Scale-Invariant Parameterizations

Karl Stratos

Last updated: February, 2026

Abstract

Scale-invariant parameterizations aim to preserve training stability as width d and depth L grow. A "primal" view controls forward activations—under specific assumptions on how the sum of random variables behaves (D). For width, we shrink the model weights at initialization (e.g., $\asymp 1/d$ variance; embedding/readout weights may differ) and through layerwise learning-rate scaling during updates (often $\asymp 1/d$ in hidden layers) (1). For depth, we shrink the residual branch (e.g., $\asymp 1/L$) and potentially the learning rate (2). Post-LN can reduce forward-scale pressure, but introduces depth-dependent backward transport through a product of LN Jacobians. This can be alleviated by a "dual" parameterization that amplifies the shortcut inside LN (e.g., $\asymp L$) to improve gradient transport conditioning across depth (2.2).

Contents

1 Width Scaling

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

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

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". The model generalizes SP and muP (Appendix A and B).

1.1 First Step

Let each layer's weight W_l be elementwise iid with mean 0 and variance

$$\text{Var}(W_l) = d^{-2b_l} \tag{2}$$

The power-law form is *a posteriori*. Then in the first forward and backward passes we have (Lemma E.4)

$$\text{Var}(h_{l+1}) = d^{l - 2(\sum_{i=0}^{l} a_i + b_i)} \qquad \forall l = 0 \dots L+1 \tag{3}$$

$$\text{Var}(h'_l) = d^{(L+1) - l - 2(\sum_{i=l}^{L+1} a_i + b_i)} \qquad \forall l = L+1 \dots 1 \tag{4}$$

$$\text{Var}(W'_l) = d^{L + [[1 \le l \le L]] - 2((\sum_{i=0}^{L+1} a_i + b_i) - b_l)} \qquad \forall l = L+1 \dots 0 \tag{5}$$

whose square roots coincide with RMS in the infinite-width regime. We define **stability** as

$$\text{RMS}(h_l) = \Theta(1) \qquad \forall l = 1 \dots L+1 \tag{6}$$

$$\text{RMS}(h_{L+2}) = O(1) \tag{7}$$

The iterative nature of (3) implies the following unique conditions for stability:

$$a_0 + b_0 = 0 \tag{8}$$

$$a_l + b_l = 1/2 \qquad \forall l = 1 \dots L \tag{9}$$

$$a_{L+1} + b_{L+1} \ge 1/2 \tag{10}$$

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

$$\text{RMS}(h'_l) \asymp d^{-(a_{L+1} + b_{L+1})} \qquad \forall l = L+1 \dots 1 \tag{11}$$

$$\text{RMS}(W'_{L+1}) \asymp d^{-a_{L+1}} \qquad \text{RMS}(W'_l) \asymp d^{-(a_{L+1} + b_{L+1} + a_l)} \qquad \forall l = L \dots 0 \tag{12}$$

(we write $f(d) \asymp g(d)$ and $f(d) = \Theta(g(d))$ interchangeably).

1.2 Second Step

Maintaining stability (6–7) in the second forward pass means

$$\text{RMS}(h_l + \Delta h_l) = \Theta(1) \qquad \forall l = 1 \dots L+1 \tag{13}$$

$$\text{RMS}(h_{L+2} + \Delta h_{L+2}) = O(1) \tag{14}$$

where Δh_l is the change in activation caused by the change in weight $\Delta W_l = -\eta_l \mathbf{OPT}(W'_l)$. Assume that the learning rate has the *a posteriori* power-law form

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

for some constant $C > 0$. The magnitude of the weight change depends on the optimizer, e.g.,

$$\begin{aligned} &\text{(SGD)} && \Delta W_l = -C d^{-c_l} W'_l && \Rightarrow && \text{RMS}(\Delta W_l) \asymp d^{-c_l} \text{RMS}(W'_l) \\ &\text{(Adam)} && \Delta W_l = -C d^{-c_l} \mathbf{sign}(W'_l) && \Rightarrow && \text{RMS}(\Delta W_l) \asymp d^{-c_l} \\ &\text{(Adafactor)} && \Delta W_l = -C d^{-c_l} \text{RMS}(W_l) \mathbf{sign}(W'_l) && \Rightarrow && \text{RMS}(\Delta W_l) \asymp d^{-(c_l + b_l)} \end{aligned} \tag{15}$$

Parameterization	param. multiplier			weight init.			LR scale		
	a_0	a_h	a_{L+1}	b_0	b_h	b_{L+1}	c_0	c_h	c_{L+1}
SP (aka. "simple muP")	0	0	0	0	1/2	1/2	0	1	1
NTK (Jacot *et al.*, 2018)	0	1/2	1/2	0	0	0	0	1/2	1/2
muP (Yang and Hu, 2020)	−1/2	0	1/2	1/2	1/2	1/2	1/2	1	1/2
MF (Mei *et al.*, 2018)	0	1/2	1	0	0	0	0	1/2	0

Table 1: Examples of scale-invariant parameterizations that ensure stability at initialization (8–10) and in subsequent steps (24–27), using momentumless Adam with full update-activation alignment.

