On Layer Normalization in the Transformer Architecture

Xiong et al. · ICML 2020

CS 886 — Learning Theory for Modern AI: Transformers and LLMs
University of Waterloo

Gavin Hull, Hanzhe Wu · October 2, 2026

Warm-up, LayerNorm, and gradient scale

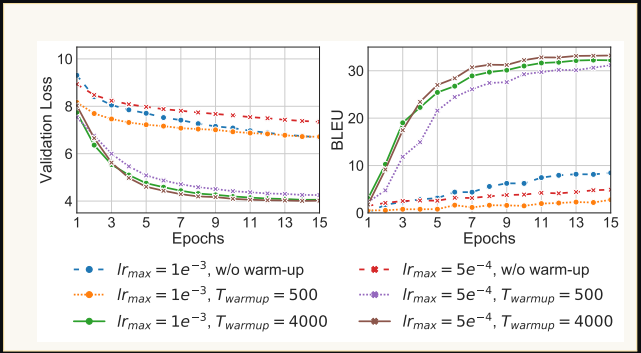
Why do Transformers rely on learning-rate warm-up?

The authors compare two architectures at initialization:

$$\text{LayerNorm placement} \longrightarrow \text{forward scale} \longrightarrow \text{gradient scale.}$$

They derive depth-dependent gradient bounds, then test the prediction experimentally.

Warm-up matters for Transformer training



Canonical warm-up in this comparison: 4000 steps.

Transformer is sensitive to warm-up.

Short or absent warm-up can perform poorly; 4000 steps is the canonical baseline here.

Can LN placement make warm-up unnecessary?

Why normalize activations?

When early layers update, the representations they send to later layers change.

Why normalize activations?

When early layers update, the representations they send to later layers change.

Later layers must therefore keep adapting to a moving input distribution during training.

Why normalize activations?

When early layers update, the representations they send to later layers change.

Later layers must therefore keep adapting to a moving input distribution during training.

Normalization helps keep these intermediate representations on a controlled scale.

Batch normalization



$H \in \mathbb{R}^{B \times d}$

One feature across the minibatch.

The statistics depend on the other examples. This couples samples and makes the minibatch estimate noisy.

Dependence on batch statistics is awkward in recurrent and variable-length sequence settings.

Layer normalization



$H \in \mathbb{R}^{B \times d}$

Features of one representation.

$$\mu = \frac{1}{d}\sum_{j=1}^{d} x_j, \quad \sigma = \sqrt{\frac{1}{d}\sum_{j=1}^{d}(x_j - \mu)^2}.$$

$$\mathrm{LN}(x) = \gamma \odot \frac{x - \mu \mathbf{1}}{\sigma} + \beta.$$

General architecture: γ and β are learnable.

Centering and rescaling

$$\mathbf{1} = (1, \ldots, 1), \qquad P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}, \qquad y = xP = x - \mu\mathbf{1}.$$

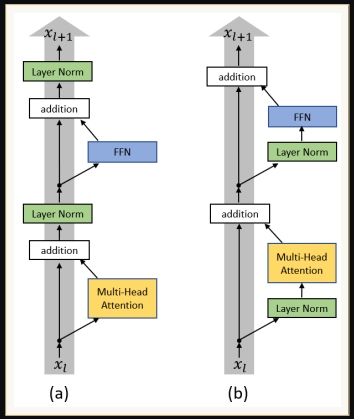
At the analysis initialization, $\gamma = \mathbf{1}$ and $\beta = 0$:

$$\mathrm{LN}(x) = \sqrt{d}\,\frac{y}{\|y\|_2}.$$

$$\mathbf{1}\,\mathrm{LN}(x)^T = 0.$$

$$\|\mathrm{LN}(x)\|_2^2 = d.$$

Post-LN and Pre-LN



(a) Post-LN

$\mathrm{LN}(x + F(x))$

(b) Pre-LN

$x + F(\mathrm{LN}(x))$

Pre-LN also applies a final LN after the stack of L blocks.

Figure 1. Arrows run upward. The analysis below uses one attention head.

Post-LN recursion

$$p_k := x_{\ell,i}^{\text{post},k}$$

$$p_1 = \text{Att}(x_{\ell,i}^{\text{post}}, [x_{\ell,j}^{\text{post}}]_{j=1}^{n}),$$

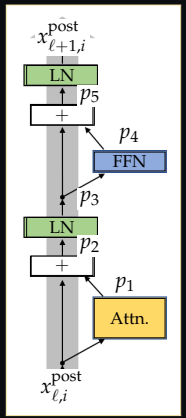
$$p_2 = x_{\ell,i}^{\text{post}} + p_1,$$

$$p_3 = \text{LN}(p_2),$$

$$p_4 = \text{ReLU}(p_3 W_{1,\ell} + b_{1,\ell}) W_{2,\ell} + b_{2,\ell},$$

$$p_5 = p_3 + p_4,$$

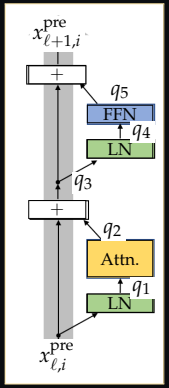
$$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5).$$



Pre-LN recursion

$q_k := x^{\text{pre,k}}_{\ell,i}$

$$
\begin{aligned}
q_1 &= \mathrm{LN}(x^{\text{pre}}_{\ell,i}), \\
q_2 &= \mathrm{Att}(q_1, [q_{1,j}]_{j=1}^{n}), \\
q_3 &= x^{\text{pre}}_{\ell,i} + q_2, \\
q_4 &= \mathrm{LN}(q_3), \\
q_5 &= \mathrm{ReLU}(q_4 W_{1,\ell} + b_{1,\ell}) W_{2,\ell} + b_{2,\ell}, \\
x^{\text{pre}}_{\ell+1,i} &= q_3 + q_5, \\
x^{\text{pre}}_{\text{Final},i} &= \mathrm{LN}(x^{\text{pre}}_{L+1,i}).
\end{aligned}
$$



Final output LN is outside the repeated block.

Analytic setup

Model

Single-head attention.

$d \times d$ matrices.

ReLU FFN.

Initialization

Independent Gaussian Xavier weights:

$W_{ab} \sim N(0, 1/d)$.

Zero biases; $\gamma = 1, \beta = 0$; Gaussian input.

Attention choice

$W_{Q,\ell} = W_{K,\ell} = 0$.

All logits vanish, $\alpha_{ij} = 1/n$.

$q_2 = (\frac{1}{n} \sum_j q_{1,j}) W_{V,\ell}$.

What the theory aims to explain

How does depth L affect the final FFN gradient at initialization?

Here d is the hidden width and $\widetilde{\mathcal{L}}$ is the mean loss.

Post-LN

$$\left\| \frac{\partial \widetilde{\mathcal{L}}}{\partial W_{2,L}} \right\|_F = O\left(d\sqrt{\log d}\right).$$

Pre-LN

$$\left\| \frac{\partial \widetilde{\mathcal{L}}}{\partial W_{2,L}} \right\|_F = O\left(d\sqrt{\frac{\log d}{L}}\right).$$

Pre-LN gains an extra $L^{-1/2}$ factor in the gradient upper bound.

Why gradient scale matters for warm-up

For gradient descent $W_{2,L}^{\text{new}} \leftarrow W_{2,L}^{\text{old}} - \eta \nabla_{W_{2,L}} \widetilde{\mathcal{L}}$ with learning rate η:

$$\Delta W_{2,L} = -\eta \nabla_{W_{2,L}} \widetilde{\mathcal{L}}, \qquad \|\Delta W_{2,L}\|_F = \eta \left\| \nabla_{W_{2,L}} \widetilde{\mathcal{L}} \right\|_F .$$

Post-LN

Large gradients can cause large initial updates. Warm-up limits their size by starting with a smaller learning rate.

Pre-LN

Accumulating residual scale gives the final LN an extra $L^{-1/2}$ factor, motivating training without warm-up.

Experiments. Pre-LN reaches comparable performance earlier without warm-up in the tested settings. The bounds alone do not guarantee stability.

Theorem 1: proof idea

For a fixed position, the chain rule separates two factors:

$$\text{FFN gradient} \lesssim \text{activation size} \times \text{LN sensitivity}.$$

1. Lemmas 1–2: forward scale. Gaussian ReLU energy determines how the residual second moment changes with depth.
2. Lemma 3: normalization sensitivity. The LN Jacobian is inversely proportional to the centered input norm.
3. Probability and aggregation. Obtain tail control by Lemma 4.2, then average positions and bound the matrix norm.

The incoming loss-gradient norm is assumed $O(1)$. We will check the centering and probability steps explicitly.

Lemma 1: Gaussian ReLU energy

Let $X \sim N(0, \sigma^2 I_d)$, with ReLU applied coordinatewise.

Lemma 1.

$$\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 = \frac{d\sigma^2}{2}.$$

Intuition. By symmetry, ReLU retains half the expected squared norm.

Lemma 1: $\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 = \dfrac{d\sigma^2}{2}$

Proof.

Given $X \sim N(0, \sigma^2 I_d)$,

$$\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 = \sum_{k=1}^{d} \mathbb{E}[\mathrm{ReLU}(X_k)^2] \quad \text{(linearity)}$$

Lemma 1: $\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 = \dfrac{d\sigma^2}{2}$

Proof.

Given $X \sim N(0, \sigma^2 I_d)$,

$$\begin{aligned}\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 &= \sum_{k=1}^{d} \mathbb{E}[\mathrm{ReLU}(X_k)^2] \quad \text{(linearity)} \\ &= d\,\mathbb{E}[X_1^2 \mathbf{1}_{\{X_1 \geq 0\}}] \quad \text{(i.i.d.)}\end{aligned}$$

Lemma 1: $\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 = \dfrac{d\sigma^2}{2}$

Proof.

Given $X \sim N(0, \sigma^2 I_d)$,

$$\begin{aligned}
\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 &= \sum_{k=1}^{d} \mathbb{E}[\mathrm{ReLU}(X_k)^2] \quad \text{(linearity)} \\
&= d\,\mathbb{E}[X_1^2 \mathbf{1}_{\{X_1 \geq 0\}}] \quad \text{(i.i.d.)} \\
&= \frac{d}{2}\mathbb{E}[X_1^2] \quad \text{(symmetry)}
\end{aligned}$$

Lemma 1: $\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 = \dfrac{d\sigma^2}{2}$

Proof.

Given $X \sim N(0, \sigma^2 I_d)$,

$$\begin{aligned}
\mathbb{E}\,\|\mathrm{ReLU}(X)\|_2^2 &= \sum_{k=1}^{d} \mathbb{E}[\mathrm{ReLU}(X_k)^2] \quad \text{(linearity)} \\
&= d\,\mathbb{E}[X_1^2 \mathbf{1}_{\{X_1 \geq 0\}}] \quad \text{(i.i.d.)} \\
&= \frac{d}{2}\mathbb{E}[X_1^2] \quad \text{(symmetry)} \\
&= \frac{d}{2}\sigma^2. \qquad \square
\end{aligned}$$

Lemma 2: forward scale

Post-LN

$$\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d.$$

Pre-LN

$$\left(1+\frac{\ell}{2}\right)d \le \mathbb{E}\left\|x_{\ell,i}^{\mathrm{pre}}\right\|_2^2 \le \left(1+\frac{3\ell}{2}\right)d.$$

Why it matters: forward scale controls the final LayerNorm Jacobian, which then enters the gradient bound.

