Post-LN Transformer 训练不稳定性的数学分析

Abstract

本文从线性代数的角度严格分析 Post-LN 与 Pre-LN Transformer 架构在训练稳定性上的差异。我们证明 LayerNorm 的 Jacobian 矩阵具有非平凡的零空间，导致 Post-LN 架构中残差连接的恒等通路被" 截断"，从而在反向传播时某些方向的梯度被完全抹除。相比之下，Pre-LN 架构通过将恒等映射置于最外层，保证了梯度在所有方向上的正下界。本文的分析为实践中观察到的 Post-LN 训练不稳定现象提供了严格的数学解释。

Contents

1 预备知识与符号约定

1.1 基本符号

设 $d \in \mathbb{N}$ 为隐藏层维度。我们约定如下符号：

- 所有向量均为列向量，$x \in \mathbb{R}^d$；$\mathbf{1} = (1, 1, \ldots, 1)^\top \in \mathbb{R}^d$ 表示全 1 向量。
- $I \in \mathbb{R}^{d \times d}$ 表示单位矩阵；$A^\top$ 表示矩阵转置。
- $\odot$ 表示 Hadamard 积（逐元素乘法）；$\mathrm{diag}(v)$ 表示对角矩阵。
- 对于可微映射 $f : \mathbb{R}^d \to \mathbb{R}^d$，记其在点 x 处的 Jacobian 矩阵为 $J_f(x) \in \mathbb{R}^{d \times d}$，其中 $(J_f(x))_{ij} = \frac{\partial f_i}{\partial x_j}$。
- $\mathrm{rank}(A)$、$\mathrm{tr}(A)$ 分别表示矩阵 A 的秩和迹。
- $\ker(A)$ 表示矩阵 A 的零空间；$\mathrm{Im}(A)$ 表示其像空间。
- $\mathrm{span}\{v_1, \ldots, v_k\}$ 表示向量组的线性张成空间。
- $\|\cdot\|_2$ 表示欧几里得范数；$\|\cdot\|$ 表示矩阵的算子范数（谱范数）。
- $\sigma_{\min}(A)$ 和 $\sigma_{\max}(A)$ 分别表示矩阵 A 的最小和最大奇异值。
- $o(\epsilon)$ 表示高阶无穷小项。

1.2 LayerNorm 的定义

定义 1.1 (LayerNorm)**.** LayerNorm 映射 $\mathrm{LN} : \mathbb{R}^d \to \mathbb{R}^d$ 定义为

$$\mathrm{LN}(z) = \frac{z - \mu(z)\mathbf{1}}{\sigma(z)}, \tag{1}$$

其中均值函数 $\mu : \mathbb{R}^d \to \mathbb{R}$ 和标准差函数 $\sigma : \mathbb{R}^d \to \mathbb{R}_{>0}$ 定义为

$$\mu(z) = \frac{1}{d}\sum_{i=1}^{d} z_i = \frac{1}{d}\mathbf{1}^\top z, \tag{2}$$

$$\sigma(z) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(z_i - \mu(z))^2} = \sqrt{\frac{1}{d}\|z - \mu(z)\mathbf{1}\|_2^2}. \tag{3}$$

注记 1.2. 在实际实现中，LayerNorm 通常还包含可学习的仿射参数 $\gamma, \beta \in \mathbb{R}^d$，即

$$\mathrm{LN}_{\gamma,\beta}(z) = \gamma \odot \mathrm{LN}(z) + \beta, \tag{4}$$

其中 $\odot$ 表示 Hadamard 积。这对 Jacobian 的影响仅是右乘对角矩阵 $\mathrm{diag}(\gamma)$，不改变零空间的结构，故我们先分析不含仿射参数的情况。