For simplicity we will focus on momentumless Adam (15).[1] We will also assume $c_l \geq b_l$ so that $\mathrm{RMS}(W_l + \Delta W_l) = O(d^{-b_l})$ (i.e., the initial weight size continues to dominate). Now we parameterize

$$\mathrm{RMS}(\Delta h_l) \asymp d^{-r_l} \tag{16}$$

and frame stability (13–14) as requiring $r_l \geq 0$ for all l. Since $\Delta h_{1,i} \in \{\pm d^{-(a_0+c_0)}\}$, we must first have

$$r_1 = a_0 + c_0 \geq 0 \tag{17}$$

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 isolate each term's alignment strength, we impose the *a posteriori* factorized forms

$$\mathrm{RMS}(W_l \Delta h_l) \asymp d^{\omega_l} \times \mathrm{RMS}(W_l) \times \mathrm{RMS}(\Delta h_l) \tag{18}$$

$$\mathrm{RMS}(\Delta W_l h_l) \asymp d^{\alpha_l} \times \mathrm{RMS}(\Delta W_l) \times \mathrm{RMS}(h_l) \tag{19}$$

$$\mathrm{RMS}(\Delta W_l \Delta h_l) \asymp d^{u_l} \times \mathrm{RMS}(\Delta W_l) \times \mathrm{RMS}(\Delta h_l) \tag{20}$$

where $\omega_l, \alpha_l, u_l \in [1/2, 1]$ (Appendix D). Then a sufficient condition to ensure $r_{l+1} \geq 0$ for $l = 1 \dots L+1$ is

$$a_l + b_l + r_l - \omega_l \geq 0 \tag{21}$$

$$a_l + c_l - \alpha_l \geq 0 \tag{22}$$

$$a_l + c_l + r_l - u_l \geq 0 \tag{23}$$

where (21) simplifies to $1/2 + r_l - \omega_l \geq 0$ for $l = 1 \dots L$ by (9). Assuming $\alpha_l = u_l = 1$ (i.e., the updates and activations are aligned), and assuming $r_l \geq 0$ is maintained iteratively $l = 1 \dots L+1$, we can intersect the conditions (17) and (21–23) against (8–10) to have

$$a_0 + c_0 \geq 0 \tag{24}$$

$$a_l + c_l \geq 1 \qquad \forall l = 1 \dots L+1 \tag{25}$$

$$\omega_l \leq 1/2 \qquad \forall l = 1 \dots L \tag{26}$$

$$a_{L+1} + b_{L+1} \geq \max(1/2, \omega_{L+1}) \tag{27}$$

Since ω_l is not configurable, we need to directly assume (26) to achieve stability (i.e., the weights and activation changes are not aligned for non-readout layers). 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 (27). 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.

1.2.1 Subsequent steps

The only autoregressive dependence we have in the second step is the fact that

$$\mathrm{RMS}(h_l) = O(1)$$

$$\mathrm{RMS}(W_l) = O(d^{-b_l})$$

for all l. The second always holds if $c_l \geq b_l$. The first holds inductively. Thus for any fixed number of gradient steps $T = O(1)$, the initial stability conditions (6–7) are maintained for all steps $t = 1 \dots T$, if we treat the interaction scales $\omega_l, \alpha_l, u_l \in [1/2, 1]$ as constant over this horizon (or at least remain stable).[2]

[1] For nonzero β_1, β_2, the Adam update fluctuates but converges to $\sqrt{\frac{1-\beta_1}{1+\beta_1}} + O(1-\beta_2)$ in expectation under mild assumptions.

[2] To see why we need a constant horizon, consider how $O(1)$ activation changes accumulate over $T = O(d)$ steps.

1.2.2 Equivalence classes

Pick any parameterization (a_l, b_l, c_l) satisfying (8–10) and (24–27). Pick any scalar $\theta_l \in \mathbb{R}$ and redefine

$$a_l \leftarrow a_l + \theta_l \qquad b_l \leftarrow b_l - \theta_l \qquad c_l \leftarrow c_l - \theta_l \tag{28}$$

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.

1.2.3 Feature learning and nontriviality

A loophole of stability is that we are only requesting $O(1)$ changes in activation to avoid their exploding in width, which admits the undesirable situation where even the cumulative changes vanish in width. To address this, the literature defines stronger conditions

$$\text{RMS}(h_{L+1}^{(T)} - h_{L+1}^{(0)}) = \text{RMS}\left(\sum_{t=1}^{T} \Delta h_{L+1}^{(t)}\right) = \Theta(1) \qquad \text{(aka. } \textbf{feature learning}\text{)}$$

$$\text{RMS}(h_{L+2}^{(T)} - h_{L+2}^{(0)}) = \text{RMS}\left(\sum_{t=1}^{T} \Delta h_{L+2}^{(t)}\right) = \Theta(1) \qquad \text{(aka. } \textbf{nontriviality}\text{)}$$

where the superscript denotes the number of gradient steps with $T = O(1)$. In the notation $\text{RMS}(\Delta h_l) \asymp d^{-r_l}$ (16), these conditions correspond to

- Stability: $r_l \geq 0$ for all l at all steps
- Feature learning: $r_{L+1} = 0$ at some step
- Nontriviality: $r_{L+2} = 0$ at some step

1.3 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 (8–10), the variance of both q and k is $\Theta(1)$.[3] 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.[4] 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.

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

[4] 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.