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

Proof.

$$\|p_3\|_2^2 = \|\mathrm{LN}(p_2)\|_2^2$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

Proof.

$$\begin{aligned}\|p_3\|_2^2 &= \|\mathrm{LN}(p_2)\|_2^2 \\ &= \|(p_2 - \mu\mathbf{1})/\sigma\|_2^2\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

Proof.

$$\begin{aligned}\|p_3\|_2^2 &= \|\mathrm{LN}(p_2)\|_2^2 \\ &= \|(p_2 - \mu\mathbf{1})/\sigma\|_2^2 \\ &= \sum_{k=1}^{d}\left(\frac{(p_2)_k - \mu}{\sigma}\right)^2\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

Proof.

$$
\begin{aligned}
\|p_3\|_2^2 &= \|\mathrm{LN}(p_2)\|_2^2 \\
&= \|(p_2 - \mu\mathbf{1})/\sigma\|_2^2 \\
&= \sum_{k=1}^{d} \left(\frac{(p_2)_k - \mu}{\sigma}\right)^2 \\
&= \frac{d}{\sigma^2}\left(\frac{1}{d}\sum_{k=1}^{d}((p_2)_k - \mu)^2\right)
\end{aligned}
$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

Proof.

$$
\begin{aligned}
\|p_3\|_2^2 &= \|\mathrm{LN}(p_2)\|_2^2 \\
&= \|(p_2 - \mu\mathbf{1})/\sigma\|_2^2 \\
&= \sum_{k=1}^{d} \left(\frac{(p_2)_k - \mu}{\sigma}\right)^2 \\
&= \frac{d}{\sigma^2}\left(\frac{1}{d}\sum_{k=1}^{d}((p_2)_k - \mu)^2\right) \\
&= d
\end{aligned}
$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

$$\mathbb{E}\,\|p_4\|_2^2 = \mathbb{E}\left\|uW_{2,\ell}\right\|_2^2$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

$$\begin{aligned}\mathbb{E}\,\|p_4\|_2^2 &= \mathbb{E}\,\left\|uW_{2,\ell}\right\|_2^2 \\ &= \mathbb{E}\left[u\,\mathbb{E}[W_{2,\ell}W_{2,\ell}^T]u^T\right] \quad \text{(independence)}\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \mathrm{Att}(\ldots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

$$\begin{aligned}\mathbb{E}\,\|p_4\|_2^2 &= \mathbb{E}\left\|uW_{2,\ell}\right\|_2^2 \\ &= \mathbb{E}\left[u\,\mathbb{E}[W_{2,\ell}W_{2,\ell}^T]u^T\right] \quad \text{(independence)} \\ &= \mathbb{E}\,\|u\|_2^2 \quad \text{(Xavier covariance)}\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \mathrm{Att}(\ldots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

$$\begin{aligned}
\mathbb{E}\,\|p_4\|_2^2 &= \mathbb{E}\,\left\|uW_{2,\ell}\right\|_2^2 \\
&= \mathbb{E}\left[u\,\mathbb{E}[W_{2,\ell}W_{2,\ell}^T]u^T\right] \quad \text{(independence)} \\
&= \mathbb{E}\,\|u\|_2^2 \quad \text{(Xavier covariance)} \\
&= d/2. \quad \text{(Lemma 1 and (1))}
\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \mathrm{Att}(\ldots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.
(2) $\mathbb{E}\,\|p_4\|_2^2 = d/2$.

Proof.

$$\mathbb{E}\,\|p_5\|_2^2 = \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 + 2\mathbb{E}[p_3 p_4^T]$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$
$p_1 = \text{Att}(\dots)$
$p_2 = x_{\ell,i}^{\text{post}} + p_1$
$p_3 = \text{LN}(p_2)$
$p_4 = \text{ReLU}(p_3 W_1) W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.
(2) $\mathbb{E}\,\|p_4\|_2^2 = d/2$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

$$\begin{aligned}\mathbb{E}\,\|p_5\|_2^2 &= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 + 2\mathbb{E}[p_3 p_4^T] \\ &= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 + 2\mathbb{E}[p_3\,\mathbb{E}[W_{2,\ell}^T]u^T]\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$
$p_1 = \mathrm{Att}(\ldots)$
$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.
(2) $\mathbb{E}\,\|p_4\|_2^2 = d/2$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

$$\begin{aligned}\mathbb{E}\,\|p_5\|_2^2 &= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 + 2\mathbb{E}[p_3 p_4^T] \\ &= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 + 2\mathbb{E}[p_3\,\mathbb{E}[W_{2,\ell}^T]u^T] \\ &= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 \quad \text{(fresh centered } W_{2,\ell}\text{)}\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$
$p_1 = \mathrm{Att}(\ldots)$
$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Lemma 2(i): $\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d$

(1) $\mathbb{E}\,\|p_3\|_2^2 = d$.
(2) $\mathbb{E}\,\|p_4\|_2^2 = d/2$.

Proof.

Let $u := \mathrm{ReLU}(p_3 W_{1,\ell})$.

$$\begin{aligned}
\mathbb{E}\,\|p_5\|_2^2 &= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 + 2\mathbb{E}[p_3 p_4^T] \\
&= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 + 2\mathbb{E}[p_3\,\mathbb{E}[W_{2,\ell}^T]u^T] \\
&= \mathbb{E}\,\|p_3\|_2^2 + \mathbb{E}\,\|p_4\|_2^2 \quad \text{(fresh centered } W_{2,\ell}\text{)} \\
&= d + d/2 = 3d/2. \quad \text{((1), (2))} \qquad \square
\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$
$p_1 = \mathrm{Att}(\ldots)$
$p_2 = x_{\ell,i}^{\text{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\text{post}} = \mathrm{LN}(p_5)$

Lemma 2(ii):

$$\left(1+\frac{\ell}{2}\right)d \le \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \le \left(1+\frac{3\ell}{2}\right)d$$

Proof.

$$\left\|q_{1,j}\right\|_2^2 = \left\|\text{LN}(x_{\ell,j}^{\text{pre}})\right\|_2^2$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii):

$$\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$$

Proof.

$$\begin{aligned}\left\|q_{1,j}\right\|_2^2 &= \left\|\text{LN}(x_{\ell,j}^{\text{pre}})\right\|_2^2 \\ &= d \quad \text{(LayerNorm identity)} \qquad (1 \leq j \leq n).\end{aligned}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\left\|q_{1,j}\right\|_2^2 = d \quad (1 \leq j \leq n)$.

Proof.

$m_\ell := n^{-1} \sum_j q_{1,j}$, so $q_2 = m_\ell W_{V,\ell}$.

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii):

$$\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$$

(3) $\|q_{1,j}\|_2^2 = d \quad (1 \leq j \leq n).$

Proof.

$m_\ell := n^{-1}\sum_j q_{1,j}$, so $q_2 = m_\ell W_{V,\ell}$.

$$\mathbb{E}\|q_3\|_2^2 = \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 + 2\mathbb{E}[q_0 q_2^T]$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\|q_{1,j}\|_2^2 = d \quad (1 \leq j \leq n)$.

Proof.

$m_\ell := n^{-1} \sum_j q_{1,j}$, so $q_2 = m_\ell W_{V,\ell}$.

$$\begin{aligned}\mathbb{E}\|q_3\|_2^2 &= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 + 2\mathbb{E}[q_0 q_2^T] \\ &= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 + 2\mathbb{E}[q_0 \mathbb{E}[W_{V,\ell}^T \mid q_{0,1:n}] m_\ell^T]\end{aligned}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\|q_{1,j}\|_2^2 = d \quad (1 \leq j \leq n)$.

Proof.

$m_\ell := n^{-1}\sum_j q_{1,j}$, so $q_2 = m_\ell W_{V,\ell}$.

$$\begin{aligned}\mathbb{E}\|q_3\|_2^2 &= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 + 2\mathbb{E}[q_0 q_2^T] \\ &= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 + 2\mathbb{E}[q_0 \mathbb{E}[W_{V,\ell}^T \mid q_{0,1:n}] m_\ell^T] \\ &= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2\end{aligned}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\left\|q_{1,j}\right\|_2^2 = d \quad (1 \leq j \leq n)$.

Proof.

$m_\ell := n^{-1} \sum_j q_{1,j}$, so $q_2 = m_\ell W_{V,\ell}$.

$$\begin{aligned}
\mathbb{E}\left\|q_3\right\|_2^2 &= \mathbb{E}\left\|q_0\right\|_2^2 + \mathbb{E}\left\|q_2\right\|_2^2 + 2\mathbb{E}[q_0 q_2^T] \\
&= \mathbb{E}\left\|q_0\right\|_2^2 + \mathbb{E}\left\|q_2\right\|_2^2 + 2\mathbb{E}[q_0 \mathbb{E}[W_{V,\ell}^T \mid q_{0,1:n}] m_\ell^T] \\
&= \mathbb{E}\left\|q_0\right\|_2^2 + \mathbb{E}\left\|q_2\right\|_2^2 \\
&= \mathbb{E}\left\|q_0\right\|_2^2 + \mathbb{E}[m_\ell \mathbb{E}[W_{V,\ell} W_{V,\ell}^T] m_\ell^T]
\end{aligned}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\|q_{1,j}\|_2^2 = d \quad (1 \leq j \leq n)$.

Proof.

$m_\ell := n^{-1} \sum_j q_{1,j}$, so $q_2 = m_\ell W_{V,\ell}$.

$$\begin{aligned}
\mathbb{E}\|q_3\|_2^2 &= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 + 2\mathbb{E}[q_0 q_2^T] \\
&= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 + 2\mathbb{E}[q_0 \mathbb{E}[W_{V,\ell}^T \mid q_{0,1:n}] m_\ell^T] \\
&= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|q_2\|_2^2 \\
&= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}[m_\ell \mathbb{E}[W_{V,\ell} W_{V,\ell}^T] m_\ell^T] \\
&= \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|m_\ell\|_2^2 .
\end{aligned}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\|q_{1,j}\|_2^2 = d$.

(4) $\mathbb{E}\|q_3\|_2^2 = \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|m_\ell\|_2^2$.

Proof.

$$0 \leq \mathbb{E}\|m_\ell\|_2^2 \leq \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left\|q_{1,j}\right\|_2^2 \quad \text{(Jensen's inequality)}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\|q_{1,j}\|_2^2 = d$.

(4) $\mathbb{E}\|q_3\|_2^2 = \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|m_\ell\|_2^2$.

Proof.

$$0 \leq \mathbb{E}\|m_\ell\|_2^2 \leq \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left\|q_{1,j}\right\|_2^2 \quad \text{(Jensen's inequality)}$$
$$= d \quad \text{((3))}.$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(3) $\|q_{1,j}\|_2^2 = d$.

(4) $\mathbb{E}\|q_3\|_2^2 = \mathbb{E}\|q_0\|_2^2 + \mathbb{E}\|m_\ell\|_2^2$.

