量化嵌入的谱平坦化

Junlin Huang[1], **Wenyi Fang**[2], **Zhenheng Tang**[3],**Yuxin Wang**[2],**Xueze Kang**[1],**Yang Zheng**[2]、**Bo Li**[3],和 **Xiaowen Chu**[1]

[1]香港科技大学（广州）[2]华为技术有限公司[3]香港科技大学

arXiv:2602.00969v1 [cs.LG] 1 Feb 2026

摘要

以超低精度训练大语言模型（大语言模型）受到严重阻碍，其不稳定性源于离散量化约束与语言数据固有的重尾谱特性之间的冲突。通过形式化齐夫统计与随机矩阵理论之间的联系，我们证明，嵌入奇异值谱中的幂律衰减是实现语义编码的基本必要条件。我们推导出理论界限，表明均匀量化会引入噪声底，不成比例地截断谱尾，从而导致谱平坦化，并使表示的稳定秩严格增加。对包括 GPT-2 和 TinyLlama 在内的多种架构进行的实证验证表明，这种几何退化会引发表示坍缩。本文不仅量化了大语言模型的谱敏感性，还确立了谱保真度是实现稳定低比特优化的必要条件。

1 引言

训练大语言模型（LLM）所需的计算资源不断增加，这使得低比特量化成为维持可扩展性的必然范式转变。近年来，业界已成功从标准 FP32 和 BF16 格式转向 FP8训练 [1, 2, 3], ，在不牺牲收敛稳定性的前提下，显著提升了内存效率和吞吐量。目前，研究前沿正进一步转向 4 位精度。值得注意的是，英伟达关于 Blackwell架构的技术报告显示，新兴的 NVFP4 格式可将内存占用降低 1.8×，使 B200 GPU 的通用矩阵乘法（GeMM）速度比基于 Hopper 的 H100 [4]，的 FP8 吞吐量提升超过 5× ，凸显了 FP4 训练在新一代基础模型中的变革潜力 [5]。

然而，将训练前沿拓展至 FP4 面临着独特的理论障碍，无法仅靠外推以往的成功经验来克服。随着精度降至 4 位，表示对动态范围和分辨率施加了指数级严格的限制。这与大语言模型中参数、激活值和梯度固有的高动态范围分布形成了根本冲突。关键在于，这些分布特性并非任意产生的伪象，而是深度学习低秩假设的内在体现 [6, 7]。尽管大语言模型的权重矩阵名义上是满秩的，但其谱结构呈现重尾特征——受齐夫定律支配 [8]——其中少数主导奇异值承载了绝大部分能量，而漫长且脆弱的尾部则编码着细粒度语义信息。

我们认为，FP4 等超低比特量化的主要风险并不止于简单的标量近似误差；它还在于谱尾的结构退化。当量化噪声底侵入这些脆弱的尾部奇异值的幅度时，模型会发生谱平坦化。至关重要的是，尽管更高的稳定秩通常与表示学习中的有益各向同性相关，但我们证明，在量化语境下，这一增长意味着判别能力的病态丧失。这一现象实际上截断了复杂推理所必需的细微特征子空间，导致我们形式化定义的表示坍缩。现有分析往往忽视了量化的离散约束与支配自然语言的连续齐夫式

定律之间的这一关键相互作用。在本文中，我们通过建立一个严谨的理论框架来量化这一特征，从而弥合这一差距：

贡献一： 我们对自然语言的统计特性与线性层的谱结构之间的联系进行了形式化描述。考虑一个线性层 $\mathbf{Y} = \mathbf{X}\mathbf{W}$，其中 $\mathbf{X}$ 表示输入词元嵌入，$\mathbf{W}$ 是权重矩阵。基于实证研究表明大规模语料库中的词频服从齐夫分布，我们证明该分布决定了 $\mathbf{X}$ 和权重梯度 $\nabla_{\mathbf{W}} = \mathbf{X}^T\mathbf{G}$ 的奇异值呈幂律衰减，其中 $\mathbf{G}$ 是相对于输出的损失梯度。具体而言，我们推导出前 r 个奇异值满足以下界：

$$\sigma_k(\mathbf{X}) \leq C \cdot k^{-\alpha/2}, \quad \sigma_k(\nabla_{\mathbf{W}}) \leq (CM) \cdot k^{-\alpha/2}, \tag{1}$$

其中 $1 \leq k \leq r, C$ 是常数，$\sigma_k(\cdot)$ 是第 k 大奇异值，α 是衰减指数，且 $\|\mathbf{G}\|_2 \leq M$。

贡献二： 我们证明，分块对称均匀量化这一现代低精度格式（如 MXFP4 [9] 和 NVFP4 [10], ）广泛采用的机制，会有效改变 $\mathbf{X}$ 和$\nabla_{\mathbf{W}}$ 的谱性质。我们证明，量化会破坏上述幂律衰减，从而提高表示的稳定秩。

本文其余部分安排如下。第2节概述统计基础。第3节运用随机矩阵理论推导大语言模型的重尾谱结构。第4节证明量化会截断谱尾并提高稳定秩。第5节通过不同模型架构的实证验证这些理论预测。第6节总结全文，并展望未来研究方向。

2 预备知识与理论基础

本节将建立分析所需的记号和概率框架。我们首先定义用于描述高维极限的渐近行为。随后，我们将语言数据的统计特性和嵌入空间的几何结构形式化。最后，我们介绍矩阵扰动理论和集中不等式中的关键工具，以便界定量化噪声对谱的影响。补充定义见附录B节。

2.1 渐近记号

为分析大规模模型的谱性质，我们考虑维数为 $d \to \infty$ 的区间。我们定义以下渐近关系，以刻画序列的收敛行为。

定义 2.1(渐近等价**)。** 若两个序列 $A(d)$ 和 $B(d)$ 满足当 $d \to \infty$ 时，其比值收敛于 *1*，则称它们渐近等价，记作 $A(d) \sim B(d)$：

$$A(d) \sim B(d) \iff \lim_{d\to\infty} \frac{A(d)}{B(d)} = 1. \tag{2}$$

或者等价地说，相对误差趋于 *0*：

$$\lim_{d\to\infty} \left| \frac{A(d) - B(d)}{B(d)} \right| = 0. \tag{3}$$

定义 2.2 (渐近上界**)。** 对于依赖于 d, $B(d)$ 的两个序列 $A(d)$ 和 $B(d)$，当且仅当存在 ,，使得对于所有 ，以下条件成立时，(d) 是 $A(d)$ 的渐近上界：

$$|A(d)| \leq C \cdot |B(d)|, \tag{4}$$

记作 $A(d) = \mathcal{O}(B(d))$。

2.2 语言表示的统计结构

为分析大语言模型中重尾谱的起源，我们首先对驱动训练动态的输入数据分布进行建模。基于关于词元频率和嵌入几何的两个广受认可的假设，我们确定数据嵌入矩阵的谱性质。

假设 2.3 ([8])。 假设数据流由大小为 V 的词汇表 $\mathcal{V} = \{w_1, w_2, \ldots, w_V\}$ 生成。令 p_k 表示第 k 个词元的出现概率。我们假设 $\{p_k\}_{1 \le k \le V}$ 渐近服从齐夫幂律分布：

$$p_k \sim k^{-\alpha}, \quad k = 1, 2, \ldots, V, \tag{5}$$

其中 $\alpha > 1$是衰减参数。

假设 2.3 以计量语言学为依据。针对大规模语料库（例如 Wikipedia [11],Common Crawl [12]）的实证研究一再表明，词频与排名之间存在幂律关系。这意味着，由这些词元构成的嵌入矩阵必然会呈现偏斜的重尾奇异值分布。

假设 2.4。令 $v_k \in \mathbb{R}^d$ 表示与第 k个词元对应的嵌入向量，并将其归一化为单位范数，使得 $\|v_k\|_2 = 1$。其协方差矩阵 Σ定义为按概率加权的外积：

$$\Sigma \triangleq \sum_{k=1}^{V} p_k v_k v_k^T. \tag{6}$$

我们提出，在高维极限（$d \to \infty$）下，这些嵌入向量表现出准正交性。具体而言，对于任意一对不同的指标 $i \neq j$:

$$|\langle v_i, v_j \rangle| = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right). \tag{7}$$

假设 2.4 基于高维空间的基本几何性质：相互独立的随机向量倾向于彼此正交。这一现象得到了高维概率理论的印证（通常与 Johnson-Lindenstrauss 引理 [13]相关）。该理论表明，两个随机单位向量的内积期望为零，方差为 $1/d$。因此，在较大 d 区间内，不同语义向量之间的余弦相似度按 $\mathcal{O}(d^{-1/2})$ 进行缩放。方程（7）是引理 3.1 证明所必需的。

2.3 谱扰动分析工具

为量化量化噪声引起的谱失真，我们借助随机矩阵理论的基本工具 [14]。

定理 2.5。 设 $\{\mu_n\}_{n \ge 1}$ 是定义在 $\mathbb{R}$上的一列概率测度，并令 $\{m_n(z)\}_{n \ge 1}$ 表示其对应的斯蒂尔斯变换。若存在函数 $m(z)$，其定义于上半复平面 $\mathbb{C}^+$上，使得对于所有 $z \in \mathbb{C}^+$，逐点极限存在：

$$\lim_{n \to \infty} m_n(z) = m(z), \tag{8}$$

且 $m(z)$ 是概率测度 μ的斯蒂尔切斯变换，则 $\{\mu_n\}$ 弱收敛于 μ:

$$\mu_n \xrightarrow{w.} \mu. \tag{9}$$

定理 2.5 使我们能够通过分析其斯蒂尔切斯变换来刻画大型协方差矩阵的极限谱分布，这是随机矩阵理论中的一项标准技术。

˜定理 2.6 ([15]). 设 $A \in \mathbb{R}^{m \times n}$ 为一个矩阵， $A = A + E$ 为其扰动矩阵，奇异值分别为$\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$ 和 $\tilde{\sigma}_1 \ge \tilde{\sigma}_2 \ge \cdots \ge 0$。则对每个 k，第 k 个奇异值之间的绝对差受扰动矩阵的谱范数界定：

$$|\tilde{\sigma}_k - \sigma_k| \le \|E\|_2. \tag{10}$$

由于复杂的特征向量相互作用，直接分析每个特定 $\tilde{\sigma}_k$ 的分布难以处理。定理 2.6 使我们能够将单个奇异值偏移与特定噪声结构解耦，并利用全局谱范数 $\|E\|_2$ 对它们进行一致界定。

定理 2.7 ([16])。 考虑一个由相互独立的中心化随机矩阵组成的有限序列 $\{X_k\} \subseteq \mathbb{R}^{d_1 \times d_2}$。假设每个矩阵的谱范数均一致有界：

$$\|X_k\|_2 \le R \quad \textit{almost surely}, \tag{11}$$

对于某个常数 R。令 δ^2 为如下定义的方差统计量：

$$\delta^2 = \max\left\{ \left\| \sum_k \mathbb{E}[X_k X_k^T] \right\|_2, \left\| \sum_k \mathbb{E}[X_k^T X_k] \right\|_2 \right\}. \tag{12}$$

那么，对于任意阈值 $t > 0$，随机矩阵 $Y \triangleq \sum_k X_k$ 的谱范数超过 t 的概率由以下指数项界定：

$$P(\|Y\|_2 \geq t) \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right). \tag{13}$$

定理 2.7 证明，累积量化噪声的谱范数不会随矩阵维度线性增长。相反，它会紧密地集中在零附近，并呈现由方差统计量 δ^2 控制的次高斯尾行为。尾概率的这种指数衰减意味着，较大的谱偏差极为罕见，因此我们有充分把握将量化噪声视为有界扰动。

3 语言模型嵌入的低秩特性

本节将严谨推导嵌入的谱性质。我们的分析分为四步：（1）将词元频率与总体协方差谱联系起来；（2）将尾谱建模为有效白噪声；（3）应用随机矩阵理论刻画样本协方差特征值；（4）推导输入嵌入和梯度的奇异值界。详细证明见附录 C 节。

3.1 总体协方差与谱分解

我们首先建立词元语言分布与其嵌入协方差矩阵谱衰减之间的联系。

引理 3.1。 在假设 *2.3* 和假设 *2.4* 下，第 k 大特征值 Σ（记为 τ_k）的渐近行为由概率权重 p_k 决定：

$$\tau_k \sim p_k \sim k^{-\alpha}, \tag{14}$$

其中 $\alpha > 1$。

引理 3.1 表明，Σ 的谱结构继承了词元概率分布 p_k 所具有的幂律衰减特征。基于此，我们在后续分析中采用以下显式谱模型：

$$\tau_k = L \cdot k^{-\alpha}, \quad \forall k \in \{1, \ldots, d\}, \tag{15}$$

其中，$L > 0$ 为能量尺度，$\alpha > 1$ 为衰减指数。
为了有效应用谱分析工具，我们必须区分主导信号分量与残余噪声底。下面的引理将这一分解形式化。

引理 3.2。 设协方差矩阵 $\Sigma \in \mathbb{R}^{d \times d}$ 的特征值遵循幂律 $\tau_k = L \cdot k^{-\alpha}$，其中 $\alpha > 1$。定义背景噪声水平 $\nu^2(d)$ 为截断前 r 个主成分后残余特征值的平均能量：

$$\nu^2(d) \triangleq \frac{1}{d-r} \sum_{j=r+1}^{d} \tau_j = \frac{L}{d-r} \sum_{j=r+1}^{d} j^{-\alpha}. \tag{16}$$