1.3 Post-LN 与 Pre-LN 的结构

定义 1.3 (Post-LN 层)**.** 设 $F : \mathbb{R}^d \to \mathbb{R}^d$ 为子层映射（Attention 或前馈网络）。Post-LN 层定义为

$$y = \mathrm{LN}(x + F(x)). \tag{5}$$

定义 1.4 (Pre-LN 层)**.** Pre-LN 层定义为

$$y = x + F(\mathrm{LN}(x)). \tag{6}$$

2 LayerNorm 的 Jacobian 分析

2.1 Jacobian 的显式计算

引理 2.1 (LayerNorm 的 Jacobian)**.** 设 $z \in \mathbb{R}^d$ 满足 $\sigma(z) > 0$。则 LayerNorm 在 z 处的 Jacobian 矩阵为

$$J_{\mathrm{LN}}(z) = \frac{1}{\sigma(z)} \left(I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top - \frac{1}{d}\hat{z}\hat{z}^\top \right), \tag{7}$$

其中 $\hat{z} = \frac{z-\mu(z)\mathbf{1}}{\sigma(z)} = \mathrm{LN}(z)$ 是标准化后的向量。

Proof. 我们分步计算。令 $\bar{z} = z - \mu(z)\mathbf{1}$（去均值后的向量），则 $\mathrm{LN}(z) = \bar{z}/\sigma(z)$。

步骤 1：计算 $\frac{\partial \mu}{\partial z}$
由 $\mu(z) = \frac{1}{d}\mathbf{1}^\top z$，直接得

$$\frac{\partial \mu}{\partial z} = \frac{1}{d}\mathbf{1}^\top \in \mathbb{R}^{1\times d}. \tag{8}$$

步骤 2：计算 $\frac{\partial \bar{z}}{\partial z}$
由 $\bar{z} = z - \mu(z)\mathbf{1}$，利用乘积法则：

$$\frac{\partial \bar{z}}{\partial z} = I - \mathbf{1}\frac{\partial \mu}{\partial z} = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top. \tag{9}$$

记这个矩阵为 $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$。注意 P 是到 $\mathbf{1}^\perp$（与 $\mathbf{1}$ 正交的子空间）的正交投影矩阵。

步骤 3：计算 $\frac{\partial \sigma}{\partial z}$
由 $\sigma(z)^2 = \frac{1}{d}\|\bar{z}\|_2^2 = \frac{1}{d}\bar{z}^\top\bar{z}$，对两边关于 z 求导：

$$2\sigma\frac{\partial \sigma}{\partial z} = \frac{2}{d}\bar{z}^\top\frac{\partial \bar{z}}{\partial z} = \frac{2}{d}\bar{z}^\top P. \tag{10}$$

由于 P 是对称矩阵且 $P\bar{z} = \bar{z}$（因为 $\bar{z}$ 已经去均值，属于 $\mathbf{1}^\perp$），我们有 $\bar{z}^\top P = \bar{z}^\top$。因此

$$\frac{\partial \sigma}{\partial z} = \frac{1}{d\sigma}\bar{z}^\top = \frac{1}{d}\hat{z}^\top, \tag{11}$$

其中最后一步用了 $\hat{z} = \bar{z}/\sigma$。

步骤 4：组合得到 J_{LN}
由 $\mathrm{LN}(z) = \bar{z}/\sigma$，利用商法则：

$$J_{\mathrm{LN}}(z) = \frac{\partial}{\partial z}\left(\frac{\bar{z}}{\sigma}\right) = \frac{1}{\sigma}\frac{\partial \bar{z}}{\partial z} - \frac{\bar{z}}{\sigma^2}\frac{\partial \sigma}{\partial z} \tag{12}$$

$$= \frac{1}{\sigma}P - \frac{\sigma\hat{z}}{\sigma^2}\cdot\frac{1}{d}\hat{z}^\top \quad (\text{代入}\bar{z} = \sigma\hat{z}) \tag{13}$$

$$= \frac{1}{\sigma}P - \frac{1}{d}\hat{z}\hat{z}^\top \tag{14}$$

$$= \frac{1}{\sigma}\left(I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top - \frac{1}{d}\hat{z}\hat{z}^\top\right). \tag{15}$$