Proof.

$$
\begin{aligned}
0 \leq \mathbb{E}\|m_\ell\|_2^2 &\leq \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left\|q_{1,j}\right\|_2^2 \quad \text{(Jensen's inequality)} \\
&= d \quad \text{((3))}. \\
\implies \mathbb{E}\|q_0\|_2^2 &\leq \mathbb{E}\|q_3\|_2^2 \leq \mathbb{E}\|q_0\|_2^2 + d \quad \text{((4))}.
\end{aligned}
$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(5) $\mathbb{E}\,\|q_0\|_2^2 \leq \mathbb{E}\,\|q_3\|_2^2 \leq \mathbb{E}\,\|q_0\|_2^2 + d.$

Proof.

$$\|q_4\|_2^2 = \|\mathrm{LN}(q_3)\|_2^2 = d \quad \text{(LayerNorm identity)}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \mathrm{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right)d \le \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \le \left(1+\frac{3\ell}{2}\right)d$

(5) $\mathbb{E}\|q_0\|_2^2 \le \mathbb{E}\|q_3\|_2^2 \le \mathbb{E}\|q_0\|_2^2 + d.$

Proof.

$$\|q_4\|_2^2 = \|\mathrm{LN}(q_3)\|_2^2 = d \quad \text{(LayerNorm identity)}$$

$$\mathbb{E}\|q_5\|_2^2 = \mathbb{E}\left\|\mathrm{ReLU}(q_4 W_{1,\ell})W_{2,\ell}\right\|_2^2$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1)W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \mathrm{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right)d \le \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \le \left(1+\frac{3\ell}{2}\right)d$

(5) $\mathbb{E}\|q_0\|_2^2 \le \mathbb{E}\|q_3\|_2^2 \le \mathbb{E}\|q_0\|_2^2 + d.$

Proof.

$$\|q_4\|_2^2 = \|\mathrm{LN}(q_3)\|_2^2 = d \quad \text{(LayerNorm identity)}$$

$$\mathbb{E}\|q_5\|_2^2 = \mathbb{E}\left\|\mathrm{ReLU}(q_4 W_{1,\ell})W_{2,\ell}\right\|_2^2 = \frac{d}{2} \quad \text{(Lemma 1 and the FFN calculation in (i)).}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1)W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \mathrm{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right)d \le \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \le \left(1+\frac{3\ell}{2}\right)d$

(5) $\mathbb{E}\|q_0\|_2^2 \le \mathbb{E}\|q_3\|_2^2 \le \mathbb{E}\|q_0\|_2^2 + d.$
(6) $\mathbb{E}\|q_5\|_2^2 = d/2.$

Proof.

$$\mathbb{E}\left\|x_{\ell+1,i}^{\text{pre}}\right\|_2^2 = \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2 + 2\mathbb{E}[q_3 q_5^T]$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}, q_0 = x_{\ell,i}^{\text{pre}}$
$q_1 = \text{LN}(q_0)$
$q_2 = m_\ell W_{V,\ell}$
$q_3 = q_0 + q_2$
$q_4 = \text{LN}(q_3)$
$q_5 = \text{ReLU}(q_4 W_1) W_2$
$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$
$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(5) $\mathbb{E}\|q_0\|_2^2 \leq \mathbb{E}\|q_3\|_2^2 \leq \mathbb{E}\|q_0\|_2^2 + d$.
(6) $\mathbb{E}\|q_5\|_2^2 = d/2$.

Proof.

Let $u := \text{ReLU}(q_4 W_{1,\ell})$.

$$\begin{aligned}\mathbb{E}\left\|x_{\ell+1,i}^{\text{pre}}\right\|_2^2 &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2 + 2\mathbb{E}[q_3 q_5^T] \\ &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2 + 2\mathbb{E}[q_3 \mathbb{E}[W_{2,\ell}^T] u^T]\end{aligned}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$
$q_1 = \text{LN}(q_0)$
$q_2 = m_\ell W_{V,\ell}$
$q_3 = q_0 + q_2$
$q_4 = \text{LN}(q_3)$
$q_5 = \text{ReLU}(q_4 W_1) W_2$
$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$
$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(5) $\mathbb{E}\|q_0\|_2^2 \leq \mathbb{E}\|q_3\|_2^2 \leq \mathbb{E}\|q_0\|_2^2 + d$.
(6) $\mathbb{E}\|q_5\|_2^2 = d/2$.

Proof.

Let $u := \text{ReLU}(q_4 W_{1,\ell})$.

$$\begin{aligned}\mathbb{E}\left\|x_{\ell+1,i}^{\text{pre}}\right\|_2^2 &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2 + 2\mathbb{E}[q_3 q_5^T] \\ &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2 + 2\mathbb{E}[q_3 \mathbb{E}[W_{2,\ell}^T] u^T] \\ &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2\end{aligned}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$
$q_1 = \text{LN}(q_0)$
$q_2 = m_\ell W_{V,\ell}$
$q_3 = q_0 + q_2$
$q_4 = \text{LN}(q_3)$
$q_5 = \text{ReLU}(q_4 W_1) W_2$
$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$
$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right)d \le \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \le \left(1+\frac{3\ell}{2}\right)d$

(5) $\mathbb{E}\|q_0\|_2^2 \le \mathbb{E}\|q_3\|_2^2 \le \mathbb{E}\|q_0\|_2^2 + d$.
(6) $\mathbb{E}\|q_5\|_2^2 = d/2$.

Proof.

Let $u := \text{ReLU}(q_4 W_{1,\ell})$.

$$\begin{aligned}\mathbb{E}\left\|x_{\ell+1,i}^{\text{pre}}\right\|_2^2 &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2 + 2\mathbb{E}[q_3 q_5^T] \\ &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2 + 2\mathbb{E}[q_3 \mathbb{E}[W_{2,\ell}^T] u^T] \\ &= \mathbb{E}\|q_3\|_2^2 + \mathbb{E}\|q_5\|_2^2\end{aligned}$$

$$\therefore \mathbb{E}\|q_0\|_2^2 + d/2 \le \mathbb{E}\left\|x_{\ell+1,i}^{\text{pre}}\right\|_2^2 \le \mathbb{E}\|q_0\|_2^2 + 3d/2. \quad \text{((5), (6))}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$
$q_1 = \text{LN}(q_0)$
$q_2 = m_\ell W_{V,\ell}$
$q_3 = q_0 + q_2$
$q_4 = \text{LN}(q_3)$
$q_5 = \text{ReLU}(q_4 W_1) W_2$
$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$
$x_{\text{Final}} = \text{LN}(x_{L+1})$

Lemma 2(ii): $\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$

(7) $\mathbb{E}\,\|q_0\|_2^2 + d/2 \leq \mathbb{E}\left\|x_{\ell+1,i}^{\text{pre}}\right\|_2^2 \leq \mathbb{E}\,\|q_0\|_2^2 + 3d/2$

(8) $\mathbb{E}\left\|x_{0,i}^{\text{pre}}\right\|_2^2 = d$

Proof.

$\left(1+\frac{\ell}{2}\right) d \leq \mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 \leq \left(1+\frac{3\ell}{2}\right) d$ (induction, (7), (8)). □

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Forward norm growth

Post-LN

$$\mathbb{E}\,\|p_5\|_2^2 = \frac{3}{2}d = \Theta(d).$$

constant scale before the closing LN

Pre-LN

$$\mathbb{E}\left\|x_{\ell,i}^{\text{pre}}\right\|_2^2 = \Theta(\ell d).$$

increasing residual-stream scale

Lemma 3: the paper's statement

For ideal LayerNorm, define

$$J_{\mathrm{LN}}(x) := \frac{\partial \,\mathrm{LN}(x)}{\partial x}.$$

Lemma 3 (as stated). The paper claims

$$\|J_{\mathrm{LN}}(x)\|_2 = O\left(\frac{\sqrt{d}}{\|x\|_2}\right).$$

The denominator needs correction. We derive the exact formula for $xP \neq 0$.

Lemma 3 and Appendix D. For a matrix, $\|\cdot\|_2$ denotes the spectral norm.

Lemma 3: $\|J_{\text{LN}}(x)\|_2 = \sqrt{d} / \|xP\|_2$ for $d \geq 3$

Proof.

Let $P := I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$ and $y := xP$.

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\text{LN}(x) = \sqrt{d}\, y / \|y\|_2$

$\|\text{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d} / \|xP\|_2$ for $d \geq 3$

Proof.

Let $P := I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$ and $y := xP$.

$$\mathrm{LN}(x)_i = \frac{y_i}{\sqrt{\frac{1}{d}\sum_{k=1}^{d} y_k^2}}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y / \|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d} / \|xP\|_2$ for $d \geq 3$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y / \|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Proof.

Let $P := I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$ and $y := xP$.

$$\begin{aligned} \mathrm{LN}(x)_i &= \frac{y_i}{\sqrt{\frac{1}{d}\sum_{k=1}^{d} y_k^2}} \\ &= \sqrt{d}\, y_i \, \|y\|_2^{-1}. \end{aligned}$$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(1) $\mathrm{LN}(x)_i = \sqrt{d}\, y_i \|y\|_2^{-1}$.

Proof.

$$\partial_{y_j} \|y\|_2^{-1} = -\|y\|_2^{-2} \partial_{y_j} \|y\|_2$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d} / \|xP\|_2$ for $d \geq 3$

(1) $\mathrm{LN}(x)_i = \sqrt{d}\, y_i \|y\|_2^{-1}$.

Proof.

$$\begin{aligned} \partial_{y_j} \|y\|_2^{-1} &= -\|y\|_2^{-2} \partial_{y_j} \|y\|_2 \\ &= -\|y\|_2^{-2} \frac{y_j}{\|y\|_2} \end{aligned}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y / \|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d} / \|xP\|_2$ for $d \geq 3$

(1) $\mathrm{LN}(x)_i = \sqrt{d}\, y_i \|y\|_2^{-1}$.

Proof.

$$\begin{aligned}
\partial_{y_j} \|y\|_2^{-1} &= -\|y\|_2^{-2} \partial_{y_j} \|y\|_2 \\
&= -\|y\|_2^{-2} \frac{y_j}{\|y\|_2} \\
&= -\frac{y_j}{\|y\|_2^3}.
\end{aligned}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y / \|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d} / \|xP\|_2$ for $d \geq 3$

(1) $\mathrm{LN}(x)_i = \sqrt{d}\, y_i \|y\|_2^{-1}$.
(2) $\partial_{y_j} \|y\|_2^{-1} = -y_j / \|y\|_2^3$.

Proof.

$$\frac{\partial \,\mathrm{LN}(x)_i}{\partial y_j} = \sqrt{d}\, \partial_{y_j} (y_i \|y\|_2^{-1}) \quad ((1))$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y / \|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(1) $\mathrm{LN}(x)_i = \sqrt{d}\, y_i \|y\|_2^{-1}$.
(2) $\partial_{y_j} \|y\|_2^{-1} = -y_j/\|y\|_2^3$.

Proof.

$$\frac{\partial \,\mathrm{LN}(x)_i}{\partial y_j} = \sqrt{d}\,\partial_{y_j}(y_i \|y\|_2^{-1}) \quad ((1))$$
$$= \sqrt{d}(\delta_{ij}\|y\|_2^{-1} - y_i y_j \|y\|_2^{-3}) \quad ((2))$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\text{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(1) $\text{LN}(x)_i = \sqrt{d}\, y_i \|y\|_2^{-1}$.
(2) $\partial_{y_j} \|y\|_2^{-1} = -y_j/\|y\|_2^3$.