2 Depth Scaling

For this section we assume $d = 1$ and a simple MLP with residual connections $h_{l+1} = h_l + f_l(h_l)$ for $l = 0 \dots L$ where $h_0 = 0$. We also assume that $f_l(h_l) = \Theta(1)$ for all l (which holds with the layerwise normalization in practice), and that the variance does not accumulate superlinearly. Then the *top* activation has the form

$$h_{L+1} = \sum_{l=1}^{L} f_l(h_l) = L\mu + \Theta_p(\sqrt{L}) \asymp L^\rho = \begin{cases} \sqrt{L} & \text{if } \mu = 0 \\ L & \text{otherwise} \end{cases} \tag{29}$$

where we allow $\mu = (1/L)\sum_{l=1}^{L} \mathbf{E}[f_l(h_l)]$ ("typical" activation across layers) to be either zero ($\rho = 1/2$), e.g., because of network symmetries, or nonzero ($\rho = 1$). To counter this dependence, we may scale the connection as

$$h_{l+1} = h_l + L^{-\rho} f_l(h_l) = L^{-\rho} \sum_{i \le l} f_i(h_i) \tag{30}$$

yielding $h_{L+1} \asymp 1$. In post-initialization, consider a per-step weight change $\Delta w_l = \eta \operatorname{sign}(w_l')$ under momentumless Adam with constant learning rate $\eta > 0$. We will assume that $f_l(h_l)$ is roughly linear in w_l so that $\Delta f_l(h_l) \asymp \Delta w_l$. Then

$$\Delta h_{L+1} = L^{-\rho} \sum_{l=1}^{L} \Delta f_l(h_l) \asymp L^{-\rho} \sum_{l=1}^{L} \Delta w_l = L^{-\rho}\eta \sum_{l=1}^{L} \operatorname{sign}(w_l') \asymp \begin{cases} L^{1-\rho}\eta & \text{(if layer gradients are correlated)} \\ L^{1/2-\rho}\eta & \text{(otherwise)} \end{cases} \tag{31}$$

where $\sum_{l=1}^{L} \operatorname{sign}(w_l')$ grows like either $\Theta(L)$ or $\Theta(\sqrt{L})$. Since $L^{1-\rho} \ge L^{1/2-\rho}$, in either case we can set $\eta = \Theta(L^{\rho-1})$ to ensure $\Delta h_{L+1} = O(1)$. Assuming $\rho = 1/2$ yields the original depth-muP that scales both the residual connection and η by $L^{-1/2}$ (Yang *et al.*, 2023); assuming $\rho = 1$ yields a simpler version where only the residual connection is scaled by L^{-1}.

2.1 Joint Scaling

We can patch together the existing scaling schemes for width, depth, and optimizer-specific hyperparameters. Dey *et al.* (2025) consider one version ("completeP"). For width scaling, they use muP in Table 1 shifted by $\theta_l = 1/2$ for $l \in \{0, L+1\}$ (28), specifically

	param. multiplier			weight init.			LR scale		
Parameterization	a_0	a_h	a_{L+1}	b_0	b_h	b_{L+1}	c_0	c_h	c_{L+1}
muP	$-1/2$	0	$1/2$	$1/2$	$1/2$	$1/2$	$1/2$	1	$1/2$
muP shifted here	0	0	1	0	$1/2$	0	0	1	0

For depth scaling, they use $\rho = 1$ (Section 2), so $h_{l+1} = h_l + \frac{1}{L} f_l(h_l)$ for all residual connections. They use coupled weight decay scaling (1), which means growing $\lambda_h \asymp d$ for hidden layers to counter $c_h = 1$. Since all layers' gradients shrink like $\Theta(d^{-1})$ due to $a_{L+1} = 1$ but the residual blocks's gradients additionally shrink like $\Theta(L^{-1})$, they use layerwise ϵ_l to match the vanishing scales of the Adam moments $m_l, \sqrt{v_l}$ (i.e., $\Theta(d^{-1}L^{-1})$ for residual blocks, $\Theta(d^{-1})$ for embedding/readout layers).

2.2 Post-LN

Depth-muP controls the *top activation* h_{L+1} (29) by scaling the residual branch *down* (30) in $h_{l+1} = h_l + f_l(h_l)$. Instead, "post-LN" can be used to explicitly normalize the activation at every layer

$$h_{l+1} = \mathrm{LN}_l(h_l + f_l(h_l)) \tag{32}$$

where $\mathrm{LN}_l(u) = \frac{u}{\mathrm{RMS}(u)} \operatorname{diag}(\gamma_l)$ with paramater $\gamma_l \in \mathbb{R}^d$. While this controls output scale layerwise, it "disrupts" the residual stream in the backward pass. Let J_{f_l} denote the Jacobian of $f_l(h_l)$, and B_l the Jacobian of $\mathrm{LN}_l(h_l + f_l(h_l))$.[5] Since the bottom activation gradient is computed as $h_1' = \prod_{l=1}^{L} \left(\frac{\partial h_{l+1}}{\partial h_l}\right) h_{L+1}'$ from the top, it changes

[5] We use the transposed convention for Jacobians, i.e., $(\frac{\partial y}{\partial x})_{i,j} = \frac{\partial y_j}{\partial x_i}$.