设 c 为任意非负实常数。对于任意固定的头部秩 $k \in [1, r]$，信号主导噪声阈值：

$$\tau_k > \nu^2(d) \cdot (1 + \sqrt{c}). \tag{17}$$

反之，对于任意固定的尾秩 $k \in (r, d]$，特征值降至阈值以下：

$$\tau_k \leq \nu^2(d) \cdot (1 + \sqrt{c}). \tag{18}$$

此外，噪声水平的量级为 $\nu^2(d) = \mathcal{O}(d^{-1})$。

3.2 矩阵相变

我们证明中的一个关键步骤是证明重尾残差谱在渐近意义上表现得如同白噪声。我们通过分析谱测度的斯蒂尔切斯变换来实现这一点。

引理 3.3。 考虑引理 *3.2* 中定义的残余特征值集合 $\mathcal{T}_A \triangleq \{\tau_j = Lj^{-\alpha}\}_{j=r+1}^{d}$。 $\mathcal{T}_A$ 的斯蒂尔切斯变换为：

$$m_{Noise}(z) \triangleq \frac{1}{d-r}\sum_{j=r+1}^{d}\frac{1}{\tau_j - z}. \tag{19}$$

考虑一个参考白噪声谱 $\mathcal{T}_B$，它由 $d-r$ 个相同的值组成，且这些值均等于背景噪声水平ν^2（d）：

$$\mathcal{T}_B \triangleq \underbrace{\{\nu^2(d), \ldots, \nu^2(d)\}}_{d-r}, \tag{20}$$

其斯蒂尔切斯变换为：

$$m_{White}(z) \triangleq \frac{1}{d-r}\sum_{j=r+1}^{d}\frac{1}{\nu^2(d) - z} = \frac{1}{\nu^2(d) - z}. \tag{21}$$

那么，当 $d \to \infty$ 时，这两个变换渐近等价：

$$\lim_{d\to\infty} |m_{Noise}(z) - m_{White}(z)| = 0. \tag{22}$$

结合引理 3.3 与斯蒂尔切斯变换的连续性（定理 2.5），我们得到噪声谱的极限分布。

引理 3.4。 将协方差矩阵 Σ分解为信号部分和噪声部分：

$$\Sigma = \Sigma_{Signal} + \Sigma_{Noise}, \tag{23}$$

其中

$$\begin{aligned} \textit{Spec}(\Sigma_{Signal}) &= (\tau_1, \ldots, \tau_r, 0, \ldots, 0), \\ \textit{Spec}(\Sigma_{Noise}) &= (0, \ldots, 0, \tau_{r+1}, \ldots, \tau_d) \end{aligned}. \tag{24}$$

当 $d \to \infty$ 时， Σ_{Noise} 的经验谱分布 （*ESD*）弱收敛到以噪声方差 $\nu^2(d)$为中心的狄拉克质量：

$$\textit{ESD}(\Sigma_{Noise}) \xrightarrow{w.} \delta_{\nu^2(d)}, \tag{25}$$

其中$\delta_{\nu^2(d)}$ 定义为：在 $\nu^2(d)$ 处的概率质量为 *1*，其他位置为 *0*。

引理 3.4 证明了 一个关键事实：尽管存在幂律衰减，协方差矩阵的尾部仍表现出谱集中现象， 其实际表现如同方差为 $\nu^2(d)$ 的白噪声。

现在，我们转而关注训练过程中观测到的样本协方差矩阵，其定义如下： ：

$$S_N = \frac{1}{N}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{d\times d}, \tag{26}$$

其中，$\mathbf{X} \in \mathbb{R}^{d\times N}$ 是词元嵌入的数据矩阵， N 表示有效批量大小。至关重要的是，$\mathbf{X}$ 的各列根据词元概率 $\{p_k\}$ 从词汇表 $\{v_k\}_{k=1}^{V}$ 中独立采样。因此， S_N 可作为假设 2.4 中定义的总体协方差 Σ 的经验估计量，并满足无偏性条件：

$$\mathbb{E}[S_N] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T\right] = \sum_{k=1}^{V} p_k v_k v_k^T = \Sigma, \tag{27}$$

其中， x_i 是$\mathbf{X}$的第 i 列。在渐近机制下，当 $d, N \to \infty$ 且比值为 $d/N \to c \in (0, \infty)$ 时， S_N 的样本特征值 λ_k 的行为由 Baik-Ben Arous-Péché （ BBP ）相变定理 [17] 刻画。

定理 3.5 （BBP 相变）。 设 τ_k 为总体协方差 Σ的第 k 大特征值， λ_k为样本协方差 S_N中对应的特征值。给定 $\textit{ESD}(\Sigma_{Noise}) \xrightarrow{w} \delta_{\nu^2(d)}$，谱相变发生在阈值 $\nu^2(d)(1+\sqrt{c})$处：

1. 超临界区域：如果总体特征值满足分离条件：

$$\tau_k > \nu^2(d)(1+\sqrt{c}), \tag{28}$$

那么相应的样本特征值 λ_k 会与谱主体分离，并几乎必然收敛到一个取决于 τ_k的确定性极限：

$$\lambda_k \xrightarrow{a.s.} \rho(\tau_k) = \tau_k \left(1 + \frac{c \cdot \nu^2(d)}{\tau_k - \nu^2(d)}\right). \tag{29}$$

2. 亚临界区间：如果总体特征值低于阈值：

$$\tau_k \leq \nu^2(d)(1 + \sqrt{c}), \tag{30}$$

那么 λ_k 会收敛到 *Marchenko-Pastur* 谱主体分布 *[18]*的右边缘：

$$\lambda_k \xrightarrow{a.s.} \nu^2(d)(1 + \sqrt{c})^2. \tag{31}$$

具体而言，$\mathit{Spec}(\Sigma_{Signal})$处于超临界区域，而 $\mathit{Spec}(\Sigma_{Noise})$ 处于亚临界区间。

定理 3.5 在 $\nu^2(d)(1 + \sqrt{c})$处定义了可检测性边界。在超临界区域，超过此阈值的信号会作为离群值从噪声谱主体中分离出来。在亚临界区间，信号会收敛到 Marchenko-Pastur 谱主体的边缘，并变得无法与噪声区分。这证明了将 Spec(Σ_{Signal}) 与 Spec(Σ_{Noise}) 分开是合理的。为简化非线性极限 $\rho(\tau_k)$，下面的引理推导出$d \to \infty$的渐近等价，说明主导样本特征值重构了总体幂律，而尾部则收敛到噪声底。

引理 3.6。样本特征值 λ_k 表现出如下渐近行为：

$$\lambda_k \sim \begin{cases} \tau_k & 1 \leq k \leq r \\ \nu^2(d)(1 + \sqrt{c})^2 & r < k \leq d \end{cases}, \tag{32}$$

其中 $\tau_k \in$ 谱(Σ_{Signal})当 $1 \leq k \leq r$，且 $\tau_k \in$ 谱(Σ_{Noise}) 当 $r < k \leq d$。

3.3 奇异值界

最后，我们将这些结果从协方差域转换回数据嵌入矩阵 $\mathbf{X}$ 的奇异值以及权重梯度 $\nabla_{\mathbf{W}} = \mathbf{X}^T\mathbf{G}$，其中 $\mathbf{G}$ 是损失相对于输出的梯度。由于 $S_N = \frac{1}{N}\mathbf{X}\mathbf{X}^T$，$\mathbf{X}$ 的奇异值通过 $\sigma_k(\mathbf{X}) = \sqrt{N\lambda_k}$ 与 S_N 的特征值相关联。

定理 3.7. 设 $\sigma_k(\mathbf{X})$ 为数据嵌入矩阵 $\mathbf{X} \in \mathbb{R}^{d\times N}$ 的第 k 大奇异值。当 d、$N \to \infty$ 且 $d/N \to c$ 时，奇异值谱遵循幂律曲线：

$$\sigma_k(\mathbf{X}) \sim \begin{cases} \sqrt{NL} \cdot k^{-\alpha/2}, & 1 \leq k \leq r \\ \sqrt{N\nu^2(d)} \cdot (1 + \sqrt{c}), & r < k \leq d \end{cases}. \tag{33}$$

此外，存在常数 C_1，使得：

$$\sigma_k(\mathbf{X}) \leq \begin{cases} C_1 \cdot k^{-\alpha/2}, & 1 \leq k \leq r \\ C_1 \cdot r^{-\alpha/2}, & r < k \leq d \end{cases}. \tag{34}$$

应用微积分链式法则和标准谱范数不等式，我们推导出权重梯度的谱界。

定理 3.8。考虑一个线性层 $\mathbf{Y} = \mathbf{X}\mathbf{W}$，其中 $\mathbf{Y}$ 是输出矩阵，$\mathbf{X}$ 是数据嵌入矩阵，$\mathbf{W}$ 是权重矩阵。设$\mathbf{G}$ 为损失相对于输出的梯度。假设训练稳定性（即$\|\mathbf{G}\|_2 \leq M$对于某个有限常数 M），则权重梯度矩阵 $\nabla_{\mathbf{W}} = \mathbf{X}^T\mathbf{G}$ 的奇异值有界于：

$$\sigma_k(\nabla_{\mathbf{W}}) \leq \begin{cases} (MC_1) \cdot k^{-\alpha/2}, & 1 \leq k \leq r \\ (MC_1) \cdot r^{-\alpha/2}, & r < k \leq d \end{cases}. \tag{35}$$

4 量化提高稳定秩

在定理 3.7 和定理 3.8 中，我们证明了训练过程的关键组成部分 $\mathbf{X}$ 和 $\nabla_{\mathbf{W}}$ 呈现幂律谱衰减。为了进行统一分析，本节将这些特定矩阵抽象为一个通用的随机矩阵 A。因此，下文讨论的矩阵 A 应被视为高精度嵌入 $\mathbf{X}$ 或梯度 $\nabla_{\mathbf{W}}$ 的理论替代物，两者都满足幂律假设。详细证明见附录 D 节。

4.1 问题设置与谱假设

令 $A \in \mathbb{R}^{m\times n}$ 表示目标高精度矩阵（代表 $\mathbf{X}$ 或 $\nabla_{\mathbf{W}}$）。与大语言模型中的实证观察以及我们在第3节中的理论推导一致，我们通过以下假设对 A 的谱结构进行形式化描述。

假设 4.1。 矩阵 A 的奇异值 $\{\sigma_k(A)\}$ 被假设遵循以下分布：

$$\sigma_k(A) \sim \begin{cases} \mu \cdot k^{-\alpha/2}, & 1 \le k \le r \\ \mu \cdot r^{-\alpha/2}, & r < k \le \min(m,n) \end{cases}, \tag{36}$$

其中，$\mu > 0$ 是谱幅度常数，$\alpha > 1$ 是该模型领域特有的衰减率参数。

令 $\tilde{A}$ 表示 A 的逐分量量化对应矩阵。量化误差矩阵定义为差值 $E = A - \tilde{A}$，其元素为 $E_{ij} = e_{ij}$。为便于进行逐分量谱分析，我们将误差矩阵 E 分解为一组独立基矩阵之和：

$$E = \sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij}, \tag{37}$$

其中，每个 $Z_{ij} \in \mathbb{R}^{m\times n}$ 都是一个稀疏矩阵，仅在条目 $(i,\ j)$ 处包含误差 e_{ij}，其余位置均为零。

在现代低精度训练机制（如 MXFP4 和 NVFP4）中，底层量化算子可以统一为分块对称均匀量化公式。具体而言，对于输入标量 $a_{ij} = A_{ij}$，量化器 $Q_s(a_{ij})$ 定义为：

$$Q_s(a_{ij}) = s \cdot \text{round}\left(\frac{a_{ij}}{s}\right), \tag{38}$$

其中，$s = a_{\max}/L$ 表示量化步长。这里，$a_{\max} = \max_{i,\ j}|A_{ij}|$ 表示动态范围，而L 是给定位宽下可表示的最大整数值（例如，$L = 7$ 用于有符号 4 位整数表示）。我们采用关于量化误差 $e_{ij} = Q_s(a_{ij}) - a_{ij}$分布的标准统计假设，如假设 4.2 所述。

假设 4.2。 假定量化误差 $e_{ij} = Q_s(a_{ij}) - a_{ij}$ 是一个均值为零 $\mathbb{E}[e_{ij}] = 0$ 且方差有限的中心化随机变量 $Var(e_{ij}) = B \propto s^2$。

4.2 噪声谱范数的界定

分析量化效应的一项主要挑战在于误差具有随机性。为推导出对误差谱范数 $\|E\|_2$的确定性控制，我们采用定理 2.7。要将该不等式应用于具体的误差分解 $E = \sum_{i,\ j} Z_{ij}$，我们必须先计算方差统计量 δ^2（方程 (12)），这需要求出期望协方差和的谱范数。下面的引理通过确立交叉项的独立性，简化了这一计算。

引理4.3。 对于分解为独立基矩阵 Z_{ij} 的量化误差矩阵 E，其中 e_{ij} 遵循假设 *4.2*，期望二阶矩矩阵之和满足：

$$\begin{aligned} \sum_{i,j}\mathbb{E}[Z_{ij}Z_{ij}^T] &= \mathbb{E}[EE^T], \\ \sum_{i,j}\mathbb{E}[Z_{ij}^T Z_{ij}] &= \mathbb{E}[E^T E]. \end{aligned} \tag{39}$$

引理4.3 至关重要，因为它消除了误差矩阵中的空间相关性。它本质上指出，由于独立性，即使在矩阵域中，和的方差也等于各项方差之和。由定理 2.7 和引理4.3，可得以下定理。