Proof.

$$\begin{aligned}
\frac{\partial \,\text{LN}(x)_i}{\partial y_j} &= \sqrt{d}\,\partial_{y_j}(y_i \|y\|_2^{-1}) \quad ((1)) \\
&= \sqrt{d}(\delta_{ij}\|y\|_2^{-1} - y_i y_j \|y\|_2^{-3}) \quad ((2)) \\
&= \frac{\sqrt{d}}{\|y\|_2}\left(\delta_{ij} - \frac{y_i y_j}{\|y\|_2^2}\right)
\end{aligned}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\text{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\text{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\text{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(1) $\text{LN}(x)_i = \sqrt{d}\, y_i \|y\|_2^{-1}$.
(2) $\partial_{y_j} \|y\|_2^{-1} = -y_j/\|y\|_2^3$.

Proof.

$$\begin{aligned}
\frac{\partial \,\text{LN}(x)_i}{\partial y_j} &= \sqrt{d}\,\partial_{y_j}(y_i \|y\|_2^{-1}) \quad ((1)) \\
&= \sqrt{d}(\delta_{ij}\|y\|_2^{-1} - y_i y_j \|y\|_2^{-3}) \quad ((2)) \\
&= \frac{\sqrt{d}}{\|y\|_2}\left(\delta_{ij} - \frac{y_i y_j}{\|y\|_2^2}\right) \\
\implies \frac{\partial \,\text{LN}(x)}{\partial y} &= \frac{\sqrt{d}}{\|y\|_2}\left(I - \frac{y^T y}{\|y\|_2^2}\right).
\end{aligned}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\text{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\text{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(3) $\partial \mathrm{LN}(x)/\partial y = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)$.

Proof.

$$\frac{\partial y}{\partial x} = P$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(3) $\partial \mathrm{LN}(x)/\partial y = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)$.

Proof.

$$\frac{\partial y}{\partial x} = P$$

$$= I - \frac{1}{d}\mathbf{1}^T\mathbf{1}.$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(3) $\partial \mathrm{LN}(x)/\partial y = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)$.

(4) $\partial y/\partial x = P$.

Proof.

$$J_{\mathrm{LN}}(x) = \frac{\partial \mathrm{LN}(x)}{\partial y}\frac{\partial y}{\partial x} \quad \text{(chain rule)}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(3) $\partial \mathrm{LN}(x)/\partial y = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)$.
(4) $\partial y/\partial x = P$.

Proof.

$$J_{\mathrm{LN}}(x) = \frac{\partial \mathrm{LN}(x)}{\partial y}\frac{\partial y}{\partial x} \quad \text{(chain rule)}$$

$$= \frac{\sqrt{d}}{\|y\|_2}\left(I - \frac{y^T y}{\|y\|_2^2}\right)P \quad \text{((3), (4))}.$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(5) $J_{\mathrm{LN}}(x) = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)P.$

Proof.
Let $u := y/\|y\|_2$ and $e := \mathbf{1}/\sqrt{d}$.

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(5) $J_{\mathrm{LN}}(x) = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)P.$

Proof.
Let $u := y/\|y\|_2$ and $e := \mathbf{1}/\sqrt{d}$.

$$\|J_{\mathrm{LN}}(x)\|_2 \leq \frac{\sqrt{d}}{\|y\|_2}\left\|I - u^T u\right\|_2\left\|I - e^T e\right\|_2 \quad \text{((5), submultiplicativity)}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\,y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\text{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(5) $J_{\text{LN}}(x) = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)P.$

Proof.
Let $u := y/\|y\|_2$ and $e := \mathbf{1}/\sqrt{d}$.

$$\|J_{\text{LN}}(x)\|_2 \leq \frac{\sqrt{d}}{\|y\|_2}\left\|I - u^T u\right\|_2\left\|I - e^T e\right\|_2 \quad \text{((5), submultiplicativity)}$$

$$= \frac{\sqrt{d}}{\|y\|_2} \quad \text{(projection norms)}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\text{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\text{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(5) $J_{\mathrm{LN}}(x) = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)P.$

Proof.
Let $u := y/\|y\|_2$ and $e := \mathbf{1}/\sqrt{d}$.

$$\begin{aligned}
\|J_{\mathrm{LN}}(x)\|_2 &\leq \frac{\sqrt{d}}{\|y\|_2}\left\|I - u^T u\right\|_2\left\|I - e^T e\right\|_2 && \text{((5), submultiplicativity)} \\
&= \frac{\sqrt{d}}{\|y\|_2} && \text{(projection norms)} \\
&= \frac{\sqrt{d}}{\|xP\|_2}.
\end{aligned}$$

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: $\|J_{\mathrm{LN}}(x)\|_2 = \sqrt{d}/\|xP\|_2$ for $d \geq 3$

(5) $J_{\mathrm{LN}}(x) = \frac{\sqrt{d}}{\|y\|_2}(I - y^T y/\|y\|_2^2)P.$

Proof.
Let $u := y/\|y\|_2$ and $e := \mathbf{1}/\sqrt{d}$.

$$\begin{aligned}
\|J_{\mathrm{LN}}(x)\|_2 &\leq \frac{\sqrt{d}}{\|y\|_2}\left\|I - u^T u\right\|_2 \left\|I - e^T e\right\|_2 && \text{((5), submultiplicativity)} \\
&= \frac{\sqrt{d}}{\|y\|_2} && \text{(projection norms)} \\
&= \frac{\sqrt{d}}{\|xP\|_2}.
\end{aligned}$$

For $d \geq 3$, equality holds: the product projects onto $\{u, e\}^\perp$.

LayerNorm

$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}$

$y = xP$

$\mathrm{LN}(x) = \sqrt{d}\, y/\|y\|_2$

$\|\mathrm{LN}(x)\|_2^2 = d$

Lemma 3: the error and the fix

Error in Appendix D. The last step replaces the centered norm $\|xP\|_2$ by $\|x\|_2$ without justification.

For centered $v \neq 0$, let $x = t\mathbf{1} + v$. Then

$$xP = v, \qquad \|x\|_2^2 = dt^2 + \|v\|_2^2 .$$

Changing the coordinate mean changes $\|x\|_2$, but leaves LN unchanged.

Fix. For $d \geq 3$ and $xP \neq 0$,

$$\|J_{\mathrm{LN}}(x)\|_2 = \frac{\sqrt{d}}{\|xP\|_2}.$$

The gradient proof therefore needs a lower bound on the centered normalization input.

Lemma 4.2: the paper's probability tool

Lemma 4.2 (Appendix E).

Let Y be integrable, and let B be a constant with $Y \le B$ almost surely. For every $a < B$,

$$\mathbb{P}(Y \le a) \le \frac{B - \mathbb{E}Y}{B - a}.$$

Proof idea. Apply Markov's inequality to $B - Y \ge 0$.

The upper bound $Y \le B$ must hold almost surely. A high-probability upper bound alone is insufficient. Full proof: Backup J.

Theorem 1

At initialization, under bounded loss derivatives and the paper's concentration assumptions:

Let $\mathcal{L}$ denote the loss, $\mathcal{L}_i$ its value at position i, and $\widetilde{\mathcal{L}} := \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i$.

Post-LN

$$\left\| \frac{\partial \widetilde{\mathcal{L}}}{\partial W_{2,L}} \right\|_F = O\left(d\sqrt{\log d}\right).$$

Pre-LN

$$\left\| \frac{\partial \widetilde{\mathcal{L}}}{\partial W_{2,L}} \right\|_F = O\left(d\sqrt{\frac{\log d}{L}}\right).$$

These are upper bounds.

Reported probability: at least $0.99 - \delta - \epsilon/(0.9 + \epsilon)$.

Theorem 1: implications for training

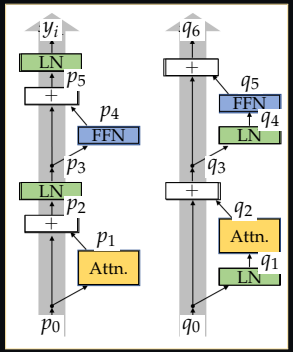
- LayerNorm placement matters. It changes the depth dependence of the initial gradient bound. Pre-LN gains an extra $L^{-1/2}$ factor for the final FFN matrix.
- A simpler training recipe. This motivates testing Pre-LN with the target learning rate from the start, without warm-up.
- Evidence from the paper. In the tested translation and BERT recipes, Pre-LN reaches comparable performance earlier without warm-up, reducing warm-up tuning.

Before the proof: the last block

Fix $\ell = L$; write $W_1 = W_{1,L}$, $W_2 = W_{2,L}$ and $F(b) = \mathrm{ReLU}(bW_1)W_2$.

Post-LN definition

$$p_0 = x_{L,i}^{\text{post}}$$
$$p_1 = \mathrm{Att}(p_0, [p_{0,j}]_j)$$
$$p_2 = p_0 + p_1$$
$$p_3 = \mathrm{LN}(p_2)$$
$$p_4 = F(p_3)$$
$$p_5 = p_3 + p_4$$
$$y_i = \mathrm{LN}(p_5)$$



Post-LN Pre-LN

Pre-LN definition

$$q_0 = x_{L,i}^{\text{pre}}$$
$$q_1 = \mathrm{LN}(q_0)$$
$$q_2 = \mathrm{Att}(q_1, [q_{1,j}]_j)$$
$$q_3 = q_0 + q_2$$
$$q_4 = \mathrm{LN}(q_3)$$
$$q_5 = F(q_4)$$
$$q_6 = q_3 + q_5$$
$$y_i = \mathrm{LN}(q_6)$$

Figure 1, relabeled. $q_6 = x_{L+1,i}^{\text{pre}}$; its final output LN is outside the pictured block.

Theorem 1: recall how the pieces fit

At the last FFN, let u_i be its ReLU activation and z_i the final LN input. With a bounded incoming loss gradient,

$$\left| \frac{\partial \mathcal{L}_i}{\partial (W_{2,L})_{pq}} \right| \lesssim |(u_i)_p| \ \left\| J_{\mathrm{LN}}(z_i) \right\|_2 .$$

- Lemmas 1–2: $\mathbb{E} \left\| z_i \right\|_2^2$ has scale d (Post-LN) or Ld (Pre-LN); Gaussian coordinate tails control u_i.
- Lemma 3: $\left\| J_{\mathrm{LN}}(z_i) \right\|_2 = \sqrt{d} / \left\| z_i P \right\|_2$ for $d \geq 3$.
- Lemma 4.2: the paper's tail bound connects $\mathbb{E} \left\| z_i P \right\|_2^2$ to a high-probability lower bound.

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

Remark: assume $\|\partial\mathcal{L}_i/\partial\,\mathrm{LN}(p_5)\|_2 = O(1)$,
with a constant bound independent of width d and depth L.

Proof.

At $\ell = L$, let $u_i := \mathrm{ReLU}(p_3 W_{1,L})$.

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\| \nabla_{W_{2,L}} \widetilde{\mathcal{L}} \right\|_F = O(d\sqrt{\log d})$$

Remark: assume $\|\partial \mathcal{L}_i / \partial \,\mathrm{LN}(p_5)\|_2 = O(1)$,
with a constant bound independent of width d and depth L.

Proof.

At $\ell = L$, let $u_i := \mathrm{ReLU}(p_3 W_{1,L})$.

$$\frac{\partial \mathcal{L}_i}{\partial (W_{2,L})_{pq}} = (u_i)_p \left[\frac{\partial \mathcal{L}_i}{\partial \,\mathrm{LN}(p_5)} J_{\mathrm{LN}}(p_5) \right]_q \qquad \text{(chain rule)}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

Remark: assume $\|\partial\mathcal{L}_i/\partial\,\mathrm{LN}(p_5)\|_2 = O(1)$,
with a constant bound independent of width d and depth L.