from

$$h'_1 = \prod_{l=1}^{L} (I_d + J_{f_l})\, h'_{L+1} = \boxed{h'_{L+1}} + (\text{other terms})$$

under $h_{l+1} = h_l + f_l(h_l)$, to

$$h'_1 = \prod_{l=1}^{L} (I_d + J_{f_l})\, B_l h'_{L+1} = \boxed{\left(\prod_{l=1}^{L} B_l\right) h'_{L+1}} + (\text{other terms})$$

under $h_{l+1} = \mathrm{LN}_l(h_l + f_l(h_l))$. Thus the connection between h'_1 and h'_{L+1} is disrupted by a chain of LN Jacobians $B = \prod_{l=1}^{L} B_l$. Its norm shrinks exponentially in L under certain simplifying assumptions (roughly: f_l preserves size but changes direction, the gating parameter is uniform and does not change much between layers, see Lemma E.5):

$$||B||_2 \leq \prod_{l=1}^{L} ||B_l||_2 \approx 2^{-L/2} \tag{33}$$

Scale invariant optimizers cannot recover from such directional distortion in layerwise projection/rescaling B_l. This led the literature to prefer non-disrupting locations for LN such as $h_{l+1} = h_l + f_l(\mathrm{LN}_l(h_l))$ ("pre-LN"). However, due to the desirable property of post-LN (that every layer is explicitly normalized), researchers have considered keeping it and fixing the depth dependence of (33) by scaling the residual stream *up* inside LN (Wang *et al.*, 2024):

$$h_{l+1} = \mathrm{LN}_l(\alpha h_l + f_l(h_l)) \tag{34}$$

Under (34),

$$h'_1 = \prod_{l=1}^{L} (\alpha I_d + J_{f_l})\, B_l h'_{L+1} = \boxed{\left(\prod_{l=1}^{L} \alpha B_l\right) h'_{L+1}} + (\text{other terms})$$

Thus the new disrupting transformation $B(\alpha) = \prod_{l=1}^{L} \alpha B_l$ can now be controlled through $\alpha > 0$. Specifically, under similar assumptions as before (Lemma E.5),

$$||B(\alpha)||_2 \leq \prod_{l=1}^{L} \alpha\, ||B_l||_2 \approx \left(1 + \frac{1}{\alpha^2}\right)^{-L/2} \tag{35}$$

thus it is natural to scale $\alpha = L$ so that $||B(L)||_2 \to 1$ as $L \to \infty$ (using the bound as a proxy for the actual scale). This framework admits pre-LN, yielding the final form $h_{l+1} = \mathrm{LN}_l(L \cdot h_l + f_l(\mathrm{LN}_{l,\mathrm{pre}}(h_l)))$ of KEEL (Chen and Wei, 2026).

References

Chen, C. and Wei, L. (2026). Post-layernorm is back: Stable, expressive, and deep. *arXiv preprint arXiv:2601.19895*.

Dey, N., Zhang, B. C., Noci, L., Li, M., Bordelon, B., Bergsma, S., Pehlevan, C., Hanin, B., and Hestness, J. (2025). Don't be lazy: Completep enables compute-efficient deep transformers. *arXiv preprint arXiv:2505.01618*.

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.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, **31**.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, **115**(33), E7665–E7671.

Taleb, N. N. (2016). *Fooled by randomness: The hidden role of chance in life and in the markets*. Editeurs divers USA.

Wang, H., Ma, S., Dong, L., Huang, S., Zhang, D., and Wei, F. (2024). Deepnet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**(10), 6761–6774.

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

Yang, G., Yu, D., Zhu, C., and Hayou, S. (2023). Tensor programs vi: Feature learning in infinite-depth neural networks. *arXiv preprint arXiv:2310.02244*.

A Standard Parameterization (SP)

An L-layer transformer without attention, normalization, and residual connections is an MLP 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

$$\begin{aligned} h_0 &= x & h'_{L+2} &= \text{softmax}(h_{L+2}) - y \\ h_{l+1} &= W_l h_l \quad \forall l = 0 \dots L+1 & h'_l &= W_l^\top h'_{l+1} \quad \forall l = L+1 \dots 1 \\ & & W'_l &= h'_{l+1} h_l^\top \quad \forall l = L+1 \dots 0 \end{aligned} \tag{36}$$

A.1 Initialization

We assume that the weights $W_{l,i,j}$ are iid with zero mean and variance $\sigma_l^2 > 0$. Let $\text{Var}(X)$ 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 E.1)

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

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

A.1.1 Hidden layers

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

$$\text{Var}(h_l) = \sigma_0^2 \qquad \forall l = 1 \dots L+2 \tag{37}$$

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

$$\text{Var}(h'_l) \asymp \sigma_{L+1}^2 \qquad \forall l = L \dots 1 \tag{38}$$

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 WLOG that $d = d_1 = \cdots d_{L+1}$ and use $\sigma_l^2 = 1/d$ for $l = 1 \dots L$.

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 $\text{Var}\left(h'_{L+2}\right) \approx \Theta(\sigma_{L+1}^2 d_{L+1} \text{Var}(h_{L+1}))$ when $h_{L+2} \approx 0$. When RMSNorm is ✓, we assume $h_l = \text{RMSNorm}(W_{l-1}h_{l-1})$ for all layers except $l = L+2$. Most studies assume the first row for SP (36), which makes activations unit order without RMSNorm.

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} (37) but shrinks the activation gradients (38). The choices of σ_0^2 and σ_{L+1}^2 together control $\text{Var}\left(W'_l\right)$. 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 $X \mapsto \text{RMSNorm}(X) = X/\text{RMS}(X)$ 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 $1/\text{RMS}(X)$ in the backward pass (Lemma E.2). For illustration, consider incorporating RMSNorm as $h_l = \text{RMSNorm}(W_{l-1}h_{l-1})$ for all layers except $l = L+2$. It turns out that (Lemma E.3)

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 $O_l = \textbf{OPT}(W'_l)$ 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 $h_{l+1}^{\text{new}} = (W_l + \Delta W_l)(h_l + \Delta h_l)$, so we have

$$\begin{aligned} \Delta h_1^{\text{new}} &= -\eta_0 O_{0,:,i} && x_i = 1 \\ \Delta h_{l+1}^{\text{new}} &= \underbrace{W_l \Delta h_l}_{①} + \underbrace{(-\eta_l O_l h_l^{\text{new}})}_{②} && \forall l = 1 \dots L+1 \end{aligned} \tag{39}$$

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 $\text{RMS}(\Delta h_1^{\text{new}}) \asymp 1$ by setting $\eta_0 \asymp O_0^{-1}$. Unfortunately in (39), ① is not controllable by the learning rate. To make analysis tractable, we enforce the following conditions.