定理 4.4。 考虑一系列维数为 $m \times n$ 的独立基矩阵 $\{Z_{ij}\}$，其中误差项 e_{ij} 满足假设 *4.2*。设 δ^2 为如下定义的方差统计量：

$$\delta^2 = \max\left\{ \left\|\sum_{i,j}\mathbb{E}[Z_{ij}Z_{ij}^T]\right\|_2, \left\|\sum_{i,j}\mathbb{E}[Z_{ij}^T Z_{ij}]\right\|_2 \right\}. \tag{40}$$

那么，对于任意阈值 $t > 0$，总和 $E = \sum_{i,\ j} Z_{ij}$ 的谱范数超过 t 的概率有如下界：

$$P(\|E\|_2 \geq t) \leq (m+n) \cdot \exp\left(-\frac{t^2}{C \cdot \max(m,n) \cdot s^2}\right), \tag{41}$$

其中，C 是与噪声的次高斯范数相关的通用常数。

上述推导得出了类似高斯的尾界，定量表明，全局谱噪声 $\|E\|_2$大致按 $\sqrt{\max(m,\ n)} \cdot s$的比例缩放。虽然这表明，对于极宽的层，绝对谱误差会增大，但表示的稳定性取决于这种噪声的相对影响，接下来我们将对此进行分析。

我们将尾界函数 $F(t)$ 定义为公式（41）的右侧：

$$F(t) \triangleq (m+n) \cdot \exp\left(-\frac{t^2}{C \cdot \max(m,n) \cdot s^2}\right). \tag{42}$$

注意，对于所有 $t > 0$，导数 $F'(t) < 0$，这确保了 $F(t)$ 严格单调递减。这种单调性对于后续分析中其反函数的存在至关重要。

4.3 奇异值的相对误差分析

在建立了全局谱误差 $\|E\|_2$的概率上界后，我们现在考察这种全局噪声如何转化为单个奇异值的扰动。为了衔接误差矩阵的谱范数与奇异值的具体偏差，我们在样本空间 Ω 上定义相对误差违约事件 $\mathcal{R}(\epsilon_k)$。具体而言，$\mathcal{R}(\epsilon_k)$ 包含所有满足相对扰动超过 ϵ_k的结果 $\omega \in \Omega$：

$$\mathcal{R}(\epsilon_k) \triangleq \left\{ \omega \in \Omega : \frac{|\sigma_k(\tilde{A}(\omega)) - \sigma_k(A)|}{\sigma_k(A)} > \epsilon_k \right\}. \tag{43}$$

为简洁起见，我们在记号中略去对 ω 的依赖，并将该事件记为 $\mathcal{R}(\epsilon_k) = \left\{ \frac{|\sigma_k(\tilde{A}) - \sigma_k(A)|}{\sigma_k(A)} > \epsilon_k \right\}$。

定理 4.5. 对于固定的置信水平 $1-\theta$，其中 $\theta \in (0,1)$，由量化引起的最大相对误差界 $\epsilon_k(\theta)$ 与奇异值 $\sigma_k(A)$的大小严格成反比：

$$\epsilon_k(\theta) \propto \frac{1}{\sigma_k(A)}. \tag{44}$$

此外，对于任意两个大小分别为 $\sigma_i(A) > \sigma_j(A) > 0$的奇异值，在相同的置信水平 $1-\theta$下，严格成立：

$$\epsilon_i(\theta) < \epsilon_j(\theta). \tag{45}$$

定理 4.5确立了我们分析的理论核心，揭示了一种根本的谱不对称性：尽管量化为整个谱添加了大致均匀的噪声底，但相对损伤高度不均匀。主导奇异值足够大，足以抑制该噪声，从而产生可忽略的相对扰动。相反，尾部奇异值处于低信噪比区间，因而极易受到失真的影响。

4.4 幂律衰减下的失效概率

在本小节中，我们通过计算显式的失效概率来具体分析相对误差。定义 θ_k 为第 k 个奇异值对应的失效概率。以下定理给出了 θ_k的上界

定理 4.6. 考虑满足假设 *4.1* 和假设 *4.2* 的随机矩阵 $A \in \mathbb{R}^{m \times n}$。设 $R_k \triangleq \frac{|\sigma_k(A) - \sigma_k(\tilde{A})|}{\sigma_k(A)}$ 为相对误差。对于任意给定的相对误差容限 $\eta > 0$，定义 θ_k 为$R_k > \eta$的概率。则 θ_k 满足以下上界：

$$\theta_k \lesssim (m+n) \cdot \exp\left(-\frac{\eta^2 \mu^2}{CDs^2} \cdot \xi_k\right), \tag{46}$$

其中

$$\xi_k = \begin{cases} k^{-\alpha}, & 1 \leq k \leq r \\ r^{-\alpha}, & r < k \leq \min(m,n) \end{cases}, \tag{47}$$

$D \triangleq \max(m,n)$，且 C 是与噪声的次高斯范数相关的通用常数 *e.*

定理 4.6 在渐近谱假设下确立了失效概率的变化规律。回顾一下，量化步长定义为 $s = a_{\max}/L$。根据假设 4.1，我们有 $\sigma_1(A) \sim \mu$。因此，步长满足缩放关系 $s \sim \mu/L$。将这一渐近关系代入公式（46），失效概率上界具有如下形式：

$$\theta_k \lesssim (m+n) \cdot \exp\left(-\frac{\eta^2 L^2}{C \cdot D} \cdot \xi_k\right). \tag{48}$$

考虑满足 $1 \le k \le r$ 的主成分。在此区间中 $\xi_k = k^{-\alpha}$。随着量化分辨率 L增大，该指数趋于负无穷：

$$\lim_{L\to\infty} \left(-\frac{\eta^2 L^2}{C \cdot D} \cdot k^{-\alpha}\right) = -\infty. \tag{49}$$

因此，失效概率趋于零：

$$\lim_{L\to\infty} \theta_k = 0, \tag{50}$$

这证明，只要位宽足够大，主导奇异值就能抵抗量化噪声。考虑满足 $r < k \le \min(m,n)$ 的残余成分。在此区间中 $\xi_k = r^{-\alpha}$。当 $r \to \infty$ 时，该极限成立：

$$\lim_{r\to\infty} (m+n) \cdot \exp\left(-\frac{\eta^2 L^2}{C \cdot D} \cdot r^{-\alpha}\right) = (m+n). \tag{51}$$

该界放宽为平凡极限（$m+n$），这意味着相对误差不再受到抑制。这从数学上形式化了谱平坦化：在尾部奇异值处，量化噪声的幅度与信号底噪相当，使得谱结构无法与噪声区分。

4.5 量化增加StableRank

在定理 4.1 下确立了量化噪声的非均匀影响后，我们现在应用这些发现来解释矩阵稳定秩的行为。

定理 4.7。设 $A \in \mathbb{R}^{m\times n}$ 是一个满足假设 *4.1* 的矩阵， A 是其量化对应矩阵。那么，矩阵 A的稳定秩在量化后严格增加：

$$S_r(\tilde{A}) > S_r(A). \tag{52}$$

5 实验

5.1 模型架构与数据集

为验证分析结果，我们采用遵循 GPT-2-124M 配置的仅解码器 Transformer 架构 [19]。具体而言，模型包含 $L = 12$ 个 Transformer 层，隐藏维度为 $d_{\text{model}} = 768$，并设有 $H = 12$ 个注意力头。上下文窗口大小设为 1024 个词元。我们在 OpenWebText 数据集 [20], 上训练所有模型；该数据集是一个大规模网络文本集合，广泛用于语言模型预训练。所有实验均使用 PyTorch 框架实现，并在配备 8 个 GPU 的计算节点上运行。我们采用 1280 的全局批次大小，通过每个 GPU 的局部批次大小为 4，并结合 40 步梯度累积来实现。模型使用 Ada mW 优化器 [21] 进行训练，权重衰减系数为 $\beta = 0.1$。学习率初始值设为 1.5×10^{-4}，并在前 20 步进行线性预热。此外，为降低梯度爆炸的风险，我们采用梯度裁剪，将最大范数设为 1。0。本文所有 FP4 训练均采用 W4A4G4 量化，其中权重、激活值和梯度均以 E2M1 NVFP4 格式表示。由于英伟达的 FP4 训练软件栈为闭源，目前尚无法进行原生的硬件支持 FP4 训练；因此，我们通过 BF16 仿真开展 NVFP4 实验。更多实验见附录 E 节。

5.2 谱展平的实证验证

我们通过对 GPT-2-124M 中多层感知机（MLP）模块进行逐层分析，实证验证了所预测的谱退化。具体而言，我们评估了 BF16 基线与其 NVFP4 量化对应矩阵的奇异值谱。

图 1 左侧面板通过逐分量失真比 ϵ_k 阐明了这一机制。我们观察到主导奇异值（前 30%，比值 ≈ 0.02–0.05）受到不成比例的衰减，这直接抬高了稳定秩

，原因是谱范数受到抑制。相反，尾部成分则相对得以保留（ $\epsilon_k \to 0.18$）。这种不对称性印证了定理 4.5，表明量化噪声底有效阻碍了头部谱的衰减。图 1 右侧面板证实，量化前嵌入呈现对数线性谱衰减，验证了定理 3.7 中的幂律结果。经过 NVFP4 量化后，我们检测到稳定秩确定性地增加（$2.52 \to 2.71$），这与定理 4.7 预测的谱平坦化一致。

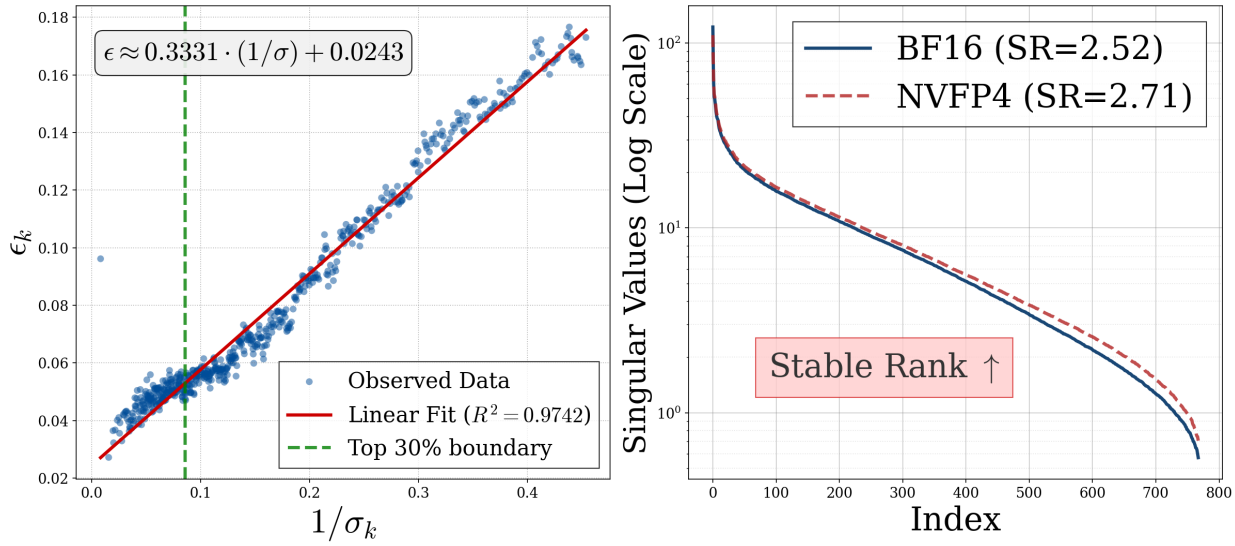


图 1：多层感知机嵌入矩阵的谱分析。（左）量化后奇异值与原始奇异值在各索引处的比值。（右）BF16 和 NVFP4 量化下奇异值的对数分布。

5.3 谱演化与表示坍缩

在定理 4.7 分析的基础上，我们研究谱性质随时间的演变。图 2 对比了多层感知机层中 BF16 和 NVFP4 嵌入矩阵在若干代表性训练步数下的奇异值谱。

实证结果揭示了谱动力学的分岔，表明低精度训练存在两个截然不同的阶段。在初始迭代阶段（步骤 0–10），观测结果印证了我们的理论预测。如图 2 上排所示，与 BF16 基线相比，量化嵌入（NVFP4）的尾谱更高。因此，量化表征的稳定秩随之增加。这证实了量化会引入噪声底，掩盖细粒度谱结构，从而使分布整体趋于平坦。

然而，随着训练推进（第 50–100 步），出现了病态发散。图 2 的底行显示，量化后嵌入的稳定秩迅速下降，偏离了 BF16 基线中观察到的单调上升趋势。到第 100 步时，BF16 模型达到了较高的稳定秩（SR=1.08），表明其有效维数得到扩展，特征编码能力有所增强。相比之下，NVFP4 嵌入退化为低秩状态（SR=1.00）。这一转变表明，累积的梯度误差使系统从静态扁平化区间转向动态表征坍缩。该过程将模型容量压缩到少数主导奇异值上，从而实际上消除了语义编码所必需的重尾谱结构。

6 结论与未来工作

本研究结合齐夫统计与随机矩阵理论，表明均匀量化会破坏大语言模型的重尾谱结构，导致谱平坦化和稳定秩严格上升。我们的实证分析显示，这种失真会引发表示坍缩相变。这一发现挑战了非结构化扰动模型：量化应被视为优化景观的几何退化，而非简单的加性噪声。