这完成了证明。 ■

2.2 零空间的刻画

定理 2.2 (LayerNorm Jacobian 的零空间)**.** 对任意 $z \in \mathbb{R}^d$（满足 $\sigma(z) > 0$），有

$$\ker(J_{\mathrm{LN}}(z)) = \mathrm{span}\{\mathbf{1}, \hat{z}\}, \tag{16}$$

其中 $\hat{z} = \mathrm{LN}(z)$。特别地，由于 $\mathbf{1}$ 与 $\hat{z}$ 正交（因 $\mathbf{1}^\top \hat{z} = 0$），它们线性无关，故 $\dim(\ker(J_{\mathrm{LN}}(z))) = 2$，进而由秩-零化度定理得 $\mathrm{rank}(J_{\mathrm{LN}}(z)) = d - 2$。

Proof. 由引理 2.1，$J_{\mathrm{LN}}(z) = \frac{1}{\sigma}(I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top - \frac{1}{d}\hat{z}\hat{z}^\top)$。由于 $\sigma > 0$，零空间完全由括号内的矩阵决定。
令 $M = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top - \frac{1}{d}\hat{z}\hat{z}^\top$，我们需要找到所有满足条件的 v，使得 $Mv = 0$。

验证 $\mathbf{1} \in \ker(M)$

$$M\mathbf{1} = \mathbf{1} - \frac{1}{d}\mathbf{1}\mathbf{1}^\top\mathbf{1} - \frac{1}{d}\hat{z}\hat{z}^\top\mathbf{1}. \tag{17}$$

由于 $\mathbf{1}^\top\mathbf{1} = d$，第二项为 $\mathbf{1}$。由于 $\hat{z} = \mathrm{LN}(z)$ 满足 $\mathbf{1}^\top\hat{z} = 0$（标准化后均值为零），第三项为 0。因此 $M\mathbf{1} = \mathbf{1} - \mathbf{1} - 0 = 0$。

验证 $\hat{z} \in \ker(M)$

$$M\hat{z} = \hat{z} - \frac{1}{d}\mathbf{1}\mathbf{1}^\top\hat{z} - \frac{1}{d}\hat{z}\hat{z}^\top\hat{z}. \tag{18}$$

第二项：由于 $\mathbf{1}^\top\hat{z} = 0$，此项为 0。第三项：由于 $\hat{z}^\top\hat{z} = \|\hat{z}\|_2^2 = d$（标准化后方差为 1，即 $\frac{1}{d}\|\hat{z}\|^2 = 1$），此项为 $\hat{z}$。因此 $M\hat{z} = \hat{z} - 0 - \hat{z} = 0$。

验证 $\mathbf{1}$ 和 $\hat{z}$ 线性无关
由于 $\mathbf{1}^\top\hat{z} = 0$，它们正交，故线性无关。

验证 $\ker(M)$ 恰为二维
矩阵 M 可以写成

$$M = I - \frac{1}{d}(\mathbf{1}\mathbf{1}^\top + \hat{z}\hat{z}^\top). \tag{19}$$

令 $Q = \frac{1}{d}(\mathbf{1}\mathbf{1}^\top + \hat{z}\hat{z}^\top)$。由于 $\mathbf{1}$ 和 $\hat{z}$ 正交且 $\|\mathbf{1}\|^2 = d$、$\|\hat{z}\|^2 = d$，矩阵 Q 是到 $\mathrm{span}\{\mathbf{1}, \hat{z}\}$ 的正交投影。因此 $M = I - Q$ 是到 $\mathrm{span}\{\mathbf{1}, \hat{z}\}^\perp$ 的正交投影，其零空间恰为 $\mathrm{span}\{\mathbf{1}, \hat{z}\}$。 ■

注记 2.3. 定理 2.2 表明 LayerNorm 的 Jacobian 是一个秩为 $d - 2$ 的矩阵，其零空间由两个正交方向张成：全 1 向量 $\mathbf{1}$（对应均值方向）和标准化后的向量 $\hat{z}$（对应输出方向）。这是 LayerNorm 操作的内在几何性质。