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

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

Condition A.1 is relatively mild given that it holds at initialization.[6] Condition A.2, however, is not easily justifiable for the readout layer intuitively because RMS normalizes by the constant vocab size V.[7] We will come back to this issue in muP (Appendix B) and assume both Condition A.1 and A.2 hold for SP.

A.2.1 Learning rates

(2) has the entry $(2)_i = -\eta_l A_{l,i}$ where $A_{l,i} = \sum_{j=1}^{d} O_{l,i,j} h_{l,j}^{\text{new}}$ is the dot product between optimizer updates and activations. Assuming $O_{l,i,j} = \Theta(1)$ (e.g., Adam) and the variance of the sum over width grows linearly, we have (Appendix D)

$$A_{l,i} = d\mu_{l,i} + \Theta(\sqrt{d})$$

where $\mu_{l,i} = (1/d)\sum_{j=1}^{d} \mathbf{E}[O_{l,i,j} h_{l,j}^{\text{new}}]$ measures if updates and activations typically drift together. If $\mu_{l,i} \neq 0$ (aligned), then $A_{l,i} = \Theta(d)$; $A_{l,i} = \Theta(\sqrt{d})$ otherwise. 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.

Lemma 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

$$\eta_0 = \Theta(1) \qquad \eta_l = \begin{cases} \Theta(1/d) & \text{if Adam is aligned} \\ \Theta(1/\sqrt{d}) & \text{otherwise} \end{cases} \qquad \forall l = 1 \dots L+1 \tag{41}$$

Then the following invariants hold at all training steps:

$$\text{RMS}(W_0) = \Theta(1) \tag{42}$$
$$\text{RMS}(W_l) = \Theta(1/\sqrt{d}) \qquad \forall l = 1 \dots L+1 \tag{43}$$
$$\text{RMS}(h_l) = \Theta(1) \qquad \forall l = 1 \dots L+2 \tag{44}$$
$$\text{RMS}(h'_{L+2}) = \Theta(1) \tag{45}$$
$$\text{RMS}(h'_l) = \Theta(1/\sqrt{d}) \qquad \forall l = L+1 \dots 1 \tag{46}$$
$$\text{RMS}(W'_{L+1}) = \Theta(1) \tag{47}$$
$$\text{RMS}(W'_l) = \Theta(1/\sqrt{d}) \qquad \forall l = L \dots 0 \tag{48}$$

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 (43–48) hold and consider a new forward/backward pass. Adam specifies $\Delta W_{l,i,j} = -\eta_l \textbf{sign}(W'_{l,i,j}) = \Theta(\eta_l)$. 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 (42) 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 (43) 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 $\Delta h_1 = \Delta W_0 x = \text{col}(\Delta W_0)$ 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

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

[6] 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 $||W_l||_2 \asymp 1$.

[7] For non-readout layers, Condition A.1 implies Condition A.2 since

$$\text{RMS}(W_l \Delta h_l) = \frac{||W_l \Delta h_l||_2}{\sqrt{d_{l+1}}} \leq ||W_l||_2 \frac{||\Delta h_l||_2}{\sqrt{d}} \asymp 1 \cdot 1 = 1 \qquad \forall l = 1 \dots L \tag{40}$$

where we inductively assume $\text{RMS}(\Delta h_l) \asymp 1$ (i.e., $||\Delta h_l||_2 = \sqrt{d}$). This breaks at $l = L+1$ since $d_{l+1} = O(1)$ so that the bound becomes $\Theta(\sqrt{d})$.

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 (41), this is $\Theta(1)$ always. Thus (44) is maintained.

For the activation gradients, (45) 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

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

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 (46) 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 (47) and (48) are maintained.

□

A.2.2 Effect of weight decay

With (decoupled) weight decay the weight change becomes

$$\Delta W_l = -\eta_l O_l - \lambda_l W_l$$

where $\lambda_l > 0$ is assumed to be close to zero. Under momentumless Adam with iid centered gradients, the noisy contraction $W_l \mapsto (1 - \lambda_l)W_l - \eta_l O_l$ converges to the stationary per-element size

$$|W_l| = \Theta\left(\frac{\eta_l}{\sqrt{\lambda_l}}\right) \tag{49}$$

Recall that keeping $\mathrm{RMS}(W_l) = \Theta(1/\sqrt{d})$ is a necessary condition for activation stability to keep $\mathrm{Var}\left(h_{l+1}\right) \approx d\mathrm{RMS}(W_l)^2\mathrm{Var}\left(h_l\right) = \Theta(1)$. (49) suggests that there may be a conflict between this condition and the stationary size in the infinite-horizon regime. To resolve the conflict, we may use

$$\lambda_l = \begin{cases} \Theta(1/d) & \text{if } \eta_l = \Theta(1/d) \\ \Theta(1) & \text{if } \eta_l = \Theta(1/\sqrt{d}) \end{cases}$$

But real training is finite-horizon and other regularizing factors like LR schedules, RMSNorm, and residuals, so it is unclear how serious this conflict is.