未来的研究应着重于设计以幂律统计为基础的非均匀量化方案，以保留尾部保真度，同时采用谱正则化来防止秩坍缩。此外，将这一谱分析扩展到 Softmax 和注意力机制等非线性函数，对于全面表征量化效应至关重要。这些方向共同为稳健的低精度训练奠定了理论基础。

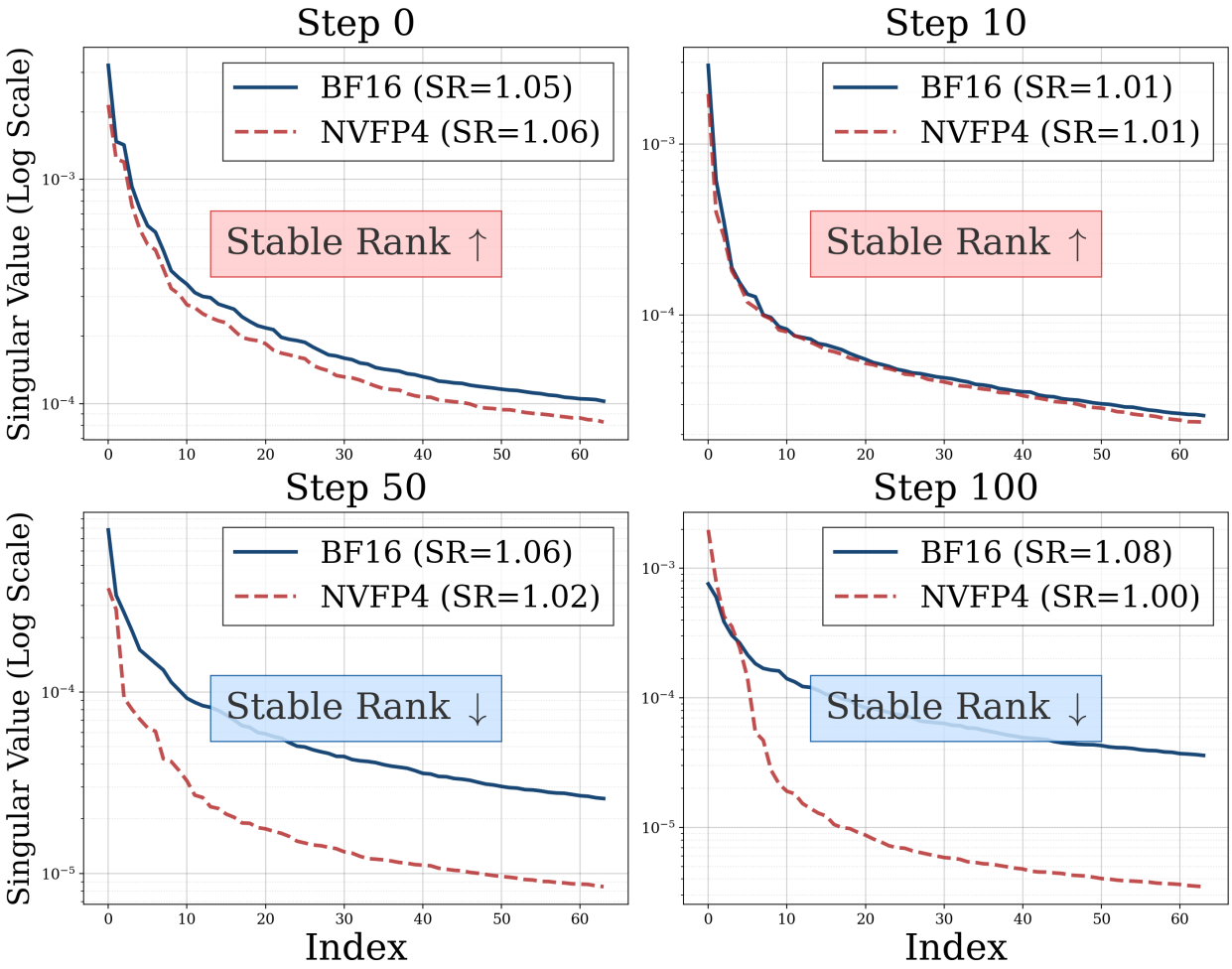


图 2：训练的谱演化。上：初始噪声使稳定秩（SR）增大。下：后续步骤显示表示坍缩。

参考文献

[1] Paulius Micikevicius、Dusan Stosic、Neil Burgess、Marius Cornea、Pradeep Dubey、Richard Grisenthwaite、Sangwon Ha、Alexander Heinecke、Patrick Judd、John Kamalu、Naveen Mellempudi、Stuart Oberman、Mohammad Shoeybi、Michael Siu 和 Hao Wu。《用于深度学习的 FP8 格式》，2022年。[2] Houwen Peng、Kan Wu、Yixuan Wei、Guoshuai Zhao、Yuxiang Yang、Ze Liu、Yifan Xiong、Ziyue Yang、Bolin Ni、Jingcheng Hu、Ruihang Li、Miaosen Zhang、Chen Li、Jia Ning、Ruizhe Wang、Zheng Zhang、Shuguang Liu、Joe Chau、Han Hu 和 Peng Cheng。《FP8-LM：训练 FP8 大语言模型》，2023年。[3] Sergio P. Perez、Yan Zhang、James Briggs、Charlie Blake、Josh Levy-Kramer、Paul Balanca、Carlo Luschi、Stephen Barlow 和 Andrew William Fitzgibbon。《使用8位浮点数训练和推理大语言模型》，2023年。[4] 英伟达公司。《NVIDIA Blackwell架构技术概述》。技术报告，英伟达公司，2024年。访问日期：2024-03-18。[5] 英伟达。《推出 NVFP4，实现高效、准确的低精度推理》。https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/，2025年6月。英伟达技术博客。[6] Armen Aghajanyan、Sonal Gupta 和 Luke Zettlemoyer。《内在维度解释了语言模型微调的有效性》。载于 Chengqing Zong、Fei Xia、Wenjie Li 和 Roberto Navigli 编，《第59届计算语言学协会年会暨第11届国际联合》论文集

自然语言处理会议（第*1*卷：长篇论文），第7319–7328页，线上，2021年8月。计算语言学协会。[7] Charles H. Martin 和 Michael W. Mahoney。深度神经网络中的隐式自正则化：来自随机矩阵理论的证据及其对学习的启示。机器学习研究期刊，22(165):1–73，2021年。[8] Yuen Ren Chao 和 George Kingsley Zipf。《人类行为与最省力原则：人类生态学导论》。《语言》，26:394，1950年。[9] Bita Darvish Rouhani、Ritchie Zhao、Ankit More、Mathew Hall、Alireza Khodamoradi、Summer Deng、Dhruv Choudhary、Marius Cornea、Eric Dellinger、Kristof Denolf、Stosic Dusan、Venmugil Elango、Maximilian Golub、Alexander Heinecke、Phil James-Roxby、Dharmesh Jani、Gaurav Kolhe、Martin Langhammer、Ada Li、Levi Melnick、Maral Mesmakhosroshahi、Andres Rodriguez、Michael Schulte、Rasoul Shafipour、Lei Shao、Michael Siu、Pradeep Dubey、Paulius Micikevicius、Maxim Naumov、Colin Verrilli、Ralph Wittig、Doug Burger 和 Eric Chung。用于深度学习的微缩放数据格式，2023年。[10] 英伟达、Felix Abecassis、Anjulie Agrusa、Dong Ahn、Jonah Alben、Stefania Alborghetti、Michael Andersch、Sivakumar Arayandi、Alexis Bjorlin、Aaron Blakeman、Evan Briones、Ian Buck、Bryan Catanzaro、Jinhang Choi、Mike Chrzanowski、Eric Chung、Victor Cui、Steve Dai、Bita Darvish Rouhani、Carlo del Mundo、Deena Donia、Burc Eryilmaz、Henry Estela、Abhinav Goel、Oleg Goncharov、Yugi Guvvala、Robert Hesse、Russell Hewett、Herbert Hum、Ujval Kapasi、Brucek Khailany、Mikail Khona、Nick Knight、Alex Kondratenko、Ronny Krashinsky、Ben Lanir、Simon Layton、Michael Lightstone、Daniel Lo、Paulius Micikevicius、Asit Mishra、Tim Moon、Deepak Narayanan、Chao Ni、Abhijit Paithankar、Satish Pasumarthi、Ankit Patel、Mostofa Patwary、Ashwin Poojary、Gargi Prasad、Sweta Priyadarshi、Yigong Qin、Xiaowei Ren、Oleg Rybakov、Charbel Sakr、Sanjeev Satheesh、Stas Sergienko、Pasha Shamis、Kirthi Shankar、Nishant Sharma、Mohammad Shoeybi、Michael Siu、Misha Smelyanskiy、Darko Stosic、Dusan Stosic、Bor-Yiing Su、Frank Sun、Nima Tajbakhsh、Shelby Thomas、Przemek Tredak、Evgeny Tsykunov、Gandhi Vaithilingam、Aditya Vavre、Rangharajan Venkatesan、Roger Waleffe、Qiyu Wan、Hexin Wang、Mengdi Wang、Lizzie Wei、Hao Wu、Evan Wu、Keith Wyss、Ning Xu、Jinze Xue、Charlene Yang、Yujia Zhai、Ruoxi Zhang、Jingyang Zhu 和 Zhongbo Zhu。使用 NVFP4 预训练大语言模型，2025年。[11] 维基媒体基金会。维基百科转储数据。https://dumps。wikimedia。org，2024年。访问日期：2025年1月1日。[12] Common Crawl 基金会。Common Crawl。https://commoncrawl。org，2024年。访问日期：2025年1月1日。[13] William B. Johnson 和 Joram Lindenstrauss。希尔伯特空间中的 Lipschitz 映射扩展。《当代数学》，26:189–206，1984年。[14] Zhidong Bai 和 Jack W Silverstein。《大维随机矩阵的谱分析》。施普林格统计学丛书。施普林格，纽约州纽约市，第2版，2010年。[15] Rajendra Bhatia。矩阵分析，数学研究生教材第169卷。施普林格出版社，纽约，1997年。[16] Joel A. Tropp。随机矩阵和的易用尾界。《*Found. Comput. Math.*》，12(4):389–434，2012年8月。[17] Jinho Baik、Gérard Ben Arous 和 Sandrine Péché。非零复样本协方差矩阵最大特征值的相变。《概率年刊》，33:1643–1697，2004年。[18] V A Marcenko 和 Leonid A. Pastur。若干随机矩阵集合的特征值分布。《苏联数学汇刊》，1:457–483，1967年。[19] Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督的多任务学习者。2019年。[20] Aaron Gokaslan 和 Vanya Cohen。OpenWebText 语料库。http://Skylion007。github。io/OpenWebTextCo rpus，2019年。[21] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于国际学习表征会议，2017年。[22] Lukasz Dkebowski。从 Zipf 定律到神经网络缩放：经由 Heaps 定律和 Hilberg 假说。2025年。[23] Kawin Ethayarajh。上下文化词表示的上下文性有多强？比较 BERT、ELMo 和 GPT-2 嵌入的几何特性。载于 Kentaro Inui、Jing Jiang、Vincent Ng 和 Xiaojun Wan 编，*2019*年自然语言处理实证方法会议与第*9*届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第55–65页，中国香港，2019年11月。计算语言学协会。[24] Jun Gao、Di He、Xu Tan、Tao Qin、Liwei Wang 和 Tie-Yan Liu。训练自然语言中的表示退化问题