Proof.

At $\ell = L$, let $u_i := \mathrm{ReLU}(p_3 W_{1,L})$.

$$\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}} = (u_i)_p \left[\frac{\partial\mathcal{L}_i}{\partial\,\mathrm{LN}(p_5)} J_{\mathrm{LN}}(p_5)\right]_q \qquad \text{(chain rule)}$$

$$\Longrightarrow \left|\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 \le |(u_i)_p|^2 \left\|\frac{\partial\mathcal{L}_i}{\partial\,\mathrm{LN}(p_5)}\right\|_2^2 \|J_{\mathrm{LN}}(p_5)\|_2^2$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

Remark: assume $\|\partial\mathcal{L}_i/\partial\,\mathrm{LN}(p_5)\|_2 = O(1)$,
with a constant bound independent of width d and depth L.

Proof.

At $\ell = L$, let $u_i := \mathrm{ReLU}(p_3 W_{1,L})$.

$$\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}} = (u_i)_p \left[\frac{\partial\mathcal{L}_i}{\partial\,\mathrm{LN}(p_5)} J_{\mathrm{LN}}(p_5)\right]_q \quad \text{(chain rule)}$$

$$\implies \left|\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 \le |(u_i)_p|^2 \left\|\frac{\partial\mathcal{L}_i}{\partial\,\mathrm{LN}(p_5)}\right\|_2^2 \|J_{\mathrm{LN}}(p_5)\|_2^2$$

$$= O\Big(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2\Big) \quad \text{(bounded)}.$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\| \nabla_{W_{2,L}} \widetilde{\mathcal{L}} \right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}} \mathcal{L}_i|^2 = O(|(u_i)_p|^2 \, \|J_{\mathrm{LN}}(p_5)\|_2^2)$.

Proof.

$$\|p_3\|_2^2 = d \quad \text{(LayerNorm identity)}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}} \mathcal{L}_i|^2 = O(|(u_i)_p|^2 \, \|J_{\mathrm{LN}}(p_5)\|_2^2)$.

Proof.

$$\|p_3\|_2^2 = d \quad \text{(LayerNorm identity)}$$
$$(p_3 W_{1,L})_p \mid p_3 \sim N(0,1) \quad \text{(Xavier initialization)}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\ldots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2)$.

Proof.

$$\|p_3\|_2^2 = d \quad \text{(LayerNorm identity)}$$

$$(p_3 W_{1,L})_p \mid p_3 \sim N(0,1) \quad \text{(Xavier initialization)}$$

$$\mathbb{P}\Big(|(u_i)_p|^2 > 2\log(100d)\Big) \le \frac{0.01}{d} \quad \text{(Gaussian tail)}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2)$.

Proof.

$$\begin{aligned}
\|p_3\|_2^2 &= d && \text{(LayerNorm identity)}\\
(p_3W_{1,L})_p \mid p_3 &\sim N(0,1) && \text{(Xavier initialization)}\\
\mathbb{P}\Big(|(u_i)_p|^2 > 2\log(100d)\Big) &\le \frac{0.01}{d} && \text{(Gaussian tail)}\\
\mathbb{P}\left(\max_{p\le d}|(u_i)_p|^2 > 2\log(100d)\right) &\le 0.01 && \text{(Union bound)}
\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3W_1)W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2)$.

Proof.

$$\|p_3\|_2^2 = d \quad \text{(LayerNorm identity)}$$

$$(p_3 W_{1,L})_p \mid p_3 \sim N(0,1) \quad \text{(Xavier initialization)}$$

$$\mathbb{P}\Big(|(u_i)_p|^2 > 2\log(100d)\Big) \le \frac{0.01}{d} \quad \text{(Gaussian tail)}$$

$$\mathbb{P}\left(\max_{p\le d}|(u_i)_p|^2 > 2\log(100d)\right) \le 0.01 \quad \text{(Union bound)}$$

$$\max_{p\le d}|(u_i)_p|^2 = O(\log d) \quad \text{for fixed } i.$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$

$p_1 = \mathrm{Att}(\dots)$

$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$

$p_3 = \mathrm{LN}(p_2)$

$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\| \nabla_{W_{2,L}} \widetilde{\mathcal{L}} \right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}} \mathcal{L}_i|^2 = O(|(u_i)_p|^2 \, \|J_{\mathrm{LN}}(p_5)\|_2^2)$.
(2) $\max_p |(u_i)_p|^2 = O(\log d)$ for fixed i.

Proof.

$$\|p_5\|_2^2 \geq \tfrac{1}{10}\mathbb{E}\,\|p_5\|_2^2 \quad \text{(w.h.p.; concentration / lower tail)}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$
$p_1 = \mathrm{Att}(\dots)$
$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2)$.
(2) $\max_p |(u_i)_p|^2 = O(\log d)$ for fixed i.

Proof.

$$\begin{aligned}\|p_5\|_2^2 &\geq \tfrac{1}{10}\mathbb{E}\,\|p_5\|_2^2 \quad \text{(w.h.p.; concentration / lower tail)}\\ &= \tfrac{3}{20}d \quad \text{(Lemma 2(i))}\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$
$p_1 = \mathrm{Att}(\dots)$
$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1)W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2)$.
(2) $\max_p |(u_i)_p|^2 = O(\log d)$ for fixed i.

Proof.

$$\begin{aligned}
\|p_5\|_2^2 &\geq \tfrac{1}{10}\mathbb{E}\,\|p_5\|_2^2 \quad \text{(w.h.p.; concentration / lower tail)}\\
&= \tfrac{3}{20}d \quad \text{(Lemma 2(i))}\\
\|J_{\mathrm{LN}}(p_5)\|_2^2 &\leq \frac{d}{\|p_5\|_2^2} \quad \text{(Lemma 3)}
\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$
$p_1 = \mathrm{Att}(\dots)$
$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1)W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2)$.
(2) $\max_p |(u_i)_p|^2 = O(\log d)$ for fixed i.

Proof.

$$\begin{aligned}
\|p_5\|_2^2 &\geq \tfrac{1}{10}\mathbb{E}\,\|p_5\|_2^2 \quad \text{(w.h.p.; concentration / lower tail)}\\
&= \tfrac{3}{20}d \quad \text{(Lemma 2(i))}\\
\|J_{\mathrm{LN}}(p_5)\|_2^2 &\leq \frac{d}{\|p_5\|_2^2} \quad \text{(Lemma 3)}\\
&= O(1).
\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$
$p_1 = \mathrm{Att}(\ldots)$
$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1)W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(1) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\,\|J_{\mathrm{LN}}(p_5)\|_2^2)$.
(2) $\max_p |(u_i)_p|^2 = O(\log d)$. (3) $\|J_{\mathrm{LN}}(p_5)\|_2^2 = O(1)$ w.h.p..

Proof.

$$\left|\frac{\partial \mathcal{L}_i}{\partial (W_{2,L})_{pq}}\right|^2 = O(\log d) \quad \text{((1), (2), (3)).}$$

Post-LN definition

$p_k = x_{\ell,i}^{\mathrm{post},k}$
$p_1 = \mathrm{Att}(\dots)$
$p_2 = x_{\ell,i}^{\mathrm{post}} + p_1$
$p_3 = \mathrm{LN}(p_2)$
$p_4 = \mathrm{ReLU}(p_3 W_1) W_2$
$p_5 = p_3 + p_4$
$x_{\ell+1,i}^{\mathrm{post}} = \mathrm{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(4) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(\log d)$ on the needed event.

Proof.

$$\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2 = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \text{Att}(\dots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \text{LN}(p_2)$

$p_4 = \text{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(4) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(\log d)$ on the needed event.

Proof.

$$\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2 = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 \quad \text{(Jensen's inequality)}$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \text{Att}(\dots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \text{LN}(p_2)$

$p_4 = \text{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(4) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(\log d)$ on the needed event.

Proof.

$$\begin{aligned}\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2 &= \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 \\ &\le \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 \quad \text{(Jensen's inequality)} \\ &= O(\log d) \quad \text{((4), simultaneous event).}\end{aligned}$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \text{Att}(\dots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \text{LN}(p_2)$

$p_4 = \text{ReLU}(p_3 W_1) W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5)$

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(5) $|\partial_{(W_2)_{pq}}\widetilde{\mathcal{L}}|^2 = O(\log d)$ ‡.

Proof.

$$\left\|\frac{\partial\widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F^2 = \sum_{p,q=1}^{d}\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \text{Att}(\dots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \text{LN}(p_2)$

$p_4 = \text{ReLU}(p_3 W_1)W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5)$

† Centered norm: Backup L. ‡ Simultaneous positions: Backup M.

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(5) $|\partial_{(W_2)_{pq}}\widetilde{\mathcal{L}}|^2 = O(\log d)$ ‡.

Proof.

$$\left\|\frac{\partial\widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F^2 = \sum_{p,q=1}^{d}\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2$$
$$= O(d^2\log d) \quad ((5))$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \text{Att}(\dots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \text{LN}(p_2)$

$p_4 = \text{ReLU}(p_3 W_1)W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5)$

† Centered norm: Backup L. ‡ Simultaneous positions: Backup M.

Theorem 1(i):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d})$$

(5) $|\partial_{(W_2)_{pq}}\widetilde{\mathcal{L}}|^2 = O(\log d)$ ‡.

Proof.

$$\left\|\frac{\partial\widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F^2 = \sum_{p,q=1}^{d}\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2$$

$$= O(d^2\log d) \quad ((5))$$

$$\Longrightarrow \left\|\frac{\partial\widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F = O\left(d\sqrt{\log d}\right). \qquad \square$$

Post-LN definition

$p_k = x_{\ell,i}^{\text{post},k}$

$p_1 = \text{Att}(\dots)$

$p_2 = x_{\ell,i}^{\text{post}} + p_1$

$p_3 = \text{LN}(p_2)$

$p_4 = \text{ReLU}(p_3 W_1)W_2$

$p_5 = p_3 + p_4$

$x_{\ell+1,i}^{\text{post}} = \text{LN}(p_5)$

† Centered norm: Backup L. ‡ Simultaneous positions: Backup M.

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

Remark: assume $\left\|\partial\mathcal{L}_i/\partial\,\mathrm{LN}(x^{\mathrm{pre}}_{L+1,i})\right\|_2 = O(1)$,
with a constant bound independent of width d and depth L.

Proof.

Let $u_i := \mathrm{ReLU}(q_4 W_{1,L})$, where $\ell = L$.

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1) W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d / L})$$

Remark: assume $\left\|\partial \mathcal{L}_i / \partial \operatorname{LN}(x^{\text{pre}}_{L+1,i})\right\|_2 = O(1)$,
with a constant bound independent of width d and depth L.

Proof.