Coupled case. The literature often assumes the coupled weight decay $\Delta W_l = -\eta_l O_l - \lambda_l^{\text{eff}} W_l$ in the original AdamW formulation where

$$\lambda_l^{\text{eff}} := \eta_l \lambda_l$$

Let us assume $\eta_l = \Theta(1/d)$ for simplicity. In this context, we can consider orthogonal strategies for scale invariance, depending on what object we want to maintain as invariant.

1. If we want $\lambda_l^{\text{eff}} = \Theta(1)$ (i.e., weight decay does not turn off as we increase width), we should set $\lambda_l = \Theta(d)$.

2. If we want the infinite-horizon weights (49) to remain $\Theta(1/\sqrt{d})$, we should set $\lambda_l = \Theta(1)$.

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 (50)	1	$1/d$	1	$\sqrt{d}$	1	$1/\sqrt{d}$	1	$1/\sqrt{d}$	$1/\sqrt{d}$
SP+readout (50)	1	$1/d$	1	1	1	$1/d$	$1/\sqrt{d}$	$1/d$	$1/d$
SP+emb/readout (50)	$1/d$	$1/d$	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 (50). 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).

B muP

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

$$\mathrm{RMS}(W_{L+1}\Delta h_{L+1}) = \Theta(\sqrt{d}) \tag{50}$$

One justification for (50) 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 $uh_{L+1}^\top$ causing W_{L+1} and Δh_{L+1} to be aligned. Since this component ((1) in (39)) is not controllable by the learning rate, the only choice we have in order to make $\mathrm{RMS}(\Delta h_{L+2}) = \Theta(1)$ is to *scale* the readout layer by $1/\sqrt{d}$. This changes the forward and backward passes as

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

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

$$\begin{aligned} u_l &= W_{l-1}h_{l-1} \in \mathbb{R}^{d_l} \\ h_l &= \phi_l(u_l) \in \mathbb{R}^{d_l} \end{aligned} \tag{51}$$

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

$$\begin{aligned} u'_{l+1} &= \phi'_{l+1}(u_{l+1}) \odot h'_{l+1} \in \mathbb{R}^{d_{l+1}} \\ h'_l &= W_l^\top u'_{l+1} \in \mathbb{R}^{d_l} \\ W'_l &= u'_{l+1}h_l^\top \in \mathbb{R}^{d_{l+1} \times d_l} \end{aligned} \tag{52}$$

We assume that ϕ_l is Λ-Lipschitz $|\phi_l(a) - \phi_l(b)| \le \Lambda\,|a - b|$ for some constant $\Lambda > 0$. Then $|\phi'_l| \le \Lambda$ (this holds at kinks using sub-gradients), so when we view (51) and (52) 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

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

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

D Asymptotic Behavior of the Sum of Random Variables

Let $Y_n = \sum_{i=1}^n X_i$ with $\sigma_n^2 = \mathrm{Var}\,(Y_n)$. We can always write

$$Y_n = n\mu_n + A_n$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[X_i]$ and $A_n = \sum_{i=1}^n X_i - \mathbf{E}[X_i]$. Since $\mathbf{E}[A_n] = 0$ and $\mathrm{Var}\,(A_n) = \sigma_n^2$, it follows immediately from Chebyshev that A_n is uniformly bounded by a constant multiple of σ_n in probability. In particular, we can write

$$Y_n = n\mu_n + O_p(\sigma_n) \tag{53}$$

where the probilistic asymptote is in $n \to \infty$. Thus (53) holds for all random variables. If we want to make a stronger statement, we may mildly assume that $A_n/\sigma_n \not\to 0$ in probability which rules out degenerate situations where the fluctuation of A_n becomes smaller than its own standard deviation. This implies $A_n \neq o_p(\sigma_n)$, thus

$$Y_n = n\mu_n + \Theta_p(\sigma_n) \tag{54}$$

is assumed to hold in most realistic scenarios.

D.1 IID Case

If $X_1 \dots X_n$ are iid with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ which are constant in n, we have $\mu_n = \mu$ and $\sigma_n^2 = n\sigma^2$ and thus

$$Y_n = n\mu + \Theta_p(\sqrt{n}) \tag{55}$$

We have two asymptotic regimes.

- $Y_n = \Theta(n)$ if $\mu \neq 0$
- $Y_n = \Theta_p(\sqrt{n})$ if $\mu = 0$

In words, as long as an actual drift "exists" at all, it will win over any constant noise (no matter how large it is) in the long run. This is a typical model in finance (Taleb, 2016). Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$ denote a random return of a portfolio at any given trading hour in a year, with $\mu = 0.000075$ and $\sigma = 0.00224$ so that $\mathrm{sign}\,(X_i)$ is nearly random. Assuming iid, the annual return has the behavior $Y_n = 0.15 \pm 0.1$ which is positive with 93% chance (using the law $Y_n \sim \mathcal{N}(0.15, 0.01)$).

D.2 Non-IID Case

In general, the "linear-drift, square-root fluctuation" picture (55) may not hold. We can easily come up with counterexamples:

- X_i is strongly correlated with each other. E.g., if $X_i = X$ for some variable X with constant variance, $\mathrm{Var}\,(Y_n) = \mathrm{Var}\,(nX) = \Theta(n^2)$ so the fluctuation becomes linear as well and $Y_n = \Theta_p(n)$ regardless of $\mu = 0$.
- Even if each X_i is independent and has constant variance, its distribution may have a dependence on n. E.g., if $\mathbf{E}[X_i] = n$, then $Y_n = n^2 + \Theta_p(\sqrt{n}) = \Theta(n^2)$ drifts quadratically.