guage g生成模型，2019年。

[25] Sanjeev Arora、Yuanzhi Li、Yingyu Liang、Tengyu Ma 和 Andrej Risteski。基于PMI的词嵌入的潜变量模型方法。计算语言学协会会刊，4:385–399，2016年。[26] Woojin Chung、Jiwoo Hong、Na Min An、James Thorne 和 Se-Young Yun。通过降低嵌入变异性实现稳定的语言模型预训练。载于 Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen 编，*2024*年自然语言处理实证方法会议论文集，第10852–10863页，迈阿密，佛罗里达州，美国，2024年11月。计算语言学协会。[27] Colin Wei、Sang Michael Xie 和 Tengyu Ma。为什么预训练语言模型有助于下游任务？对头部微调和提示微调的分析。载于第*35*届神经信息处理系统国际会议论文集，NIPS '21，美国纽约州雷德胡克，2021年。Curran Associates Inc. [28] Jiaqi Mu、Suma Bhat 和 Pramod Viswanath。去除顶部成分法：一种简单有效的词表示后处理方法，2018年。[29] Bohan Li、Hao Zhou、Junxian He、Mingxuan Wang、Yiming Yang 和 Lei Li。关于预训练语言模型生成的句子嵌入，2020年。[30] Jianlin Su、Jiarun Cao、Weijie Liu 和 Yangyiwen Ou。通过白化句子表示提升语义表现并加快检索，2021年。[31] Sho Yokoi、Han Bao、Hiroto Kurita 和 Hidetoshi Shimodaira。齐夫白化。载于 A. Globerson、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. Tomczak 和 C. Zhang 编，神经信息处理系统进展，第37卷，第122259–122291页。Curran Associates, Inc.，2024年。[32] Armen Aghajanyan、Sonal Gupta 和 Luke Zettlemoyer。内在维度解释了语言模型微调的有效性。载于 Chengqing Zong、Fei Xia、Wenjie Li 和 Roberto Navigli 编，第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议（第*1*卷：长篇论文）论文集，第7319–7328页，线上，2021年8月。计算语言学协会。[33] Nandan Kumar Jha 和 Brandon Reagen。语言模型中的谱缩放定律：前馈网络对其潜在空间的利用效率如何？，2025年。[34] Levent Sagun、Utku Evci、V. Ugur Guney、Yann Dauphin 和 Leon Bottou。过参数化神经网络海森矩阵的实证分析，2018年。[35] Behrooz Ghorbani、Shankar Krishnan 和 Ying Xiao。通过海森矩阵特征值密度探究神经网络优化，2019年。[36] Sepp Hochreiter 和 Jürgen Schmidhuber。平坦极小值。神经计算，9(1):1–42，1997年1月。[37] Yann LeCun、John Denker 和 Sara Solla。最优脑损伤。载于 D. Touretzky 编，神经信息处理系统进展，第2卷。Morgan-Kaufmann，1989年。[38] Zhen Dong、Zhewei Yao、Amir Gholami、Michael W. Mahoney 和 Kurt Keutzer。HAWQ：面向混合精度神经网络的海森感知量化。*2019*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第293–302页，2019年。[39] Jingyuan Liu、Jianlin Su、Xingcheng Yao、Zhejun Jiang、Guokun Lai、Yulun Du、Yidao Qin、Weixin Xu、Enzhe Lu、Junjie Yan、Yanru Chen、Huabin Zheng、Yibo Liu、Shaowei Liu、Bohong Yin、Weiran He、Han Zhu、Yuzhi Wang、Jianzhou Wang、Mengnan Dong、Zheng Zhang、Yongsheng Kang、Hao Zhang、Xinran Xu、Yutao Zhang、Yuxin Wu、Xinyu Zhou 和 Zhilin Yang。Muon可扩展用于大语言模型训练，2025年。[40] Jun Gao、Di He、Xu Tan、Tao Qin、Liwei Wang 和 Tie-Yan Liu。自然语言生成模型训练中的表示退化问题。载于国际学习表征会议（*ICLR*），2019年。[41] Yelysei Bondarenko、Markus Nagel 和 Tijmen Blankevoort。理解并克服高效Transformer量化的挑战。载于 Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih 编，*2021*年自然语言处理实证方法会议论文集，第7947–7969页，线上及多米尼加共和国蓬塔卡纳，2021年11月。计算语言学协会。[42] Tim Dettmers、Mike Lewis、Younes Belkada 和 Luke Zettlemoyer。LLM.int8()：大规模Transformer的8位矩阵乘法，2022年。[43] Mingjie Sun、Xinlei Chen、J. Zico Kolter 和 Zhuang Liu。大语言模型中的大规模激活，2024年。[44] Bernard Widrow 和 István Kollár。量化噪声：数字计算、信号处理、控制和通信中的舍入误差。2008年6月。

[45] Xiuying Wei、Yunchen Zhang、Xiangguo Zhang、Ruihao Gong、Shanghang Zhang、Qi Zhang、Fengwei Yu 和 Xianglong Liu。《离群值抑制：突破低比特变换器语言模型的极限》。载于 S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh 编，神经信息处理系统进展，第35卷，第17402–17414页。Curran Associates, Inc.，2022年。[46] Guangxuan Xiao、Ji Lin、Mickael Seznec、Hao Wu、Julien Demouth 和 Song Han。《SmoothQuant：大语言模型的精准高效训练后量化》。载于 Andreas Krause、Emma Brunskill、Kyunghyun Cho、Barbara Engelhardt、Sivan Sabato 和 Jonathan Scarlett 编，第*40*届国际机器学习会议论文集，机器学习研究论文集第202卷，第38087–38099页。PMLR，2023年7月23日至29日。[47] Ji Lin、Jiaming Tang、Haotian Tang、Shang Yang、Wei-Ming Chen、Wei-Chen Wang、Guangxuan Xiao、Xingyu Dang、Chuang Gan 和 Song Han。《AWQ：面向大语言模型压缩与加速的激活感知权重量化》，2024年。[48] Jerry Chee、Yaohui Cai、Volodymyr Kuleshov 和 Christopher M De Sa。《QuIP：具有保证的大语言模型2比特量化》。载于 A. Oh、T. Naumann、A. Globerson、K. Saenko、M. Hardt 和 S. Levine 编，神经信息处理系统进展，第36卷，第4396–4429页。Curran Associates, Inc.，2023年。[49] Albert Tseng、Jerry Chee、Qingyao Sun、Volodymyr Kuleshov 和 Christopher De Sa。《QuIP#：借助Hadamard非相干性和格码本实现更优的大语言模型量化》。载于 Ruslan Salakhutdinov、Zico Kolter、Katherine Heller、Adrian Weller、Nuria Oliver、Jonathan Scarlett 和 Felix Berkenkamp 编，第*41*届国际机器学习会议论文集，机器学习研究论文集第235卷，第48630–48656页。PMLR，2024年7月21日至27日。[50] William B. Johnson 和 Joram Lindenstrauss。《到希尔伯特空间中的Lipschitz映射的扩张》。《当代数学》，26：189–206，1984年。[51] Saleh Ashkboos、Amirkeivan Mohtashami、Maximilian L. Croci、Bo Li、Pashmina Cameron、Martin Jaggi、Dan Alistarh、Torsten Hoefler 和 James Hensman。《QuaRot：旋转大语言模型的无离群值4比特推理》。载于 A. Globerson、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. Tomczak 和 C. Zhang 编，神经信息处理系统进展，第37卷，第100213–100240页。Curran Associates, Inc.，2024年。[52] Zhenda Xie、Yixuan Wei、Huanqi Cao、Chenggang Zhao、Chengqi Deng、Jiashi Li、Damai Dai、Huazuo Gao、Jiang Chang、Kuai Yu、Liang Zhao、Shangyan Zhou、Zhean Xu、Zhengyan Zhang、Wangding Zeng、Shengding Hu、Yuqing Wang、Jingyang Yuan、Lean Wang 和 Wenfeng Liang。《mHC：流形约束超连接》，2025年。[53] 英伟达。《NVIDIA Blackwell架构技术简报》。https://resources。nvidia。com/en-us-blackwell-architecture，2025年。[54] Augustin-Louis Cauchy。《style id='31'>《巴黎综合理工学院分析课程》。剑桥图书馆藏书系列——数学。剑桥大学出版社，2009年。[55] Edward Charles Titchmarsh。《style id='34'>《黎曼ζ函数理论》。牛津大学出版社，牛津，第2版，1986年。

A 相关研究

大语言模型（LLM）超低比特量化面临的挑战，在于极端压缩条件下如何保留谱结构。本节将回顾这一挑战的理论基础，追溯从自然语言的统计物理到神经网络谱性质的发展脉络，并进一步分析旨在缓解谱退化的量化算法的演进。

A.1 统计与谱基础

大语言模型的谱属性并非随机初始化造成的单纯伪象，而是由训练数据的统计分布内在塑造的。正如 [8], 所确立的，自然语言中的词频遵循严格的幂律分布 $f \propto r^{-\alpha}$。这一分布已在包括 Common Crawl [12] 和 Wikipedia [11] 在内的网络规模语料库中得到广泛验证。 [22] 最近的理论分析表明，观测到的神经缩放定律是这些齐夫统计的演绎结果。
在神经表征的语境中，这种重尾数据分布会在嵌入空间中引发几何各向异性。 [23] 证明，上下文表征呈现出狭窄锥体拓扑，而 [24] 则将这一现象形式化为表示退化。尽管早期框架曾假设这一生成过程主导几何结构 [25], , 但 [26] 最近的分析通过实证证实，与稀有词元相比，高频词元表现出不同的变化模式，由此确立了词元频率决定嵌入流形的局部曲率。这一现象往往会受到 softmax 层固有频率偏差的进一步加剧 [27]。

从时间脉络来看，缓解这种坍塌的策略从启发式减法 [28] 逐渐发展到密度感知校准，例如归一化流 [29] 和标准白化 [30]。最近， [31] 提出了齐夫白化，通过明确利用词频统计来展平谱，推进了这一方法。[32] 利用由此产生的低内在维度，证明模型可以在低秩子空间中进行微调；[33] 则认为，脆弱的长尾——常被视为噪声而弃之不顾——通过尾部优先增长机制编码了关键的语义区分。对重尾谱的这一关注，与随机矩阵理论（RMT）的严格刻画相吻合。 [7] 将重尾自正则化（HT-SR）确认为高性能深度网络的标志。关于 Hessian 谱的平行研究 [34, 35] 证实，损失景观由庞大的主体和少数极端离群值构成，而后者通常与平坦极小值相关 [36]。 [37]关于最优脑损伤（OBD）的早期研究利用这些性质进行剪枝。随后， [38] 将这一思路用于海森感知量化，证明谱范数较大的层对噪声格外敏感。最近，Muon 优化器 [39]在训练过程中明确管理这些谱性质，进一步凸显了为实现可扩展学习而维持特定谱特征的必要性。表 1 总结了统计与谱基础方面的文献。

A.2 量化病理与算法演进

这些理论性质在实践中表现为离群值现象 [41], , 它构成了低比特量化的主要障碍。 [42]观察到，当参数量超过 6.7B 时，会出现系统性大规模激活值[43]。根据经典量化理论 [44], , 这些离群值迫使均匀量化器采用过大的步长，导致分布主体中的细粒度信息崩塌。 [45] 进一步指出，层归一化是放大这些方差的关键因素。

为应对离群值带来的谱挑战，量化策略已从简单分解演进为复杂的几何变换。早期的缓解策略，例如 LLM.int8() [42],, 采用混合精度分解来分离离群值。后续采用数学等价变换的方法，例如 SmoothQuant[46] 和 AWQ [47],, 旨在将量化难度从激活值转移到权重上。

近期研究已转向几何变换，以从根本上改变谱分布。QuIP [48]和 QuIP# [49]等方法引入随机哈达玛变换，对分布进行谱白化，并借助 Johnson-Lindenstrauss 引理 [50]将其近似为高斯分布。 [51] 将这一方法扩展至 QuaRot 中的端到端推理。最近， [52] 提出通过流形约束限制信号传播谱范数。

与此同时，硬件感知格式也不断演进，以支持这些算法进展。早期的探索着眼于 FP8格式 [1], , 而较新的标准（如 NVFP4 [53] 和 MXFP4 [9] ）采用分块缩放来进行局部谱归一化，从而在低精度区间保留表征能力。尽管工程技术取得了这些进展，但量化模型推理能力崩溃与齐夫式谱尾脆弱性之间的具体理论机制仍未得到解释。表 2 总结了有关量化病理与算法演进的文献。

表 1：统计与谱基础相关文献综述。

类别	主要贡献与发现	引用文献
齐夫统计	在自然界中确立幂律分布（$f \propto r^{-\alpha}$）语言，并在网络规模语料库上验证；将齐夫定律与神经缩放定律联系起来。	[8, 12, 11, 22]
各向异性与几何	指出表示退化和窄锥效应；将词元频率与局部流形曲率联系起来以及softmax偏差.	[23, 24, 40, 25, 26, 27]
各向同性恢复	从启发式减法（去除顶部成分法）演进到归一化流。	[28, 29, 30, 31]
维数与长尾	关于内在维度与以下因素重要性的争论：脆弱的长尾对于语义区分和容量的重要性增长。	[32, 33]
谱理论（随机矩阵理论）	刻画了以下方面的重尾自正则化（重尾自正则化）训练权重；分析经验谱密度。	[7]
海森矩阵与优化	分析海森矩阵谱（主体与离群值、平坦极小值-ima）；最优脑损伤（OBD）；海森感知量化（HAWQ）；通过 Muon 优化器进行谱控制器 .	[34, 35, 36, 37, 38, 39]

表 2：量化病理与算法演进相关文献综述。

类别	主要贡献与机制	引用
离群值病理	识别离群值现象和大规模激活值,大语言模型（>6.7B）；分析层归一化的作用及其在方差放大中的作用。	[41, 42, 43, 45]
经典理论	假设服从均匀分布的基础量化噪声理论（这一假设常被大语言模型中的离群值打破）。	[44]
分解与迁移	**分解**：混合精度（INT8/FP16）隔离。**迁移**：以数学方式将量化难度从激活值转移到权重（SmoothQuant、AWQ）。	[42, 46, 47]
旋转与白化	使用随机哈达玛变换（RHT）和 JL 引理对白化谱分布并分散离群值（QuIP、QuIP#、QuaRot）。	[48, 49, 50, 51]
流形约束	通过以下方式明确限制信号传播谱范数：流形投影（mHC）。	[52]
硬件格式	从 FP8 到分块缩放格式（NVFP4、MXFP4）的演进，旨在实现局部谱归一化。	[1, 53, 9]

B 补充定义

为确保本文内容自成一体，我们对谱分析中使用的数学概念和理论框架给出正式定义。

定义 B.1 (稳定秩)。 对于随机矩阵 $A \in \mathbb{R}^{m\times n}$，其稳定秩记为 $S_r(A)$，定义为其弗罗贝尼乌斯范数的平方与谱范数的平方之比：

$$S_r(A) \triangleq \frac{\|A\|_F^2}{\|A\|_2^2}, \tag{53}$$

其中， $\|A\|_F = \sqrt{\sum_{i,\ j} A_{ij}^2}$ 表示弗罗贝尼乌斯范数， $\|A\|_2 = \sigma_{\max}(A)$表示谱范数。

定义 B.2 (斯蒂尔切斯变换)。 概率测度 μ 的斯蒂尔切斯变换定义在 $z \in \mathbb{C} \setminus \mathit{supp}(\mu)$ 上，如下：

$$m_\mu(z) = \int_{\mathbb{R}} \frac{1}{\lambda - z} d\mu(\lambda). \tag{54}$$

对于具有 n 个离散特征值 $\{\lambda_i\}_{i=1}^n$ 的矩阵，该变换的经验形式为：

$$m_n(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i - z}. \tag{55}$$

斯蒂尔切斯变换用于证明引理 3.3 和引理 3.4 中噪声模型的谱收敛 .