3 Post-LN 的梯度分析

3.1 Jacobian 的结构

引理 3.1 (Post-LN 的 Jacobian)**.** 设 Post-LN 层为 $y = \mathrm{LN}(x + F(x))$，令 $z = x + F(x)$。则从输入 x 到输出 y 的 Jacobian 为

$$J^{(\mathrm{Post})}(x) = J_{\mathrm{LN}}(z) \cdot (I + J_F(x)). \tag{20}$$

Proof. 这是链式法则的直接应用。定义复合映射：

- $g(x) = x + F(x)$，则 $J_g(x) = I + J_F(x)$；
- $h(z) = \mathrm{LN}(z)$，则 $J_h(z) = J_{\mathrm{LN}}(z)$；
- $y = h(g(x))$。

由链式法则：

$$J^{(\mathrm{Post})}(x) = J_h(g(x)) \cdot J_g(x) = J_{\mathrm{LN}}(z) \cdot (I + J_F(x)). \tag{21}$$

■

3.2 秩的退化

定理 3.2 (Post-LN Jacobian 的秩上界)**.** 对任意 x 和任意可微子层 F，有

$$\mathrm{rank}(J^{(\mathrm{Post})}(x)) \le d - 2. \tag{22}$$

Proof. 由矩阵乘积的秩不等式：对任意矩阵 A, B，有

$$\mathrm{rank}(AB) \le \min(\mathrm{rank}(A), \mathrm{rank}(B)). \tag{23}$$

由定理 2.2，$\mathrm{rank}(J_{\mathrm{LN}}(z)) = d - 2$。因此

$$\mathrm{rank}(J^{(\mathrm{Post})}(x)) = \mathrm{rank}(J_{\mathrm{LN}}(z) \cdot (I + J_F(x))) \le \mathrm{rank}(J_{\mathrm{LN}}(z)) = d - 2. \tag{24}$$

■

定理 3.3 (秩-零化维数定理)**.** 对于矩阵 $M \in \mathbb{R}^{d \times d}$，其核空间与秩满足

$$\dim(\ker(M)) + \mathrm{rank}(M) = d. \tag{25}$$

定义 3.4 (非平凡零空间)**.** 设 $M \in \mathbb{R}^{m \times n}$。若 $\ker(M) \neq \{0\}$，则称 M 的零空间为**非平凡的**（non-trivial），等价地，$\dim(\ker(M)) \ge 1$。

推论 3.5 (Post-LN Jacobian 的零空间非平凡)**.** 由定理 2.2及定理 3.3可知，对于任意输入 x，Post-LN 层的 Jacobian $J^{(\mathrm{Post})}(x)$ 的核空间至少有两个维度：

$$\dim(\ker(J^{(\mathrm{Post})}(x))) \ge 2. \tag{26}$$

这意味着存在至少两个线性无关的方向 $v_1, v_2 \in \mathbb{R}^d$，使得 $J^{(\mathrm{Post})}(x)v_i = 0$，即沿这些方向的梯度信息无法通过该层传播。

3.3 梯度消失方向的刻画

定理 3.6 (Post-LN 的梯度消失方向)**.** 设 $z = x + F(x)$，$\hat{z} = \mathrm{LN}(z)$。则对任意向量 $w \in \mathbb{R}^d$，有

$$J^{(\mathrm{Post})}(x) \cdot w = 0 \iff (I + J_F(x))w \in \mathrm{span}\{\mathbf{1}, \hat{z}\}. \tag{27}$$

Proof. 由定理 2.2，$\ker(J_{\mathrm{LN}}(z)) = \mathrm{span}\{\mathbf{1}, \hat{z}\}$。于是

$$J^{(\mathrm{Post})}(x) \cdot w = J_{\mathrm{LN}}(z) \cdot (I + J_F(x)) \cdot w = 0 \tag{28}$$

当且仅当

$$(I + J_F(x))w \in \ker(J_{\mathrm{LN}}(z)) = \mathrm{span}\{\mathbf{1}, \hat{z}\}. \tag{29}$$

