{Var}\left(d^{-a_l}\sum_{j=1}^{d} W_{l,j,i} \tilde{h}'_{l+1,j}\right) && (\tilde{h}'_{l+1} \text{ is a ghost variable as defined in the proof of Lemma D.1}) \\
&= d^{-2a_l}\mathrm{Var}\left(\sum_{j=1}^{d} W_{l,j,i} \tilde{h}'_{l+1,j}\right) \\
&= d^{-2a_l}\sum_{j=1}^{d}\mathrm{Var}\left(W_{l,j,i}\right)\mathrm{Var}\left(\tilde{h}'_{l+1,j}\right) && (\text{since } \tilde{h}_{l+1} \text{ and } W_l \text{ are independent and elementwise iid}) \\
&= d^{1-2(a_l+b_l)} d^{L-l-2(\sum_{k=l+1}^{L+1} a_k+b_k)} \\
&= d^{(L+1)-l-2(\sum_{k=l}^{L+1} a_k+b_k)}
\end{aligned}$$

Likewise for the weight gradients, the base case is

$$\begin{aligned}
\mathrm{Var}\left(W'_{L+1,i,j}\right) = \mathrm{Var}\left(d^{-a_{L+1}}\tilde{h}'_{L+2,i}h_{L+1,j}\right) &= d^{-2a_{L+1}}\mathrm{Var}\left(\tilde{h}'_{L+2,i}\right)\mathrm{Var}\left(h_{L+1,j}\right) \\
&= \Theta(d^{-2a_{L+1}} d^{L-2(\sum_{k=0}^{L} a_k+b_k)}) = \Theta(d^{L-2((\sum_{k=0}^{L+1} a_k+b_k)-b_{L+1})})
\end{aligned}$$

For $l = L \dots 1$,

$$\begin{aligned}
\mathrm{Var}\left(W'_{l,i,j}\right) &= \mathrm{Var}\left(d^{-a_l}\tilde{h}'_{l+1,i}h_{l,j}\right) \\
&= d^{-2a_l}\mathrm{Var}\left(\tilde{h}'_{l+1,i}\right)\mathrm{Var}\left(h_{l,j}\right) \\
&= d^{-2a_l} \times d^{(L+1)-(l+1)-2(\sum_{k=l+1}^{L+1} a_k+b_k)} \times d^{(l-1)-2(\sum_{k=0}^{l-1} a_k+b_k)} \\
&= d^{(L+1)-2((\sum_{k=0}^{L+1} a_k+b_k)-b_l)}
\end{aligned}$$

Finally for $l = 0$,

$$\mathrm{Var}\left(W'_{0,i,j}\right) = \mathrm{Var}\left(d^{-a_0}h'_{1,i}x_j\right) = \begin{cases} 0 & \text{if } x_j = 0 \\ d^{-2a_0} d^{L-2(\sum_{k=1}^{L+1} a_k+b_k)} = d^{L-2((\sum_{k=0}^{L+1} a_k+b_k)-b_0)} & \text{if } x_j = 1 \end{cases}$$

$\square$