Let $u_i := \operatorname{ReLU}(q_4 W_{1,L})$, where $\ell = L$.

$$\frac{\partial \mathcal{L}_i}{\partial (W_{2,L})_{pq}} = (u_i)_p \left[\frac{\partial \mathcal{L}_i}{\partial \operatorname{LN}(x^{\text{pre}}_{L+1,i})} J_{\operatorname{LN}}(x^{\text{pre}}_{L+1,i})\right]_q \qquad \text{(chain rule)}$$

Pre-LN definition

$q_k = x^{\text{pre},k}_{\ell,i}$, $q_0 = x^{\text{pre}}_{\ell,i}$

$q_1 = \operatorname{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \operatorname{LN}(q_3)$

$q_5 = \operatorname{ReLU}(q_4 W_1) W_2$

$x^{\text{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\text{Final}} = \operatorname{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

Remark: assume $\left\|\partial\mathcal{L}_i/\partial\,\mathrm{LN}(x^{\mathrm{pre}}_{L+1,i})\right\|_2 = O(1)$,
with a constant bound independent of width d and depth L.

Proof.

Let $u_i := \mathrm{ReLU}(q_4 W_{1,L})$, where $\ell = L$.

$$\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}} = (u_i)_p \left[\frac{\partial\mathcal{L}_i}{\partial\,\mathrm{LN}(x^{\mathrm{pre}}_{L+1,i})} J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right]_q \qquad \text{(chain rule)}$$

$$\therefore \left|\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 = O\left(|(u_i)_p|^2 \left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2\right) \qquad \text{(bounded)}.$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1) W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(6) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2 \left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2)$.

Proof.

$$\|q_4\|_2^2 = d \quad \text{(LayerNorm identity)}$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1) W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d / L})$$

(6) $|\partial_{(W_2)_{pq}} \mathcal{L}_i|^2 = O(|(u_i)_p|^2 \left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2)$.

Proof.

$$\|q_4\|_2^2 = d \quad \text{(LayerNorm identity)}$$

$$\max_{p \le d} |(u_i)_p|^2 = O(\log d) \quad \text{(Gaussian calculation in part (i))}.$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1) W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(6) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2 \left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2)$.

(7) $\max_p |(u_i)_p|^2 = O(\log d)$.

Proof.

$$\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 \geq \tfrac{1}{10}\mathbb{E}\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 \quad \text{(w.h.p.; concentration / lower tail)}$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1)W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(6) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2 \left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2)$.

(7) $\max_p |(u_i)_p|^2 = O(\log d)$.

Proof.

$$\begin{aligned}\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 &\geq \tfrac{1}{10}\mathbb{E}\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 \quad \text{(w.h.p.; concentration / lower tail)}\\ &\geq \tfrac{1}{10}\left(1+\tfrac{L+1}{2}\right)d \quad \text{(Lemma 2(ii))}\end{aligned}$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1)W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(6) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2 \left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2)$.

(7) $\max_p |(u_i)_p|^2 = O(\log d)$.

Proof.

$$\begin{aligned}
\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 &\geq \tfrac{1}{10}\mathbb{E}\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 && \text{(w.h.p.; concentration / lower tail)}\\
&\geq \tfrac{1}{10}\left(1+\tfrac{L+1}{2}\right)d && \text{(Lemma 2(ii))}\\
\left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2 &\leq \frac{d}{\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2} && \text{(Lemma 3)}
\end{aligned}$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1)W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii): $\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d / L})$

(6) $|\partial_{(W_2)_{pq}} \mathcal{L}_i|^2 = O(|(u_i)_p|^2 \left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2)$.

(7) $\max_p |(u_i)_p|^2 = O(\log d)$.

Proof.

$$\begin{aligned}
\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 &\geq \tfrac{1}{10}\mathbb{E}\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2 \quad \text{(w.h.p.; concentration / lower tail)} \\
&\geq \tfrac{1}{10}\left(1 + \tfrac{L+1}{2}\right) d \quad \text{(Lemma 2(ii))} \\
\left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2 &\leq \frac{d}{\left\|x^{\mathrm{pre}}_{L+1,i}\right\|_2^2} \quad \text{(Lemma 3)} \\
&= O(1/L).
\end{aligned}$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1) W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(6) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(|(u_i)_p|^2\left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2)$.

(7) $\max_p |(u_i)_p|^2 = O(\log d)$. (8) $\left\|J_{\mathrm{LN}}(x^{\mathrm{pre}}_{L+1,i})\right\|_2^2 = O(1/L)$ w.h.p.

Proof.

$$\left|\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 = O\left(\frac{\log d}{L}\right) \quad \text{((6), (7), (8)).}$$

Pre-LN definition

$q_k = x^{\mathrm{pre},k}_{\ell,i}$, $q_0 = x^{\mathrm{pre}}_{\ell,i}$

$q_1 = \mathrm{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \mathrm{LN}(q_3)$

$q_5 = \mathrm{ReLU}(q_4 W_1)W_2$

$x^{\mathrm{pre}}_{\ell+1,i} = q_3 + q_5$

$x_{\mathrm{Final}} = \mathrm{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(9) $|\partial_{(W_2)_{pq}}\mathcal{L}_i|^2 = O(\log d/L)$.

Proof.

$$\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2 \le \frac{1}{n}\sum_i\left|\frac{\partial\mathcal{L}_i}{\partial(W_{2,L})_{pq}}\right|^2 \quad \text{(Jensen's inequality)}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1)W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Theorem 1(ii): $\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$

(9) $|\partial_{(W_2)_{pq}} \mathcal{L}_i|^2 = O(\log d/L)$.

Proof.

$$\left|\frac{\partial \widetilde{\mathcal{L}}}{\partial (W_{2,L})_{pq}}\right|^2 \leq \frac{1}{n}\sum_i \left|\frac{\partial \mathcal{L}_i}{\partial (W_{2,L})_{pq}}\right|^2 \quad \text{(Jensen's inequality)}$$

$$= O(\log d/L) \quad \text{((9), simultaneous event).}$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(10) $|\partial_{(W_2)_{pq}}\widetilde{\mathcal{L}}|^2 = O(\log d/L)$.

Proof.

$$\left\|\frac{\partial\widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F^2 = \sum_{p,q=1}^{d}\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1)W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Theorem 1(ii):

$$\left\|\nabla_{W_{2,L}}\widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$$

(10) $|\partial_{(W_2)_{pq}}\widetilde{\mathcal{L}}|^2 = O(\log d/L)$.

Proof.

$$\left\|\frac{\partial\widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F^2 = \sum_{p,q=1}^{d}\left|\frac{\partial\widetilde{\mathcal{L}}}{\partial(W_{2,L})_{pq}}\right|^2$$
$$= O(d^2\log d/L) \quad ((10))$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1)W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Theorem 1(ii): $\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F = O(d\sqrt{\log d/L})$

(10) $|\partial_{(W_2)_{pq}} \widetilde{\mathcal{L}}|^2 = O(\log d/L)$.

Proof.

$$\left\|\frac{\partial \widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F^2 = \sum_{p,q=1}^{d} \left|\frac{\partial \widetilde{\mathcal{L}}}{\partial (W_{2,L})_{pq}}\right|^2$$

$$= O(d^2 \log d/L) \quad ((10))$$

$$\implies \left\|\frac{\partial \widetilde{\mathcal{L}}}{\partial W_{2,L}}\right\|_F = O\left(d\sqrt{\frac{\log d}{L}}\right). \qquad \square$$

Pre-LN definition

$q_k = x_{\ell,i}^{\text{pre},k}$, $q_0 = x_{\ell,i}^{\text{pre}}$

$q_1 = \text{LN}(q_0)$

$q_2 = m_\ell W_{V,\ell}$

$q_3 = q_0 + q_2$

$q_4 = \text{LN}(q_3)$

$q_5 = \text{ReLU}(q_4 W_1) W_2$

$x_{\ell+1,i}^{\text{pre}} = q_3 + q_5$

$x_{\text{Final}} = \text{LN}(x_{L+1})$

Theorem 1: gaps, redundancy, and suboptimality

Let z_i be the final LN input and P the centering projection.

1. Gap: the denominator. The LN Jacobian depends on $\|z_i P\|_2$. A lower bound on $\|z_i\|_2$ alone does not control it.
2. Gap: simultaneous positions. The 0.99 activation event is for a fixed position. The termwise averaging step needs a joint event over all positions, with its probability justified.
3. Redundancy: concentration. Gaussian initialization yields the required centered lower-tail bounds, removing the extra concentration assumption.
4. Suboptimality: entrywise bounds. Bounding the matrix directly removes $\sqrt{\log d}$ and improves the width dependence for both architectures.

The repair: centered tail estimates

Under Gaussian initialization, with $d \geq 3$, let $z_i = p_5$ (Post-LN) or $z_i = x^{\mathrm{pre}}_{L+1,i}$ (Pre-LN). From input x_0, there are $L+1$ blocks; the Pre-LN threshold below uses the weaker factor L.

Post-LN

$$\mathbb{P}\left(\|z_i P\|_2^2 < \frac{d}{2} \right) \leq e^{-(d-1)/4}.$$

A normalized residual has squared norm d. Fresh Gaussian FFN noise rarely cancels it in every centered direction.

Pre-LN

$$\mathbb{P}\left(\|z_i P\|_2^2 < \frac{(d-1)L}{32} \right) \leq e^{-(d-1)/128} + 2(L+1)e^{-d/32}.$$

Fresh FFN noise accumulates across blocks. Iterate a conditional Gaussian estimate to control cancellation.

These bounds follow from initialization, without an extra concentration assumption.

A sharper proof: keep the matrix structure

At $\ell = L$, write $u_i = \mathrm{ReLU}(p_3 W_{1,L})$ (Post-LN) or $u_i = \mathrm{ReLU}(q_4 W_{1,L})$ (Pre-LN), and $y_i = \mathrm{LN}(z_i)$.

Let $r_i = g_i J_{\mathrm{LN}}(z_i)$, where $g_i = \nabla_{y_i} \mathcal{L}_i$ and $\|g_i\|_2 \le G$ almost surely.

$$\nabla_{W_{2,L}} \mathcal{L}_i = u_i^T r_i, \qquad \left\| u_i^T r_i \right\|_F = \|u_i\|_2 \, \|r_i\|_2 .$$

$$\left\| \nabla_{W_{2,L}} \widetilde{\mathcal{L}} \right\|_F \le \frac{G\sqrt{d}}{n} \sum_{i=1}^{n} \frac{\|u_i\|_2}{\|z_i P\|_2}.$$

Use the event $\|u_i\|_2^2 \le d$ to control the full activation vector. This removes $\sqrt{\log d}$ and improves the width bound further.

The corrected last-FFN gradient bound

For $d \geq 3$, under Gaussian initialization and $\|g_i\|_2 \leq G$:

Post-LN

$$\|z_i P\|_2^2 \geq d/2$$

$$\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F \leq 2G\sqrt{d}.$$

Pre-LN

$$\|z_i P\|_2^2 \geq (d-1)L/32$$

$$\left\|\nabla_{W_{2,L}} \widetilde{\mathcal{L}}\right\|_F \leq 8G\sqrt{d/L}.$$

Combine the centered lower tails with $\mathbb{P}(\|u_i\|_2^2 > d) \leq e^{-d/32}$. Union bounds over tokens require no independence.