■

注记 3.7 (梯度消失的几何解释). 定理 3.6揭示了 Post-LN 梯度消失的几何机制：若输入扰动 w 经过残差连接和子层后，其像 $(I + J_F(x))w$ 恰好落在 LayerNorm 的" 盲区"——均值方向 $\mathbf{1}$ 和标准化输出方向 $\hat{z}$ 张成的二维子空间，则该方向的梯度信息会被 LayerNorm 完全过滤，无法继续向前传播。这一结构性缺陷在多层堆叠时会持续存在，导致训练不稳定。

4 Pre-LN 的梯度分析

4.1 Jacobian 的结构

引理 4.1 (Pre-LN 的 Jacobian)**.** 设 Pre-LN 层为 $y = x + F(\mathrm{LN}(x))$。则从输入 x 到输出 y 的 Jacobian 为

$$J^{(\mathrm{Pre})}(x) = I + J_h(g(x)) \cdot J_g(x) = I + J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x). \tag{30}$$

4.2 最小奇异值的下界

在深度网络中，Jacobian 矩阵 $J \in \mathbb{R}^{d\times d}$ 描述了输出向量 y 对输入向量 x 的局部线性映射：

$$y \approx Jx \quad \text{在输入 } x_0 \text{ 附近.}$$

定义 4.2 (奇异值分解)**.** 任意矩阵 $J \in \mathbb{R}^{d\times d}$ 都可以进行奇异值分解（SVD）：

$$J = U\Sigma V^\top,$$

其中 $U, V \in \mathbb{R}^{d\times d}$ 是正交矩阵，$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$，且 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$ 为 J 的奇异值。

注记 4.3 (奇异值与梯度变化). 设输入的扰动方向为 $v \in \mathbb{R}^d$，则线性近似下输出的变化为

$$\delta y = Jv = U\Sigma V^\top v.$$

若令 $v = Ve_i$（e_i 为标准基向量），则

$$\delta y = U\Sigma V^\top (Ve_i) = U\Sigma e_i = \sigma_i U e_i.$$

由于 U 是正交矩阵，有

$$\|\delta y\| = \|\sigma_i U e_i\| = \sigma_i \|U e_i\| = \sigma_i \|e_i\| = \sigma_i.$$

同时，$\|v\| = \|Ve_i\| = \|e_i\| = 1$，因此

$$\|\delta y\| = \sigma_i \|v\|.$$

即梯度沿 $v = Ve_i$ 方向被放大或缩小的比例正好为奇异值 σ_i。

引理 4.4 (Weyl 不等式的推论)**.** 对任意矩阵 $A \in \mathbb{R}^{d\times d}$，其算子范数定义为 $\|A\| = \sigma_{\max}(A)$，则有

$$\sigma_{\min}(I + A) \geq 1 - \|A\|, \tag{31}$$

其中 $\sigma_{\min}(\cdot)$ 表示最小奇异值。

Proof. 我们使用 Weyl 不等式的奇异值形式。对于一般矩阵，经典 Weyl 不等式的推广表明：对任意 $B, C \in \mathbb{R}^{d\times d}$，

$$\sigma_i(B + C) \geq \sigma_i(B) - \sigma_{\max}(C), \quad i = 1, \ldots, d. \tag{32}$$

推导思路：不等式 eq. (32) 可由对称矩阵特征值的 Weyl 不等式推出。关键观察是：任意矩阵 M 的奇异值 $\sigma_i(M)$ 等于对称矩阵 $M^\top M$ 的特征值的平方根，即 $\sigma_i(M) = \sqrt{\lambda_i(M^\top M)}$。通过对 $(B + C)^\top(B + C)$、$B^\top B$、$C^\top C$ 应用对称矩阵的 Weyl 特征值不等式

$$\lambda_i(H_1 + H_2) \geq \lambda_i(H_1) + \lambda_d(H_2), \tag{33}$$

并利用矩阵范数与奇异值的关系 $\|C\| = \sigma_{\max}(C)$，即可得到 eq. (32)。

应用到本引理：取 $B = I$、$C = A$、$i = d$，由于单位矩阵的所有奇异值为 1，即 $\sigma_d(I) = 1$，代入 eq. (32) 得

$$\sigma_{\min}(I + A) = \sigma_d(I + A) \geq \sigma_d(I) - \|A\| = 1 - \|A\|. \tag{34}$$