Nonetheless, we often assume X_i is "iid enough" (after some normalization) so that we can just invoke the form (55). This is used in both width and depth scaling in this note.

D.2.1 Width scaling

In this context, $Y_n = u^\top v \in \mathbb{R}$ is a dot product between random vectors $u, v \in \mathbb{R}^n$ with RMS $\Theta(1)$ (this can be always enforced if not true already, e.g., by taking $u \mapsto u/\text{RMS}(u)$) whose length is growing in width n. Thus $X_i = u_i v_i$. The implicit assumptions are:

- The covariance terms do not add up faster than linearly. At initialization, all dimensions are literally independent. They are assumed to remain weakly correlated throughout training (e.g., because the network learns "orthogonal features").
- $\text{RMS}(u) = \Theta(1)$ implies $u_i \approx \Theta(1)$ for all i (i.e., every coordinate is "typical"). Technically, the premise only gives us $\sum_{i=1}^n u_i^2 = \Theta(n)$, so we may have, e.g., $u_1 = \Theta(n)$ and $u_2 = \Theta(1/n)$. We simply assume that this manner of crazy concentration never happens.

Under the assumptions, we have $\text{Var}(X_i) = \Theta(1)$ and can say $Y_n \approx n\mu + \Theta_p(\sqrt{n})$ where

$$\mu = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[u_i v_i] = O(1)$$

The condition $\mu \neq 0$ is called **alignment** between u, v to emphasize that their elementwise product is typically drifting in some direction. Thus Y_n grows like $\Theta(n)$ if u, v are aligned, $\Theta_p(\sqrt{n})$ otherwise.

D.2.2 Depth scaling

In this context, $Y_n = \sum_{i=1}^n X_i \in \mathbb{R}$ is the activation after n residual layers. Since X_i is usually a direct transformation of X_{i-1}, these terms are decidedly dependent. But each term is often normalized (e.g., to RMS 1). Assuming the same weak correlation, we may still invoke the form $Y_n \approx n\mu + \Theta_p(\sqrt{n})$. By network symmetries, it is often assumed that $\mu = 0$ (i.e., a "typical activation" has mean zero), and thus $Y_n = \Theta_p(\sqrt{n})$ (i.e., square-root growth in depth).

E Lemmas

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

- *(Activations)*: $\mathbf{E}[h_l] = 0_{d_l}$ and $\text{Cov}(h_l) = \sigma_{l-1}^2 \mathbf{E}[||h_{l-1}||^2] I_{d_l}$ for $l = 1 \dots L+2$.
- *(Logit gradient)*: $\mathbf{E}[h'_{L+2}] = (1/V)1_V - y$. A first-order approximation of $\text{Cov}\left(h'_{L+2}\right) = \text{Cov}(\text{softmax}(h_{L+2}))$ 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 $\text{Cov}(h'_l) = \sigma_l^2 \mathbf{E}[||h'_{l+1}||^2] I_{d_l}$ for $l = L+1 \dots 1$.
- *(Weight gradients)*: $\mathbf{E}[W'_l] = 0_{d_{l+1} \times d_l}$ and $\text{Var}(W'_{l,i,j}) = \mathbf{E}[(h'_{l+1,i})^2]\mathbf{E}[h_{l,j}^2]$ 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 $\text{Cov}(h_l) = \mathbf{E}[h_l h_l^\top] = \mathbf{E}[W_{l-1}h_{l-1}h_{l-1}^\top W_{l-1}^\top]$ 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 $p = \text{softmax}(h_{L+2})$. 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 $\mathbf{E}[p_i] = 1/V$.[8] Thus $\mathbf{E}[h'_{L+2}] = \mathbf{E}[p] - y = (1/V)1_V - y$. Since $\text{Cov}\left(h'_{L+2}\right) = \text{Cov}(p)$ and the covariance of random variables bounded in $[0,1]$

[8] 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 $\mathbf{E}[\text{softmax}(h_{L+2})] = \text{softmax}(\mathbf{E}[h_{L+2}])$ (not true in general) and Jensen's inequality (exact only for constants and linear functions).

cannot exceed 1/4, each entry is accordingly bounded. Let $J := \nabla_h \text{softmax}(h)|_{h=0_V} = (1/V)I_V - (1/V^2)1_V 1_V^\top$ 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 $\text{Cov}\left(h'_{L+2}\right) = \text{Cov}\left(p\right) \approx \text{Cov}\left(\hat{p}\right) = J\text{Cov}\left(h_{L+2}\right)J^\top = \sigma_{L+1}^2 \mathbf{E}[||h_{L+1}||^2]JJ^\top$ 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 $\text{Cov}(h'_{l,i}, h'_{l,j}) = \mathbf{E}[h'_{l,i}h'_{l,j}] = \sum_{k,t} \mathbf{E}[W_{l,k,i}W_{l,t,j}h'_{l+1,k}h'_{l+1,t}] = \sum_{k,t} \mathbf{E}[W_{l,k,i}W_{l,t,j}\tilde{h}'_{l+1,k}\tilde{h}'_{l+1,t}] = \sum_{k,t} \mathbf{E}[W_{l,k,i}W_{l,t,j}]\mathbf{E}[\tilde{h}'_{l+1,k}\tilde{h}'_{l+1,t}]$. 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 $\text{Cov}(W'_{l,i,j}, W'_{l,k,t}) = \mathbf{E}[W'_{l,i,j}W'_{l,k,t}] = \mathbf{E}[h'_{l+1,i}h'_{l+1,k}h_{l,j}h_{l,t}] = \mathbf{E}[\tilde{h}'_{l+1,i}\tilde{h}'_{l+1,k}h_{l,j}h_{l,t}] = \mathbf{E}[h'_{l+1,i}h'_{l+1,k}]\mathbf{E}[h_{l,j}h_{l,t}]$. 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 E.2. Let