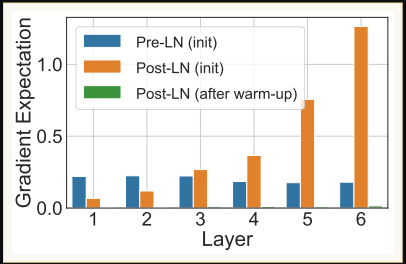
For each architecture, the bound holds with probability at least

$$1 - 4n(L+1)e^{-(d-1)/128}.$$

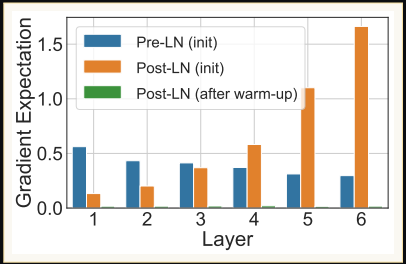
Measured gradients across layer position

IWSLT14, 6 encoder / 6 decoder layers. Figure 3(a,b) shows decoder profiles of the norm of minibatch-averaged gradients.

W_1: first FFN matrix



W_2: second FFN matrix

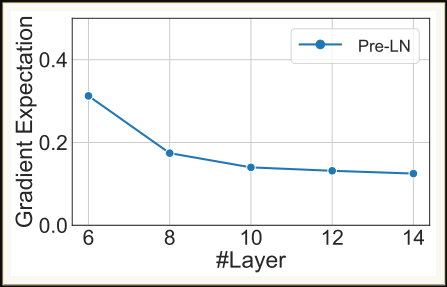


Post-LN gradients rise toward the output; Pre-LN profiles are flatter.

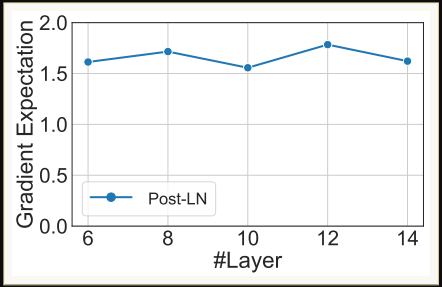
Final-layer gradients as depth increases

Figure 3(c,d): encoder/decoder depths 6/6 to 14/14; ten initialization seeds.

Pre-LN

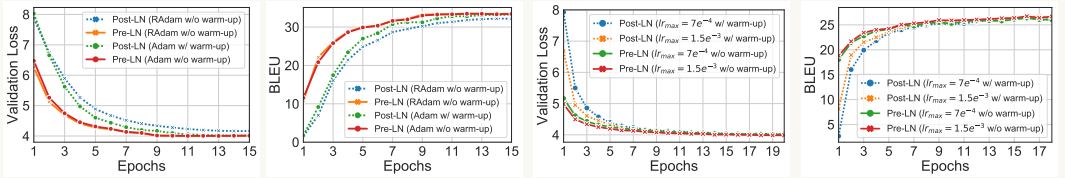


Post-LN



The final W_2 gradient decreases with depth in Pre-LN in these runs.

Translation: IWSLT14 and WMT14



(a) Validation Loss (IWSLT) (b) BLEU (IWSLT) (c) Validation Loss (WMT) (d) BLEU (WMT)

IWSLT14 De–En

Pre-LN reaches comparable loss and BLEU earlier (about epoch 9 versus 15 with Adam). RAdam trains Post-LN without warm-up but changes Pre-LN little.

WMT14 En–De

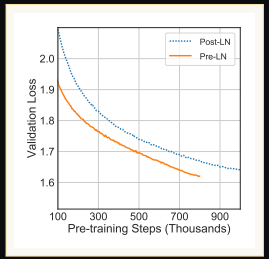
Both tested rates, 7×10^{-4} and 1.5×10^{-3}, show earlier progress with Pre-LN and competitive final BLEU.

Figure 4. Adam: Pre-LN without warm-up; Post-LN with 4000 steps. Decay schedules differ.

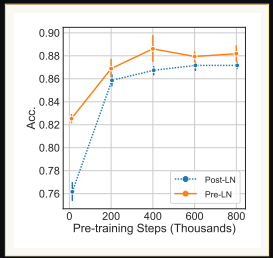
BERT: earlier progress in loss and accuracy

Figure 5(a–c). Accuracy: fine-tuned checkpoints, six seeds, 95% confidence intervals.

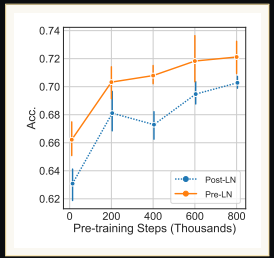
BERT validation loss



MRPC accuracy



RTE accuracy



Post-LN: 10k warm-up steps, peak rate 10^{-4}.

Pre-LN: no warm-up, initial rate 3×10^{-4}.

Limitations of the theory

- Simplified model. Gaussian initialization, uniform single-head attention, square FFNs, and ideal LayerNorm.
- Limited target. The theorem bounds the last FFN's W_2 *at initialization*. Earlier layers require full Jacobian products. No guarantee that the bound holds after training begins.
- Probability limits. High-probability upper bounds do not alone bound expected gradients from below.
- Optimization remains open. No ordering of realized gradients, convergence guarantee, or proof that warm-up is (un)necessary.

Conclusion: why LayerNorm placement matters

1. Mechanism. LN placement changes residual scale and the initial gradient bounds. Pre-LN's final LN supplies an extra depth-dependent factor.

2. Evidence. Measured Pre-LN gradients are more uniform across layers. The tested translation and BERT recipes train without warm-up and reach comparable performance earlier.

3. Contribution. The paper links initialization analysis to a simpler training recipe, reducing warm-up tuning while motivating further study of Transformer optimization.

LN placement is a consequential choice for optimization.

Thank you

Questions are welcomed.

Backup A · exact Post-LN recursion

$$
\begin{aligned}
x_{\ell,i}^{\text{post},1} &= \mathrm{Att}(x_{\ell,i}^{\text{post}}, [x_{\ell,j}^{\text{post}}]_{j=1}^{n}),\\
x_{\ell,i}^{\text{post},2} &= x_{\ell,i}^{\text{post}} + x_{\ell,i}^{\text{post},1},\\
x_{\ell,i}^{\text{post},3} &= \mathrm{LN}(x_{\ell,i}^{\text{post},2}),\\
x_{\ell,i}^{\text{post},4} &= \mathrm{ReLU}(x_{\ell,i}^{\text{post},3} W_{1,\ell} + b_{1,\ell}) W_{2,\ell} + b_{2,\ell},\\
x_{\ell,i}^{\text{post},5} &= x_{\ell,i}^{\text{post},3} + x_{\ell,i}^{\text{post},4},\\
x_{\ell+1,i}^{\text{post}} &= \mathrm{LN}(x_{\ell,i}^{\text{post},5}).
\end{aligned}
$$

Backup B · exact Pre-LN recursion

$$
\begin{aligned}
x_{\ell,i}^{\text{pre},1} &= \mathrm{LN}(x_{\ell,i}^{\text{pre}}),\\
x_{\ell,i}^{\text{pre},2} &= \mathrm{Att}(x_{\ell,i}^{\text{pre},1}, [x_{\ell,j}^{\text{pre},1}]_{j=1}^{n}),\\
x_{\ell,i}^{\text{pre},3} &= x_{\ell,i}^{\text{pre}} + x_{\ell,i}^{\text{pre},2},\\
x_{\ell,i}^{\text{pre},4} &= \mathrm{LN}(x_{\ell,i}^{\text{pre},3}),\\
x_{\ell,i}^{\text{pre},5} &= \mathrm{ReLU}(x_{\ell,i}^{\text{pre},4} W_{1,\ell} + b_{1,\ell}) W_{2,\ell} + b_{2,\ell},\\
x_{\ell+1,i}^{\text{pre}} &= x_{\ell,i}^{\text{pre},3} + x_{\ell,i}^{\text{pre},5},\\
x_{\text{Final,i}}^{\text{pre}} &= \mathrm{LN}(x_{L+1,i}^{\text{pre}}).
\end{aligned}
$$

Backup C · assumption taxonomy

Model idealization	One attention head; $d \times d$ weights; ReLU FFN.
Initialization	Independent $N(0, 1/d)$ entries, zero biases, $\gamma = 1$, $\beta = 0$, Gaussian inputs in the stated setup.
Analytic attention	$W_Q = W_K = 0$ initially, so $\alpha_{ij} = 1/n$ and attention is an average followed by W_V.
Theorem 1	Bounded derivatives and concentration for $\left\|x_{L,i}^{\text{post},5}\right\|_2^2$ and $\left\|x_{L+1,i}^{\text{pre}}\right\|_2^2$.

These are analytical premises, not asserted properties of ordinary trained Transformers.

Backup D · LayerNorm geometry

$$P = I - \frac{1}{d}\mathbf{1}^T\mathbf{1}, \quad P^2 = P = P^T, \quad P\mathbf{1}^T = 0.$$

$$y = xP, \quad \mathrm{LN}(x) = \sqrt{d}\,\frac{y}{\|y\|_2}, \quad \mathbf{1}\,\mathrm{LN}(x)^T = 0, \quad \|\mathrm{LN}(x)\|_2^2 = d.$$

$$\mathrm{LN}(x + t\mathbf{1}) = \mathrm{LN}(x), \qquad \mathrm{LN}(cx) = \mathrm{LN}(x) \quad (c > 0).$$

Centering translation invariance explains why $\|x\|$ cannot control the Jacobian.

Backup E · conditional FFN energy

For a fixed row v with $\|v\|_2^2 = d$, let $U = \mathrm{ReLU}(vW_1)$.

$$vW_1 \mid v \sim N(0, I_d), \qquad \mathbb{E}[\|U\|_2^2 \mid v] = d/2.$$

$$\mathbb{E}[W_2 W_2^T] = I_d, \qquad \mathbb{E}[\|UW_2\|_2^2 \mid U] = U\mathbb{E}[W_2 W_2^T]U^T = \|U\|_2^2.$$

$$\mathbb{E}[\|UW_2\|_2^2 \mid v] = d/2.$$

Fresh W_2 is independent of U and has zero mean.

Backup F · conditional attention energy

$u_j = \mathrm{LN}(x^{\mathrm{pre}}_{\ell,j})$, $m = n^{-1}\sum_j u_j$, $q_2 = mW_V$.

$$\mathbb{E}[\|mW_V\|_2^2 \mid x^{\mathrm{pre}}_{\ell,1:n}] = m\mathbb{E}[W_V W_V^T]m^T = \|m\|_2^2 .$$

$$\mathbb{E}[x^{\mathrm{pre}}_{\ell,i}(mW_V)^T \mid x^{\mathrm{pre}}_{\ell,1:n}] = x^{\mathrm{pre}}_{\ell,i}\mathbb{E}[W_V^T]m^T = 0.$$

$$0 \le \mathbb{E}\,\|m\|_2^2 \le \frac{1}{n}\sum_j \mathbb{E}\,\|u_j\|_2^2 = d \qquad \text{(Jensen)}.$$

Backup G · second residual cross term

Condition on the previous layers and $W_{1,\ell}$. Then q_3, q_4 and $U = \mathrm{ReLU}(q_4 W_{1,\ell})$ are fixed, whereas $W_{2,\ell}$ is fresh and centered.

$$\mathbb{E}[q_3 q_5^T \mid q_3, q_4, W_{1,\ell}] = q_3 \mathbb{E}[W_{2,\ell}^T] U^T = 0.$$

$$\mathbb{E}\,\|q_3 + q_5\|_2^2 = \mathbb{E}\,\|q_3\|_2^2 + \mathbb{E}\,\|q_5\|_2^2 .$$

Backup H · the projection fact

For a unit row a, $Q = a^T a$ is a rank-one orthogonal projection:

$$Q^2 = Q = Q^T, \qquad Qa^T = a^T, \qquad Qv = 0 \text{ if } av = 0.$$

Thus $I - Q$ has eigenvalue 0 on $\operatorname{span}(a^T)$ and eigenvalue 1 on its orthogonal complement.

$$\left\| I - a^T a \right\|_2 = 1 \qquad \text{when the ambient dimension is at least two.}$$

Backup I · exact centered Jacobian

With $u = y/\|y\|_2$ and $e = \mathbf{1}/\sqrt{d}$, $u \perp e$:

$$J_{\mathrm{LN}}(x) = \frac{\sqrt{d}}{\|xP\|_2}(I - u^T u)(I - e^T e).$$

The projections commute because $u \perp e$. On $(\mathrm{span}\{u, e\})^\perp$, their product is the identity. For $d \geq 3$, this space is nonzero, hence

$$\|J_{\mathrm{LN}}(x)\|_2 = \frac{\sqrt{d}}{\|xP\|_2}.$$

For $d = 2$, the idealized Jacobian vanishes where defined.