■

注记 4.5 (Pre-LN 梯度稳定性). 从直观上看，矩阵 $I + A$ 可以被视为恒等映射加上一个非线性扰动 A。Weyl 不等式保证，只要扰动的算子范数 $\|A\|$ 不超过 1，最小奇异值不会归零。这正体现了 Pre-LN 结构中残差连接对梯度的保护作用：无论网络深度如何堆叠，只要每层扰动满足 $\|A^{(\ell)}\| < 1$，整体 Jacobian 仍然满秩，梯度沿每个方向均有正下界。

定理 4.6 (Pre-LN 的梯度下界)**.** 考虑 Pre-LN 层 $y = x + F(\mathrm{LN}(x))$。由引理 4.1，其 Jacobian 为

$$J^{(\mathrm{Pre})}(x) = I + J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x).$$

定义 $A(x) = J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x)$，则 $J^{(\mathrm{Pre})}(x) = I + A(x)$。若算子范数满足 $\|A(x)\| < 1$，则由引理 4.4，Jacobian 的最小奇异值满足

$$\sigma_{\min}(J^{(\mathrm{Pre})}(x)) \geq 1 - \|A(x)\| > 0. \tag{35}$$

这一正下界保证了以下性质：

(i) **满秩性**：$J^{(\mathrm{Pre})}(x)$ 是可逆的，$\ker(J^{(\mathrm{Pre})}(x)) = \{0\}$。

(ii) **梯度保持**：对任意非零向量 $w \in \mathbb{R}^d$，有

$$\|J^{(\mathrm{Pre})}(x) \cdot w\| \geq \sigma_{\min}(J^{(\mathrm{Pre})}(x))\|w\| \geq (1 - \|A(x)\|)\|w\| > 0, \tag{36}$$

即梯度在所有方向上均有正的下界，不存在梯度完全消失的方向。

Proof. 由引理 4.1，$J^{(\mathrm{Pre})}(x) = I + A(x)$，其中 $A(x) = J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x)$。
由引理 4.4（Weyl 不等式）：

$$\sigma_{\min}(I + A(x)) \geq 1 - \|A(x)\|. \tag{37}$$

若 $\|A(x)\| < 1$，则 $\sigma_{\min}(J^{(\mathrm{Pre})}(x)) > 0$，因此 $J^{(\mathrm{Pre})}(x)$ 可逆，从而 $\ker(J^{(\mathrm{Pre})}(x)) = \{0\}$。
对任意 $w \neq 0$，由奇异值的定义：

$$\|J^{(\mathrm{Pre})}(x)w\| \geq \sigma_{\min}(J^{(\mathrm{Pre})}(x))\|w\| \geq (1 - \|A(x)\|)\|w\| > 0. \tag{38}$$

■

5 Post-LN 与 Pre-LN 的对比总结

Post-LN 与 Pre-LN 的主要区别体现在每一层的 Jacobian 结构及其对梯度传播的影响，各项结构性差异和梯度性质的对比总结如表格 1所示。

表格 1: Post-LN 与 Pre-LN 的结构性对比

性质	**Post-LN**	**Pre-LN**
Jacobian 形式	$J_{\mathrm{LN}}(z) \cdot (I + J_F(x))$	$I + J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x)$
恒等通路 I 的位置	被 J_{LN} 左乘	直接加在最外层
秩的上界	$\leq d - 2$	$= d$（在 $\|A\| < 1$ 时）
零空间维度下界	≥ 2	$= 0$（在 $\|A\| < 1$ 时）
梯度完全消失的方向	存在（至少 2 个独立方向）	不存在