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

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

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

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

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

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

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

□

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

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

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:

- $\text{Var}\left(h_l\right) = 1$ for $l = 1 \dots L+1$ and $\text{Var}\left(h_{L+2}\right) = \Omega(1)$.
- $\mathbf{E}[\|h'_{L+2}\|^2] = \Theta(1)$, $\text{Var}\left(W'_{L+1}\right) = \Theta(1)$, and $\text{Var}\left(h'_{L+1}\right) = \Theta(\sigma_{L+1}^2)$.
- $\text{Var}\left(h'_l\right) = \Theta(\sigma_{L+1}^2)$ and $\text{Var}\left(W'_l\right) = \Theta(\frac{\sigma_{L+1}^2}{\sigma_l^2 d})$ for $l = L \dots 1$.
- $\text{Var}\left(W'_0\right) = \Theta(\frac{\sigma_{L+1}^2}{\sigma_0^2})$.

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

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

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

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

At initialization $\text{Var}\left(h''_l\right) = \Theta(\text{Var}\left(h'_l\right))$.[9] 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:

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

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

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

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

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

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

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

□

Lemma E.4. Under (1) and (2), (3–5) hold.

Proof. For the forward pass, the base case is

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

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

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

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

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

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

For $l = L \dots 1$,

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

Likewise for the weight gradients, the base case is

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

For $l = L \dots 1$,

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

Finally for $l = 0$,

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

□

Lemma E.5. Let $o = \alpha h + f(h)$ under the same setup in Lemma E.6. For $\tau \in \mathbb{R}^d$ with $\tau^{\max} \approx \gamma^{\max}$,

$$\left|\left|\frac{\partial \text{LN}_\tau(o)}{\partial o}\right|\right|_2 \lesssim \frac{1}{\sqrt{\alpha^2+1}}$$

Proof. Let $\Pi = I_d - oo^\top / ||o||^2$ denote the projection onto the orthogonal complement of span $(o) \subset \mathbb{R}$. Then

$$\begin{aligned}
\left|\left|\frac{\partial \text{LN}_\tau(o)}{\partial o}\right|\right|_2 &= \left|\left|\frac{1}{\text{RMS}(o)}\Pi\text{diag}\left(\tau\right)\right|\right|_2 && (\text{transposed Jacobian convention}) \\
&\leq \frac{||\Pi||_2 \, ||\text{diag}\left(\tau\right)||_2}{\text{RMS}(o)} \\
&= \frac{||\tau||_\infty}{\text{RMS}(o)} && (\Pi \text{ is a projection}) \\
&\approx \frac{\tau^{\max}}{\sqrt{\alpha^2+1}\gamma^{\max}} && (\text{Lemma E.6}) \\
&\approx \frac{1}{\sqrt{\alpha^2+1}} && (\text{using } \tau^{\max} \approx \gamma^{\max})
\end{aligned}$$

□

Lemma E.6. Let $o = \alpha h + f(h)$ where $\alpha > 0$ and $h = \mathrm{LN}_\gamma(u)$ for some u. If

- $h^\top f(h) \approx 0$
- $||f(h)|| \approx ||h||$
- The gating parameter $\gamma \in \mathbb{R}^d$ of $\mathrm{LN}_\gamma(u) := \frac{u}{\mathrm{RMS}(u)}\mathrm{diag}\,(\gamma)$ is roughly uniform: $\gamma \approx \gamma^{\max} I_d$.

then

$$\mathrm{RMS}(o) \approx \sqrt{\alpha^2 + 1}\gamma^{\max}$$

Proof.

$$\begin{aligned}
||o||^2 &= \alpha^2 ||h||^2 + ||f(h)||^2 + 2h^\top f(h) \\
&\approx \alpha^2 ||h||^2 + ||f(h)||^2 && \text{(using } h^\top f(h) \approx 0\text{)} \\
&\approx (\alpha^2 + 1) ||h||^2 && \text{(using } ||f(h)|| \approx ||h||\text{)} \\
&= (\alpha^2 + 1) \left|\left| \frac{u}{\mathrm{RMS}(u)}\mathrm{diag}\,(\gamma) \right|\right|^2 \\
&\approx (\alpha^2 + 1)(\gamma^{\max})^2 \left|\left| \frac{u}{\mathrm{RMS}(u)} \right|\right|^2 && \text{(using } \gamma_i \approx \gamma^{\max}\text{)} \\
&= (\alpha^2 + 1) d (\gamma^{\max})^2
\end{aligned}$$

This implies

$$\mathrm{RMS}(o) = \frac{||o||}{\sqrt{d}} \approx \frac{\sqrt{\alpha^2 + 1}\sqrt{d}\gamma^{\max}}{\sqrt{d}} = \sqrt{\alpha^2 + 1}\gamma^{\max}$$

□