定义 B.3 （经验谱分布，ESD）。 对于一个特征值为 λ_1、.....、 λ_d 的对称矩阵 $M \in \mathbb{R}^{d\times d}$，经验谱分布定义如下：

$$\mu_M = \frac{1}{d}\sum_{i=1}^{d} \delta_{\lambda_i} \tag{56}$$

其中 δ 是狄拉克δ测度。当 $d \to \infty$ 时，经验谱分布收敛于一个确定性测度。

经验谱分布用于刻画特征值的全局密度，是证明重尾量化噪声在引理 3.4 和定理 3.5 中收敛于确定性噪声主体的基础。

C 数学证明语言模型嵌入的低秩性质

引理 3.1 的证明。 我们将公式（6）写成矩阵形式。令 $U = [v_1, v_2, \ldots, v_V] \in \mathbb{R}^{d\times V}$ 为特征向量矩阵，令 $\Lambda_p = \mathrm{diag}(p_1, \ldots, p_V)$ 为概率对角矩阵。于是，

$$\Sigma = U\Lambda_p U^T. \tag{57}$$

根据假设 2.3，概率权重满足 $p_1 > p_2 > \cdots > p_V > 0$，且 $p_k \sim k^{-\alpha}$，这意味着 p_k 具有较大的谱间隙。根据线性代数中奇异值分解的对偶性质，Σ 的非零特征值与矩阵 $M = \Lambda_p^{1/2} U^T U \Lambda_p^{1/2}$ 的非零特征值完全相同。我们分析矩阵 $M \in \mathbb{R}^{V\times V}$。令 $G = U^T U$ 为特征向量的格拉姆矩阵 [15]。由于向量归一化，

$$(G)_{ii} = \|v_i\|^2 = 1. \tag{58}$$

根据假设 2.4，$(G)_{ij} = \langle v_i,\ v_j\rangle$ $(i \neq j)$，且 $|(G)_{ij}| = \mathcal{O}(1/\sqrt{d})$。$G$ 可写为 ：

$$G = I_V + E, \tag{59}$$

其中， I_V是单位矩阵， E 是完全干扰矩阵，其对角元素等于 0，非对角元素的绝对值等于 $\mathcal{O}(1/\sqrt{d})$，即：

$$E = \begin{cases} |E_{ij}| = 0, & i = j \\ |E_{ij}| = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right), & i \neq j \end{cases}. \tag{60}$$

因此，目标矩阵 M 变为：

$$M = \Lambda_p^{1/2}(I_V + E)\Lambda_p^{1/2} = \Lambda_p + \Lambda_p^{1/2} E \Lambda_p^{1/2}. \tag{61}$$

Let:

$$A = \Lambda_p = \text{diag}(p_1, \ldots, p_V), \quad \Delta = \Lambda_p^{1/2} E \Lambda_p^{1/2}. \tag{62}$$

由于 A 是对角矩阵， A 的特征值为 p_k。根据定理 2.6， τ_k 与 p_k 之差的上界为扰动矩阵 Δ的 2-范数：

$$|\tau_k - p_k| \leq \|\Delta\|_2. \tag{63}$$

我们需要计算 $\|\Delta\|_2$的上界，其中

$$\Delta = \begin{cases} (\Delta)_{ij} = \sqrt{p_i p_j} \cdot \langle v_i, v_j \rangle, & i \neq j \\ (\Delta)_{ij} = 0, & i = j \end{cases}. \tag{64}$$

利用矩阵范数不等式：

$$\|\Delta\|_2 \leq \|\Delta\|_F = \sqrt{\sum_{i \neq j} p_i p_j \langle v_i, v_j \rangle^2}, \tag{65}$$

其中 $||\cdot||_F$是弗罗贝尼乌斯范数。根据定理 2.4， $|\langle v_i, \ v_j \rangle| = \mathcal{O}(1/\sqrt{d})$。 然后，

$$\|\Delta\|_2 \leq \mathcal{O}\left(\frac{1}{\sqrt{d}}\right) \cdot \sqrt{\sum_{i \neq j} p_i p_j} < \mathcal{O}\left(\frac{1}{\sqrt{d}}\right) \cdot \sum_{i=1}^{V} p_i. \tag{66}$$

由于概率归一化 $\sum p_i = 1$：

$$\|\Delta\|_2 \leq \mathcal{O}\left(\frac{1}{\sqrt{d}}\right) \cdot 1 = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right). \tag{67}$$

根据方程 (63)和方程 (67)，

$$|\tau_k - p_k| \leq \mathcal{O}\left(\frac{1}{\sqrt{d}}\right). \tag{68}$$

注意， $d \to \infty$ 会导致 $\mathcal{O}\left(1/\sqrt{d}\right) \to 0$。因此，对于固定的 p_k：

$$\frac{|\tau_k - p_k|}{p_k} \to 0 \quad (d \to \infty). \tag{69}$$

根据定义 2.1，渐近等价关系成立：

$$\tau_k \sim p_k. \tag{70}$$

根据假设 2.3，以下关系成立：

$$\tau_k \sim p_k \sim k^{-\alpha}. \tag{71}$$

□

引理 3.2 的证明。 根据柯西积分判别法，对于单调递减函数 $f(x) = x^{-\alpha}$，存在如下不等式：

$$\sum_{j=r+1}^{d} j^{-\alpha} < \int_r^d x^{-\alpha} dx. \tag{72}$$

计算定积分：

$$\int_r^d x^{-\alpha} dx = \left[\frac{x^{1-\alpha}}{1-\alpha}\right]_r^d = \frac{1}{\alpha - 1}\left(r^{1-\alpha} - d^{1-\alpha}\right). \tag{73}$$

由于 $d > r$ 且 $\alpha > 1$, $d^{1-\alpha} > 0$。为进一步缩放，舍弃减法项 $-d^{1-\alpha}$：

$$\int_r^d x^{-\alpha} dx < \frac{r^{1-\alpha}}{\alpha - 1}. \tag{74}$$

将公式（74）代入公式（16）可得：

$$\nu^2(d) < \frac{L}{d-r} \cdot \frac{r^{1-\alpha}}{\alpha - 1}. \tag{75}$$

对于固定的头部秩 $k \in [1, r]$，定义信噪比函数 $R(d)$ 如下：

$$R(d) \triangleq \frac{\tau_k}{\nu^2(d)}. \tag{76}$$

根据公式（75）：

$$R(d) = \frac{L \cdot k^{-\alpha}}{\nu^2(d)} > \frac{L \cdot k^{-\alpha}}{\left(\frac{L}{d-r} \cdot \frac{r^{1-\alpha}}{\alpha-1}\right)} = \frac{k^{-\alpha}(\alpha-1)}{r^{1-\alpha}} \cdot (d-r). \tag{77}$$

由于 r 和 $k^{-\alpha}(\alpha-1)\, r^{1-\alpha}$ 是正实数，因此以下极限成立：

$$\lim_{d\to\infty} \frac{k^{-\alpha}(\alpha-1)}{r^{1-\alpha}} \cdot (d-r) = +\infty. \tag{78}$$

根据比较判别法，由公式（77）可得

$$\lim_{d\to\infty} R(d) = +\infty. \tag{79}$$

由于 $\lim_{d\to\infty} R(d) = \lim_{d\to\infty} \frac{\tau_k}{\nu^2(d)} = +\infty$，对于任意给定的常数 $1 + \sqrt{c}$，存在一个维数 d_1，使得对于所有 $d > d_1$，以下关系成立：

$$\frac{\tau_k}{\nu^2(d)} > 1 + \sqrt{c}. \tag{80}$$

即：

$$\tau_k > \nu^2(d) \cdot (1 + \sqrt{c}). \tag{81}$$

另一方面，当 $k \in (r, d]$时，$\lim_{d\to\infty} \tau_k = 0$。对于任意给定的常数 $1 + \sqrt{c}$，存在一个维数 d_2，使得对所有 $d > d_2$，下列关系成立：

$$\tau_k \leq \nu^2(d) \cdot (1 + \sqrt{c}). \tag{82}$$

取 $d = \max\{d_1, d_2\}$，则公式（81）与公式（82）同时成立。此外，回顾 $\nu^2(d)$的定义：

$$\nu^2(d) = \frac{L}{d-r} \sum_{j=r+1}^{d} j^{-\alpha}. \tag{83}$$

当 $d \to \infty$时，$\sum_{j=r+1}^{d} j^{-\alpha}$ 可视为黎曼ζ函数 $\zeta(\alpha)$ [55]的尾部：

$$\zeta(\alpha) = \sum_{j=1}^{\infty} j^{-\alpha}. \tag{84}$$

由于 $\zeta(\alpha)$收敛，故 $\sum_{j=r+1}^{d} j^{-\alpha}$ 收敛。由于 L, r 为常数，存在常数 $C > 0, C, d_0 \in \mathbb{R}$，使得对所有 $d > d_0$，下列条件成立

$$|\nu^2(d)| = \left| \frac{L}{d-r} \sum_{j=r+1}^{d} j^{-\alpha} \right| \leq C \cdot \left| \frac{1}{d} \right|. \tag{85}$$

根据定义 2.2，即 $\nu^2(d) = \mathcal{O}(d^{-1})$。 □

引理 3.3 的证明。对于任意 a和 z $(a \neq z)$，以下恒等式成立：

$$\frac{1}{a-z} = -\frac{1}{z} - \frac{a}{z^2} + \frac{a^2}{z^2(a-z)}. \tag{86}$$

将 $a = \tau_j$ 代入公式（86）并展开 m_{Noise}（z），得到：

$$\begin{aligned} m_{Noise}(z) &= \frac{1}{d-r}\sum_{j=r+1}^{d}\left(-\frac{1}{z} - \frac{\tau_j}{z^2} + \frac{\tau_j^2}{z^2(\tau_j - z)}\right) \\ &= -\frac{1}{z} - \frac{1}{z^2}\left(\frac{1}{d-r}\sum_{j=r+1}^{d}\tau_j\right) + \frac{1}{d-r}\sum_{j=r+1}^{d}\frac{\tau_j^2}{z^2(\tau_j - z)}, \\ &= -\frac{1}{z} - \frac{\nu^2(d)}{z^2} + R_{Noise}(z) \end{aligned} \tag{87}$$

其中

$$R_{Noise}(z) \triangleq \frac{1}{d-r}\sum_{j=r+1}^{d}\frac{\tau_j^2}{z^2(\tau_j - z)}. \tag{88}$$

将 $a = \nu^2$（d）代入公式（86）并展开 m_{White}（z），得到：

$$m_{White}(z) = -\frac{1}{z} - \frac{\nu^2(d)}{z^2} + \frac{(\nu^2(d))^2}{z^2(\nu^2(d) - z)} \quad . \tag{89}$$

记

$$R_{White}(z) \triangleq \frac{(\nu^2(d))^2}{z^2(\nu^2(d) - z)}. \tag{90}$$

现在考察差值 Δ（z） $\triangleq |m_{Noise}$（z） $- m_{White}$（z）$|$。公式（87）和公式（89）的前两项相同，因此根据三角不等式：

$$\Delta(z) = |R_{Noise}(z) - R_{White}(z)| \leq |R_{Noise}(z)| + |R_{White}(z)|. \tag{91}$$

对于固定的 z，由于 $\tau_j \in \{\tau_j = Lj^{-\alpha}\}_{j=r+1}^{d}$,

$$\frac{\tau_j^2}{z^2(\tau_j - z)} = \mathcal{O}(\tau_j). \tag{92}$$

因此,

$$|R_{Noise}| = \mathcal{O}\left(\frac{1}{d-r}\sum_{j=r+1}^{d}\tau_j\right) = \mathcal{O}\left(\frac{L}{d-r}\sum_{j=r+1}^{d}j^{-\alpha}\right). \tag{93}$$

当 $d \to \infty$ 位于公式（93）中时，$\sum_{j=r+1}^{d} j^{-\alpha}$ 可视为黎曼ζ函数 $\zeta(\alpha)$ 的尾项。由于 $\zeta(\alpha)$ 收敛，$\sum_{j=r+1}^{d} j^{-\alpha}$ 也收敛。因此,

$$\frac{L}{d-r}\sum_{j=r+1}^{d}j^{-\alpha} = \mathcal{O}(d^{-1}) \tag{94}$$

方程（87）和方程（94）给出：

$$|R_{Noise}| = \mathcal{O}(\mathcal{O}(d^{-1})) = \mathcal{O}(d^{-1}). \tag{95}$$

同理,

$$|R_{White}| = \mathcal{O}(\nu^2(d)). \tag{96}$$

引理 3.2 表明 $\nu^2(d) = \mathcal{O}(d^{-1})$，因此

$$|R_{White}| = \mathcal{O}(\mathcal{O}(d^{-1})) = \mathcal{O}(d^{-1}). \tag{97}$$

将方程 (95) 和方程 (97) 代入方程 (91)，得到：

$$\lim_{d\to\infty} \Delta(z) = \lim_{d\to\infty} [\mathcal{O}(d^{-1}) + \mathcal{O}(d^{-1})] = 0 \tag{98}$$