6 补充：多层堆叠网络的 Jacobian 分析

6.1 局部线性化的理论基础

在分析多层网络时，我们采用**局部线性化近似**（local linearisation approximation）。这一方法的理论依据如下：

假设 6.1 **(局部线性化假设).** 设可微映射 $g : \mathbb{R}^d \to \mathbb{R}^d$ 在点 x_0 处的一阶泰勒展开为

$$g(x_0 + \delta x) = g(x_0) + J_g(x_0)\delta x + o(\|\delta x\|), \tag{39}$$

其中 $J_g(x_0)$ 是 Jacobian 矩阵，$o(\|\delta x\|)$ 表示高阶无穷小项满足

$$\lim_{\|\delta x\| \to 0} \frac{\|o(\|\delta x\|)\|}{\|\delta x\|} = 0. \tag{40}$$

在分析梯度传播时，我们关注的是 Jacobian 矩阵 $J_g(x_0)$ 的性质，它刻画了函数在 x_0 附近的**局部线性行为**。

注记 6.2 (多层复合的线性化). 对于 L 层网络 $x^{(L)} = f^{(L)} \circ \cdots \circ f^{(1)}(x^{(0)})$，设第 ℓ 层在点 $x^{(\ell-1)}$ 处的 Jacobian 为 $J^{(\ell)} = J_{f^{(\ell)}}(x^{(\ell-1)})$。由链式法则，从输入到输出的 Jacobian 为

$$\frac{\partial x^{(L)}}{\partial x^{(0)}} = J^{(L)} \cdots J^{(1)} = \prod_{\ell=1}^{L} J^{(\ell)}. \tag{41}$$

6.2 L 层 Post-LN 的 Jacobian

定理 6.3 (L 层 Post-LN 的 Jacobian)**.** 设 L 层 Post-LN 网络，第 ℓ 层的输出为

$$x^{(\ell)} = \mathrm{LN}(x^{(\ell-1)} + F^{(\ell)}(x^{(\ell-1)})). \tag{42}$$

记局部线性化 Jacobian 为

$$J^{(\ell)} = \frac{\partial x^{(\ell)}}{\partial x^{(\ell-1)}} = J_{\mathrm{LN}}(z^{(\ell)}) \cdot (I + J_{F^{(\ell)}}(x^{(\ell-1)})), \quad z^{(\ell)} = x^{(\ell-1)} + F^{(\ell)}(x^{(\ell-1)}). \tag{43}$$

则从 $x^{(0)}$ 到 $x^{(L)}$ 的 Jacobian 为矩阵乘积

$$J^{(1:L)} = \prod_{\ell=1}^{L} J^{(\ell)} = \prod_{\ell=1}^{L} J_{\mathrm{LN}}(z^{(\ell)}) \cdot (I + J_{F^{(\ell)}}(x^{(\ell-1)})). \tag{44}$$

由矩阵乘积的秩不等式 $\mathrm{rank}(AB) \le \min(\mathrm{rank}(A), \mathrm{rank}(B))$ 以及单层 Post-LN Jacobian 的秩上界（定理 3.2：$\mathrm{rank}(J^{(\ell)}) \le d - 2$），有

$$\mathrm{rank}(J^{(1:L)}) \le \min_{\ell=1,\dots,L} \mathrm{rank}(J^{(\ell)}) \le d - 2. \tag{45}$$

因此，无论堆叠多少层，Post-LN 网络至少存在 2 个梯度被完全抹除的方向。

6.3 L 层 Pre-LN 的 Jacobian

定理 6.4 (L 层 Pre-LN 的 Jacobian)**.** 设 L 层 Pre-LN 网络，第 ℓ 层的输出为

$$x^{(\ell)} = x^{(\ell-1)} + F^{(\ell)}(\mathrm{LN}(x^{(\ell-1)})). \tag{46}$$

在局部线性化近似下，定义

$$A^{(\ell)} = J_{F^{(\ell)}}(\mathrm{LN}(x^{(\ell-1)})) \cdot J_{\mathrm{LN}}(x^{(\ell-1)}), \qquad J^{(\ell)} = I + A^{(\ell)}, \quad J^{(1:L)} = \prod_{\ell=1}^{L} J^{(\ell)}. \tag{47}$$