Backup J · Lemma 4.2 in one line

If $0 \le Y \le B$ almost surely, set $X = B - Y \ge 0$. For $a < B$, Markov gives

$$\mathbb{P}(Y \le a) = \mathbb{P}(X \ge B - a) \le \frac{\mathbb{E}X}{B - a} = \frac{B - \mathbb{E}Y}{B - a}.$$

The almost-sure premise matters when applying this lemma to a squared norm.

Backup K · exact probability and definition

Theorem 1 reports probability at least

$$0.99 - \delta - \frac{\epsilon}{0.9 + \epsilon}$$

for each final-layer gradient bound under its bounded-derivative and (ϵ, δ) assumptions. The printed Definition 1 is one-sided: $\mathbb{P}((Z - \mathbb{E}Z)/\mathbb{E}Z \leq \epsilon) \geq 1 - \delta$. The

supplied absolute-value formulation is stronger. The supplement bounds Z above only on a high-probability event, then applies the almost-sure Lemma 4.2 as if the upper bound were deterministic. This step needs a further argument.

Backup L · why the centered norm matters

Take nonzero centered v and put $x = t\mathbf{1} + v$. Then

$$xP = v, \qquad \|x\|_2^2 = dt^2 + \|v\|_2^2 .$$

For $d \geq 3$,

$$\|J_{\mathrm{LN}}(x)\|_2 = \frac{\sqrt{d}}{\|v\|_2}.$$

As $|t| \to \infty$, this Jacobian stays fixed while $\sqrt{d}/\|x\|_2 \to 0$. There is no uniform deterministic replacement of $\|xP\|$ by $\|x\|$. A model-specific centered-norm lower bound could repair the intended bridge; the displayed norm hypotheses do not supply one.

Backup M · sequence-level probability

For one fixed position i, a union bound over d coordinates yields

$$\mathbb{P}\left(\max_{p\le d}|h_{i,p}|^2 > 2\log(100d)\right) \le 0.01.$$

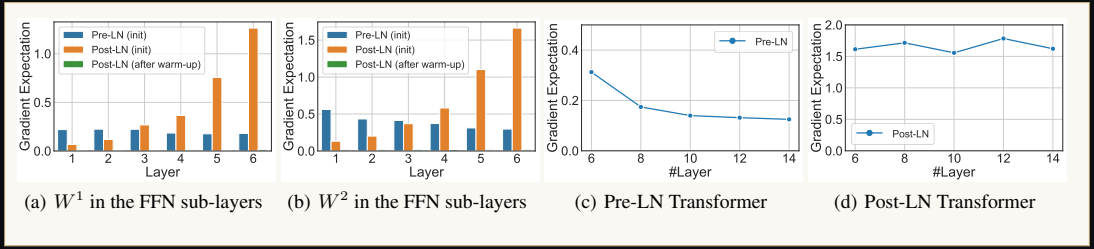
To average positions using the same pointwise estimate, one needs a simultaneous event. A direct union bound costs $0.01n$; a threshold $2\log(100nd)$ instead gives probability 0.99. The claimed $\log d$ with the same 0.99 event needs more control if n varies; the supplied proof does not provide it.

Backup N · experimental schedules

Task	Post-LN baseline	Pre-LN comparison
IWSLT14 De→En	Adam; 4000-step warm-up, inverse-square-root decay	No warm-up; initial 5×10^{-4}, factor 0.1 at epoch 8
WMT14 En→De	Adam; 4000-step warm-up, inverse-square-root decay	No warm-up; initial 7×10^{-4} or 1.5×10^{-3}; decay at epoch 6 then inverse-square-root
BERT	10k warm-up steps, peak 10^{-4}	No warm-up; linear decay from 3×10^{-4}

The architecture comparison also changes schedules; it does not isolate placement alone.

Backup O · all Figure 3 panels



(a) W^1 in the FFN sub-layers (b) W^2 in the FFN sub-layers (c) Pre-LN Transformer (d) Post-LN Transformer

Figure 3(a): W_1; (b): W_2 across layers; (c), (d): final W_2 versus depth.

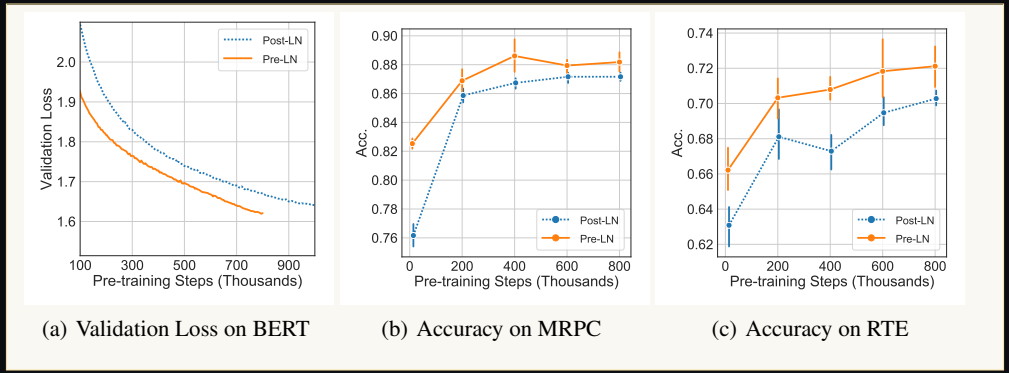
Backup P · additional IWSLT14 result

The paper also reports 35.5 test BLEU for Pre-LN on IWSLT14 De–En with a separately tuned recipe.

Setting	Tuned configuration
Initial learning rate	7.5×10^{-4}
Learning-rate decay	After 8000 updates; inverse-square-root decay
Dropout	0.3; attention dropout 0.1
Batch size	8192 tokens

Appendix A.1, footnote. This is a separate configuration from the Figure 4 comparison; the paper gives no matched Post-LN result for this recipe.

Backup Q · BERT and downstream panels



(a) Validation Loss on BERT (b) Accuracy on MRPC (c) Accuracy on RTE

Figure 5: BERT validation loss and MRPC/RTE downstream accuracy.

Backup R · softmax cross-entropy

Let $y \in \mathbb{R}^{1\times d}$ be a token's output. For each vocabulary item $v \in \mathcal{V}$, let e_v be its readout row and v_* the correct label.

$$s_v = y e_v^T \quad \text{(score)}, \qquad p_v = \frac{e^{s_v}}{\sum_{u\in\mathcal{V}} e^{s_u}} \quad \text{(probability)}.$$

Cross-entropy penalizes low probability for the correct label:

$$\mathcal{L}_i(y) = -\log p_{v_*} = -s_{v_*} + \log \sum_{u\in\mathcal{V}} e^{s_u}.$$

$$\frac{\partial \mathcal{L}_i}{\partial s_v} = p_v - \mathbf{1}_{\{v=v_*\}}.$$

The derivative is the predicted probability minus the target indicator.
The readout is held fixed when differentiating with respect to the token output.

Backup S · the gradient entering the Transformer

With $s_v = y e_v^T$ and softmax probabilities p_v, the chain rule gives

$$g_i := \nabla_y \mathcal{L}_i = \sum_{v \in \mathcal{V}} (p_v - \mathbf{1}_{\{v = v_*\}}) e_v$$

$$= \sum_{v \in \mathcal{V}} p_v e_v - e_{v_*}.$$

Since $p_v \geq 0$ and $\sum_v p_v = 1$,

$$\|g_i\|_2 \leq \sum_v p_v \|e_v\|_2 + \|e_{v_*}\|_2 \leq 2 \max_v \|e_v\|_2 .$$

If $\max_v \|e_v\|_2 \leq C$ almost surely, we may take $G = 2C$.

No independence from hidden states is needed. The original proof uses $G = O(1)$.
A coordinate bound $|(g_i)_j| \leq C$ alone gives only $\|g_i\|_2 \leq C\sqrt{d}$.

Backup T · earlier layers: the chain rule

Stack all token states into a column $H_\ell \in \mathbb{R}^{nd}$. Define

$$B_\ell := \frac{\partial H_{\ell+1}}{\partial H_\ell}, \qquad r_\ell := \nabla_{H_\ell} \widetilde{\mathcal{L}}.$$

$$r_\ell = B_\ell^T r_{\ell+1}, \qquad \|r_\ell\|_2 \le \|r_{L+1}\|_2 \prod_{j=\ell}^{L} \|B_j\|_2 .$$

For a parameter vector θ_ℓ used only in block ℓ,

$$\nabla_{\theta_\ell} \widetilde{\mathcal{L}} = \left(\frac{\partial H_{\ell+1}}{\partial \theta_\ell} \right)^T r_{\ell+1}.$$

Earlier parameter gradients require both local derivatives and later blocks.

Appendix F. For Pre-LN, r_{L+1} includes the final output LN derivative.
The loss-gradient bound G enters once through r_{L+1}.

Backup U · earlier layers: residual Jacobians

Let H be the stacked column state, F an attention or FFN sublayer, and $\mathcal{N}$ tokenwise LayerNorm.

Post-LN

$$T_{\text{post}}(H) = \mathcal{N}(H + F(H)).$$

$$J_{T_{\text{post}}}(H) = J_{\mathcal{N}}(H + F(H)) \cdot (I + J_F(H)).$$

Pre-LN

$$T_{\text{pre}}(H) = H + F(\mathcal{N}(H)).$$

$$J_{T_{\text{pre}}}(H) = I + J_F(\mathcal{N}(H)) J_{\mathcal{N}}(H).$$

Pre-LN retains an identity term in each residual derivative.

Appendix F. A complete block composes attention and FFN, so its Jacobian is the product of the two sublayer Jacobians.

Backup V · earlier layers: limits of the heuristic

Post-LN. A typical LN scale near $\sqrt{2/3}$ motivates decay. The other factors in the residual Jacobian can offset that contraction.

Pre-LN. Jacobians near the identity motivate stable gradients. Small deviations can still accumulate over many blocks. Even scalars give

$$\prod_{j=1}^{L}\left(1+\frac{c}{\sqrt{j}}\right)=\exp\bigl(\Theta(\sqrt{L})\bigr),\qquad c>0 \text{ fixed.}$$

Stability requires control of the complete Jacobian products.

Appendix F provides the architectural intuition. The final-FFN theorem does not prove gradient decay or stability in earlier layers.