根据夹逼定理，

$$\lim_{d\to\infty} |m_{Noise}(z) - m_{White}(z)| = 0 \tag{99}$$

□

引理 3.4 的证明。 考察归一化矩阵 $\tilde{\Sigma} = \frac{1}{\nu^2(d)}\Sigma$：

$$\tilde{\Sigma} = \frac{1}{\nu^2(d)}\Sigma_{Signal} + \frac{1}{\nu^2(d)}\Sigma_{Noise}. \tag{100}$$

定义属于 $\frac{1}{\nu^2(d)}\Sigma_{Noise}$ 的特征值集合

$$\tilde{\mathcal{T}}_A = \{\tilde{\tau}_j = \frac{\tau_j}{\nu^2(d)}\}_{j=r+1}^{d}. \tag{101}$$

$\tilde{\mathcal{T}}_A$ 的斯蒂尔切斯变换为：

$$\tilde{m}_{Noise}(z) \triangleq \frac{1}{d-r}\sum_{j=r+1}^{d}\frac{1}{\tilde{\tau}_j - z}. \tag{102}$$

代入 $\tilde{\tau}_j$ 的定义，并将分子和分母同时乘以 $\nu^2(d)$，得到：

$$\tilde{m}_{Noise}(z) = \frac{1}{d-r}\sum_{j=r+1}^{d}\frac{1}{\frac{\tau_j}{\nu^2(d)} - z} \tag{103}$$

$$= \frac{1}{d-r}\sum_{j=r+1}^{d}\frac{\nu^2(d)}{\tau_j - \nu^2(d)z} \tag{104}$$

$$= \nu^2(d)\cdot\underbrace{\left(\frac{1}{d-r}\sum_{j=r+1}^{d}\frac{1}{\tau_j - \nu^2(d)z}\right)}_{m_{Noise}(\nu^2(d)z)} \tag{105}$$

剩余的求和项为 $m_{Noise}(z)$，其中自变量变为 $\nu^2(d)z$。因此，以下关系成立：

$$\tilde{m}_{Noise}(z) = \nu^2(d)\cdot m_{Noise}(\nu^2(d)z) \tag{106}$$

根据引理 3.3，$\lim_{d\to\infty}|m_{Noise}(z) - m_{White}(z)| = 0$，因此该关系成立：

$$\lim_{d\to\infty}\tilde{m}_{Noise}(z) = \lim_{d\to\infty}\nu^2(d)\cdot\left(\frac{1}{\nu^2(d) - (\nu^2(d)z)}\right) \tag{107}$$

$$= \lim_{d\to\infty}\nu^2(d)\cdot\frac{1}{\nu^2(d)(1-z)} \tag{108}$$

$$= \lim_{d\to\infty}\frac{\nu^2(d)}{\nu^2(d)}\cdot\frac{1}{1-z} \tag{109}$$

$$= \lim_{d\to\infty}\frac{1}{1-z} \tag{110}$$

$$= \frac{1}{1-z}. \tag{111}$$

定理 2.5 指出，若一列斯蒂尔切斯变换 $m_n(z)$ 收敛至极限 $m(z)$，且 $m(z)$ 对应于唯一的概率测度 μ，则底层分布 μ_n 必须弱收敛至 μ。方程(107)给出了噪声矩阵变换的渐近极限：

$$\lim_{d\to\infty} \tilde{m}_{Noise}(z) = \frac{1}{1-z}. \tag{112}$$

注意到，极限函数 $(1-z)^{-1}$ 是狄拉克质量 δ_1唯一的斯蒂尔切斯变换。事实上，

$$m_{\delta_1}(z) = \int_{\mathbb{R}} \frac{1}{t-z} d\delta_1(t) = \frac{1}{1-z}. \tag{113}$$

因此，将经验谱分布识别为测度序列 $\{\mu_n\}$后，我们调用定理 2.5，将代数极限直接转化为谱的几何坍缩：

$$\text{ESD}\left(\frac{1}{\nu^2(d)}\Sigma_{Noise}\right) \xrightarrow{w} \delta_1. \tag{114}$$

根据随机变量的线性性质，Σ_{Noise} 的经验谱分布弱收敛到集中于$\nu^2(d)$处的狄拉克测度，即，

$$\text{ESD}(\Sigma_{Noise}) \xrightarrow{w.} \delta_{\nu^2(d)}, \quad d\to\infty \tag{115}$$

□

证明 **定理 3.5**。根据引理 3.4，经验谱分布(Σ_{Noise}) 弱收敛到狄拉克测度$\delta_{\nu^2(d)}$ ：

$$\text{ESD}(\Sigma_{Noise}) \xrightarrow{w.} \delta_{\nu^2(d)}, \quad d\to\infty \tag{116}$$

显然，$\delta_{\nu^2(d)}$ 是一个确定性概率测度。由于 $\nu^2(d) = \mathcal{O}(d^{-1})$，支撑集 $\{\nu^2(d)\}$ 是一个紧集。因此，取 $\delta_{\nu^2(d)}$ 为 μ。根据引理 3.2，对于 $d\to\infty$，$k=1,\ldots r$，有如下关系：

$$\tau_k > \nu^2(d)\cdot(1+\sqrt{c}) \tag{117}$$

成立，这意味着 t hat Spec$(\Sigma_{signal}) = (\tau_1,\ldots,\tau_r,0,\ldots,0)$ 落入超临界区域。对于 $d\to\infty, k = r+1,\ldots,d$, 关系 n:

$$\tau_k \le \nu^2(d)\cdot(1+\sqrt{c}) \tag{118}$$

成立，这意味着 Spec$(\Sigma_{tail}) = (0,\ldots,0,\tau_{r+1},\ldots,\tau_d)$ 落入亚临界区间。 □

引理 3 的证明 **.6.** 根据定理 3.5，由于 Spec(Σ_{Signal}) 落入超临界区域，因此该关系成立 s:

$$\lambda_k \xrightarrow{a.s.} \rho(\tau_k) = \tau_k\left(1+\frac{c\cdot\nu^2(d)}{\tau_k-\nu^2(d)}\right). \tag{119}$$

注意：

$$\frac{1}{\tau_k-\nu^2(d)} = \frac{1}{\tau_k\left(1-\frac{\nu^2(d)}{\tau_k}\right)} = \frac{1}{\tau_k}\cdot\left(1-\frac{\nu^2(d)}{\tau_k}\right)^{-1} \tag{120}$$

应用 幂级数展开公式 $(1-x)^{-1} = 1+x+x^2+\mathcal{O}(x^3)$，其中 $x = {}^{\nu^2}{}_{(d)\,\tau_k}$，得到 —— s:

$$\left(1-\frac{\nu^2(d)}{\tau_k}\right)^{-1} = 1+\frac{\nu^2(d)}{\tau_k}+\left(\frac{\nu^2(d)}{\tau_k}\right)^2+\mathcal{O}(\tau_k^{-3}). \tag{121}$$

将公式（121）代入公式（119），得到：

$$\begin{aligned}\rho(\tau_k) &= \tau_k \left[1 + c\nu^2(d) \cdot \left(\frac{1}{\tau_k}\left(1 + \frac{\nu^2(d)}{\tau_k} + \frac{(\nu^2(d))^2}{\tau_k^2} + \dots\right)\right)\right] \\ &= \tau_k \left[1 + \frac{c\nu^2(d)}{\tau_k} + \frac{c(\nu^2(d))^2}{\tau_k^2} + \frac{c(\nu^2(d))^4}{\tau_k^3} + \mathcal{O}(\tau_k^{-4})\right] \\ &= \tau_k + c\nu^2(d) + \frac{c(\nu^2(d))^2}{\tau_k} + \mathcal{O}(\tau_k^{-2})\end{aligned} \tag{122}$$

根据几乎必然收敛的定义，

$$\lim_{\tau_k \to \infty} \frac{\lambda_k}{\tau_k} = \lim_{\tau_k \to \infty} \left(\frac{\tau_k + c\nu^2(d) + \frac{c(\nu^2(d))^2}{\tau_k} + O(\tau_k^{-2})}{\tau_k}\right) = 1 \tag{123}$$

根据定义 2.1，关系成立：

$$\lambda_k \sim \tau_k, \quad k = 1, ..., r. \tag{124}$$

类似地，由于 Spec(Σ_{tail}) 落在定理 3.5 的亚临界区间内，

$$\lambda_k \sim \nu^2(d)(1 + \sqrt{c})^2, \quad k = r+1, ..., r. \tag{125}$$

□

定理 3.7 的证明。 设$\mathbf{X}$的奇异值分解为$\mathbf{X} = U\Sigma_{svd}V^T$，其中 Σ_{svd}是一个对角矩阵，其对角元素为奇异值 $\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \cdots \geq \sigma_d(\mathbf{X}) \geq 0$。回忆一下，样本协方差矩阵定义为：

$$S_N = \frac{1}{N}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{d \times d}. \tag{126}$$

将$\mathbf{X} = U\Sigma_{svd}V^T$代入 S_N，得到：

$$\begin{aligned}S_N &= \frac{1}{N}(U\Sigma_{svd}V^T)(U\Sigma_{svd}V^T)^T \\ &= \frac{1}{N}U\Sigma_{svd}V^TV\Sigma_{svd}^TU^T\end{aligned} \tag{127}$$

由于 V是一个正交矩阵，而 Σ_{svd}是一个对角矩阵， $V^TV = I$, $\Sigma_{svd}^T = \Sigma_{svd}$。因此：

$$\begin{aligned}S_N &= \frac{1}{N}U(\Sigma_{svd}\Sigma_{svd})U^T \\ &= U\left(\frac{1}{N}\Sigma_{svd}^2\right)U^T\end{aligned} \tag{128}$$

注意，方程 (128) 是 S_N 的奇异值分解，其中 U 是特征向量矩阵。因此，

$$\lambda_k = \frac{1}{N}\sigma_k^2(\mathbf{X}), \tag{129}$$

由此可得：

$$\sigma_k(\mathbf{X}) = \sqrt{N \cdot \lambda_k}. \tag{130}$$

由于 N 是一个标量常数，在 d， $N \to \infty$ 的极限过程中，它只影响振幅，不影响分布的形状。根据引理 3.6，

$$\sigma_k(\mathbf{X}) \sim \sqrt{N \cdot \tau_k} = \begin{cases}\sqrt{NL} \cdot k^{-\alpha/2}, & k \leq r \\ \sqrt{N\nu^2(d)} \cdot (1 + \sqrt{c}), & r < k \leq d\end{cases}. \tag{131}$$

根据定义 2.2， $\sqrt{NL} \cdot k^{-\alpha/2} \sim \mathcal{O}(k^{-\alpha/2})$。回顾 $\nu^2(d)$ 的定义：

$$\nu^2(d) = \frac{1}{d-r}\sum_{j=r+1}^{d}\tau_j = \frac{L}{d-r}\sum_{j=r+1}^{d}j^{-\alpha}. \tag{132}$$

将 $\nu^2(d)$ 的定义代入 $\sqrt{N\nu^2(d)} \cdot (1+\sqrt{c})$，当 $d,\ N \to \infty$ 时，

$$\begin{aligned}\sqrt{N\nu^2(d)} \cdot (1+\sqrt{c}) &\to (1+\sqrt{\frac{d}{N}}) \cdot \sqrt{\frac{NL}{d-r}\sum_{j=r+1}^{d} j^{-\alpha}} \\ &= \sqrt{\frac{NL}{d-r}\sum_{j=r+1}^{d} j^{-\alpha}} + \sqrt{\frac{dL}{d-r}\sum_{j=r+1}^{d} j^{-\alpha}}\end{aligned} \tag{133}$$

由于 $\sum_{j=r+1}^{d} j^{-\alpha}$ 是黎曼ζ函数 $\zeta(\alpha)$ 的尾项，因此当 $d \to \infty$ 时它收敛。因此，

$$\sqrt{\frac{NL}{d-r}\sum_{j=r+1}^{d} j^{-\alpha}} = \mathcal{O}(d^{-\frac{1}{2}}), \quad \sqrt{\frac{dL}{d-r}\sum_{j=r+1}^{d} j^{-\alpha}} = \mathcal{O}(1). \tag{134}$$

将方程 (134) 代入方程 (133)，得到：

$$\sqrt{N\nu^2(d)} \cdot (1+\sqrt{c}) \sim \mathcal{O}(1). \tag{135}$$

因此，

$$\sigma_k(\mathbf{X}) \sim \begin{cases}\mathcal{O}(k^{-\alpha/2}), & 1 \le k \le r \\ \mathcal{O}(1), & r < k \le d\end{cases}. \tag{136}$$

根据定义 2.2，存在两个常数 $C_1,\ C_2$，使得：

$$\sigma_k(\mathbf{X}) \le \begin{cases}C_1 \cdot k^{-\alpha/2}, & 1 \le k \le r \\ C_2, & r < k \le d\end{cases}. \tag{137}$$

由于对所有 $k \in (r,\ d]$，都有 $\sigma_{k+1}(\mathbf{X}) \le \sigma_k(\mathbf{X})$，

$$\sigma_k(\mathbf{X}) \le \sigma_r(\mathbf{X}) \le C_1 \cdot r^{-\alpha/2} \tag{138}$$

取 $C_2 = C_1 \cdot r^{-\alpha/2}$ 得到：

$$\sigma_k(\mathbf{X}) \le \begin{cases}C_1 \cdot k^{-\alpha/2}, & 1 \le k \le r \\ C_1 \cdot r^{-\alpha/2}, & r < k \le d\end{cases} \tag{139}$$