若对所有 ℓ 有 $\|A^{(\ell)}\| < 1$，则每层 Jacobian 满秩，且

$$\sigma_{\min}(J^{(1:L)}) \ge \prod_{\ell=1}^{L} \sigma_{\min}(I + A^{(\ell)}) \ge \prod_{\ell=1}^{L} (1 - \|A^{(\ell)}\|) > 0. \tag{48}$$

因此，即使网络很深，Pre-LN 的梯度在任意方向上都不会被完全抹除，保证了梯度下界。

Proof. 由局部线性化的乘积公式 $J^{(1:L)} = \prod_{\ell=1}^{L}(I + A^{(\ell)})$，每个因子 $(I + A^{(\ell)})$ 的最小奇异值满足 $\sigma_{\min}(I + A^{(\ell)}) \ge 1 - \|A^{(\ell)}\|$（引理 4.4）。奇异值乘法性质保证

$$\sigma_{\min}\Big(\prod_{\ell=1}^{L}(I + A^{(\ell)})\Big) \ge \prod_{\ell=1}^{L} \sigma_{\min}(I + A^{(\ell)}) \ge \prod_{\ell=1}^{L} (1 - \|A^{(\ell)}\|) > 0, \tag{49}$$

当 $\|A^{(\ell)}\| < 1$ 对所有 ℓ 成立时。由 $\sigma_{\min}(J^{(1:L)}) > 0$ 可知 $J^{(1:L)}$ 满秩，梯度在任意方向上均非零。 ■

7 结论

本文从线性代数的角度严格分析了 Post-LN 与 Pre-LN Transformer 架构的梯度传播性质。核心推断可概括如下：

1. **LayerNorm 的 Jacobian 结构**：

$$\ker(J_{\mathrm{LN}}(z)) = \mathrm{span}\{\mathbf{1}, \hat{z}\}, \quad \dim(\ker(J_{\mathrm{LN}}(z))) = 2,$$

即 LayerNorm 的 Jacobian 在均值方向 $\mathbf{1}$ 和标准化输出方向 $\hat{z}$ 上具有二维零空间。这一内在几何性质是后续分析的基础。

2. **Post-LN 与 Pre-LN 的结构性差异**：

- 在 **Post-LN** 中，LayerNorm 作用于残差连接之后，整层 Jacobian 为

$$J^{(\mathrm{Post})}(x) = J_{\mathrm{LN}}(z) \cdot (I + J_F(x)),$$

由于左乘 $J_{\mathrm{LN}}(z)$，恒等通路被 LayerNorm 的零空间" 过滤"，导致整层 Jacobian 秩满足

$$\mathrm{rank}(J^{(\mathrm{Post})}) \leq d - 2,$$

即至少存在两个方向的梯度在该层被完全消除。

- 在 **Pre-LN** 中，LayerNorm 作用于子层输入，整层 Jacobian 为

$$J^{(\mathrm{Pre})}(x) = I + J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x),$$

恒等映射 I 位于最外层，未被 LayerNorm 影响。若 $\|J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x)\| < 1$，则

$$\sigma_{\min}(J^{(\mathrm{Pre})}(x)) \geq 1 - \|J_F(\mathrm{LN}(x)) \cdot J_{\mathrm{LN}}(x)\| > 0,$$

保证 Jacobian 满秩，梯度在所有方向上均有正下界。

3. **实践启示**：Post-LN 架构中 LayerNorm 的零空间直接导致整层梯度退化，这解释了其训练不稳定性（需要小学习率、warm-up、对深度敏感）。相比之下，Pre-LN 通过将恒等映射置于外层，避免了零空间的传播，使其更适合深层网络训练。