□

定理 3.8 的证明。 根据链式法则，权重矩阵 $\nabla_{\mathbf{W}}$ 的梯度可通过输入矩阵的转置与误差信号矩阵的乘积计算得到：

$$\nabla_{\mathbf{W}} = \mathbf{X}^\top \mathbf{G}, \tag{140}$$

其中 $\mathbf{X}^\top \in \mathbb{R}^{N \times d}$。根据矩阵奇异值的乘积性质，对于任意两个矩阵 A 和 B，k-th 大的奇异值 AB 满足：

$$\sigma_k(AB) \le \|A\|_2 \cdot \sigma_k(B). \tag{141}$$

由于转置不会改变矩阵的奇异值，

$$\sigma_k(\nabla_{\mathbf{W}}) = \sigma_k(\nabla_{\mathbf{W}}^\top) = \sigma_k(\mathbf{G}^\top \mathbf{X}) \le \|\mathbf{G}^\top\|_2 \cdot \sigma_k(\mathbf{X}). \tag{142}$$

由于 $\|\mathbf{G}^\top\|_2 = \|\mathbf{G}\|_2$，

$$\sigma_k(\nabla_{\mathbf{W}}) \le \|\mathbf{G}\|_2 \cdot \sigma_k(\mathbf{X}) \le M \cdot \sigma_k(\mathbf{X}) \tag{143}$$

根据定理 3.7，$\mathbf{X}$ 的第 k 大奇异值满足以下渐近行为 r:

$$\sigma_k(\mathbf{X}) \le \begin{cases}C_1 \cdot k^{-\alpha/2}, & 1 \le k \le r \\ C_1 \cdot r^{-\alpha/2}, & r < k \le d\end{cases}, \tag{144}$$

其中 C_1 是常数。将方程（144）代入方程（143）得到：

$$\sigma_k(\nabla_{\mathbf{W}}) \leq \begin{cases}(MC_1)\cdot k^{-\alpha/2}, & 1 \leq k \leq r \\ (MC_1)\cdot r^{-\alpha/2}, & r < k \leq d\end{cases} \tag{145}$$

□

D 量化提高稳定秩的数学证明

定理 4 的证明。 **3.** 我们给出第一个等式的证明；第二个等式由对称性可得。根据定义 ，代入$E = \sum_{i,j}$ 将Z_{ij} 代入期望可得：

$$\mathbb{E}[EE^T] = \mathbb{E}\left[\left(\sum_{i,j} Z_{ij}\right)\left(\sum_{p,q} Z_{pq}\right)^T\right] \tag{146}$$

$$= \mathbb{E}\left[\sum_{i,j}\sum_{p,q} Z_{ij}Z_{pq}^T\right]. \tag{147}$$

利用期望算子的线性性，我们将期望移入双重求和内部：

$$\mathbb{E}[EE^T] = \sum_{i,j}\sum_{p,q}\mathbb{E}[Z_{ij}Z_{pq}^T]. \tag{148}$$

我们根据索引对（i，j）和（p，q）分析求和项 $\mathbb{E}[Z_{ij}Z_{pq}^T]$。当（$i$，$j$）$\neq$（$p$，$q$）时，由于量化噪声彼此独立，期望可以进行因式分解：

$$\mathbb{E}[Z_{ij}Z_{pq}^T] = \mathbb{E}[Z_{ij}]\mathbb{E}[Z_{pq}^T]. \tag{149}$$

由于 $\mathbb{E}[Z_{ij}] = 0$，这些项恒等为零（$0 \cdot 0 = 0$）。当（i，j）$=$（p，q）时，这些项对应于自方差，并保持为

$$\mathbb{E}[Z_{ij}Z_{ij}^T]. \tag{150}$$

因此，求和中的所有非对角交叉项均为零，表达式归结为对基矩阵进行单次求和：

$$\mathbb{E}[EE^T] = \sum_{i,j}\mathbb{E}[Z_{ij}Z_{ij}^T]. \tag{151}$$

类似地， $\mathbb{E}[E^TE] = \sum_{i,\ j}\mathbb{E}[Z_{ij}^TZ_{ij}]$ 在上述分析之后成立。引理4.3的证明至此完成。 □

定理 4.4 的证明。 在引理4.3得证之后，我们继续计算 δ^2 的值。我们逐元素计算矩阵$\mathbb{E}[EE^T]$。（i，j）项乘积矩阵 EE^T 的值由 E 的第 i 行与第 j 行的内积给出：

$$(EE^T)_{ij} = \sum_{k=1}^{n} e_{ik}e_{jk}. \tag{152}$$

对该元素取期望，并应用线性性：

$$\mathbb{E}[(EE^T)_{ij}] = \sum_{k=1}^{n}\mathbb{E}[e_{ik}e_{jk}]. \tag{153}$$

我们再次利用噪声项 e_{ik} 和 e_{jk}的独立性。对于 $i \neq j$ 的非对角元素，噪声项相互独立，因此

$$\mathbb{E}[e_{ik}e_{jk}] = \mathbb{E}[e_{ik}]\mathbb{E}[e_{jk}] = 0. \tag{154}$$

因此， $\mathbb{E}[EE^T]$ 是对角矩阵。对于满足 $i = j$的对角元素，期望变为方差和：$\mathbb{E}[e_{ik}^2] = B$。求和得到：

$$\sum_{k=1}^{n} \mathbb{E}[e_{ik}^2] = \sum_{k=1}^{n} B = nB. \tag{155}$$

因此， $\mathbb{E}[EE^T]$ 是大小为 $m \times m$ 、对角值为 nB的缩放单位矩阵。其谱范数就是该对角值：

$$\left\| \sum_{i,j} \mathbb{E}[Z_{ij}Z_{ij}^T] \right\|_2 = nB. \tag{156}$$

通过对 $\mathbb{E}[E^T E]$作对称论证，我们得到范数 mB。方差统计量 δ^2 由这两个值中的最大值决定：

$$\delta^2 = \max(nB, mB) = \max(m, n) \cdot B. \tag{157}$$

将这一特定的 δ^2 代入定理 2.7，我们得到量化误差的尾界。对于满足 $m, n \gg 1$的大规模深度学习模型，分母中的方差项 δ^2 主导有界性项 $Rt/3$ 。忽略高阶小量后，存在常数 C ，使得该界简化为：

$$P(\|E\|_2 \geq t) \leq (m+n) \cdot \exp\left(-\frac{t^2}{C \cdot \max(m,n) \cdot s^2}\right), \tag{158}$$

证明至此结束。 □

定理 4.5 的证明。 我们的目标是推导相对误差的概率上界，使得稳定性以至少为 $1-\theta$的概率成立。回顾可知， $\mathcal{R}(\epsilon_k)$表示相对误差超过阈值 ϵ_k的事件。我们将该违反条件代数地转化为绝对误差不等式：

$$\begin{aligned} \mathcal{R}(\epsilon_k) &\iff \left\{ \frac{|\sigma_k(\tilde{A}) - \sigma_k(A)|}{\sigma_k(A)} > \epsilon_k \right\} \\ &\iff \left\{ |\sigma_k(\tilde{A}) - \sigma_k(A)| > \epsilon_k \cdot \sigma_k(A) \right\}. \end{aligned} \tag{159}$$

根据定理 2.6，噪声 $\|E\|_2$ 的谱范数是绝对扰动 $|\sigma_k(A) - \sigma_k(A)|$的一个确定性上界。这带来了一个关键的逻辑推论：如果绝对扰动超过 $\epsilon_k \cdot \sigma_k(A)$，那么噪声范数 $\|E\|_2$ 必然也会超过这一阈值。用集合论意义上的语言来说，相对误差违约事件是谱范数尾事件的子集：

$$\left\{ |\sigma_k(\tilde{A}) - \sigma_k(A)| > \epsilon_k \cdot \sigma_k(A) \right\} \subseteq \{ \|E\|_2 > \epsilon_k \cdot \sigma_k(A) \}. \tag{160}$$

因此，根据概率测度的单调性，相对误差违约的概率严格小于谱范数的尾概率：

$$P(\mathcal{R}(\epsilon_k)) \leq P(\|E\|_2 > \epsilon_k \cdot \sigma_k(A)). \tag{161}$$

回顾公式（42）中推导出的尾界函数 $F(t) = P(\|E\|_2 \geq t)$，可得 ：

$$P(\mathcal{R}(\epsilon_k)) \leq F(\epsilon_k \cdot \sigma_k(A)). \tag{162}$$

为了保证指定的置信水平 $1-\theta$，我们必须确保违例概率不超过 θ。我们将上界严格等同于 θ：

$$F(\epsilon_k \cdot \sigma_k(A)) = \theta. \tag{163}$$

该方程定义了我们置信区间的边界。由于尾函数 $F(t)$ 在 $t > 0$ 上严格单调递减，其反函数 F^{-1} 存在。求解特定的相对误差阈值 ϵ_k 可得：

$$\epsilon_k \cdot \sigma_k(A) = F^{-1}(\theta) \implies \epsilon_k(\theta) \triangleq \epsilon_k = \frac{F^{-1}(\theta)}{\sigma_k(A)}. \tag{164}$$

在这一最终表达式中，分子 $F^{-1}(\theta)$ 表示一个仅由矩阵维度、噪声方差和置信水平 θ 决定的恒定噪声预算。它与索引 k 无关。这明确展示了反比例关系：

$$\epsilon_k(\theta) \propto \frac{1}{\sigma_k(A)}. \tag{165}$$

因此，对于大小为 $\sigma_i > \sigma_j > 0$ 的任意两个奇异值，在相同的置信水平 $1-\theta$ 下，严格成立：

$$\epsilon_i(\theta) < \epsilon_j(\theta). \tag{166}$$

至此，证明完成。 □

定理 4.6 的证明。 该证明依赖于扰动理论和集中不等式。根据定理 4.4，噪声范数 $||E||_2$ 超过阈值 t 的概率有界为：

$$P(\|E\|_2 \geq t) \leq (m+n) \cdot \exp\left(-\frac{t^2}{C \cdot D \cdot s^2}\right). \tag{167}$$

根据定理 2.6，绝对扰动由噪声范数界定：

$$|\sigma_k(\tilde{A}) - \sigma_k(A)| \leq \|E\|_2. \tag{168}$$

因此，失效条件 $R_k > \eta$ 蕴含 $\|E\|_2 > \eta \cdot \sigma_k(A)$。将 $t = \eta \cdot \sigma_k(A)$ 代入集中界可得：

$$\theta_k = P(R_k > \eta) \leq (m+n) \cdot \exp\left(-\frac{\eta^2 \cdot \sigma_k^2(A)}{C \cdot D \cdot s^2}\right). \tag{169}$$

现在通过代入假设 4.1 中的分段渐近谱分布来计算方程 (169)。

情形 1： 在区间 $1 \leq k \leq r$， $\sigma_k(A) \sim \mu \cdot k^{-\alpha/2}$。将该项平方得到 $\sigma_k^2(A) \sim \mu^2 \cdot k^{-\alpha}$。将其代入指数项：

$$\begin{aligned} \frac{\eta^2 \cdot \sigma_k^2(A)}{C \cdot D \cdot s^2} &\sim \frac{\eta^2 \cdot (\mu \cdot k^{-\alpha/2})^2}{C \cdot D \cdot s^2} \\ &= \frac{\eta^2 \cdot \mu^2 \cdot k^{-\alpha}}{C \cdot D \cdot s^2} \\ &= \underbrace{\frac{\eta^2 \mu^2}{CDs^2}}_{\text{Constant term}} \cdot \underbrace{k^{-\alpha}}_{\text{Decay term}}. \end{aligned} \tag{170}$$

将方程 (170) 应用于方程 (169) 得到：

$$\theta_k \lesssim (m+n) \cdot \exp\left(-\frac{\eta^2 \mu^2}{CDs^2} \cdot \frac{1}{k^\alpha}\right), \tag{171}$$

情形 2： 在区间 $r < k \leq \min(m,\ n)$ 中，奇异值变平至第 r 个分量的水平： $\sigma_k(A) \sim \mu \cdot r^{-\alpha/2}$。因此， $\sigma_k^2(A) \sim \mu^2 \cdot r^{-\alpha}$。将其代入指数项可得一个与 k 无关的常数界：

$$\theta_k \lesssim (m+n) \cdot \exp\left(-\frac{\eta^2 \mu^2}{CDs^2} \cdot \frac{1}{r^\alpha}\right). \tag{172}$$

综合这两种情况，显式的失效概率上界为：

$$\theta_k \lesssim (m+n) \cdot \exp\left(-\frac{\eta^2 \mu^2}{CDs^2} \cdot \xi_k\right), \tag{173}$$

其中

$$\xi_k = \begin{cases} k^{-\alpha}, & 1 \le k \le r \\ r^{-\alpha}, & r < k \le \min(m,n) \end{cases}, \tag{174}$$

至此，证明得证。 □

定理 4.7 的证明。为了分析稳定秩的变化，我们引入能量集中函数 $F_A(k)$，其定义为前 k 个主成分所捕获的总能量占比：

$$F_A(k) \triangleq \frac{H_k(A)}{H_k(A) + T_k(A)}, \tag{175}$$

其中，$H_k(A) = \sum_{i=1}^{k} \sigma_i^2(A)$ 和 $T_k(A) = \sum_{j=k+1}^{r} \sigma_j^2(A)$ 分别表示头部能量和尾部能量。对于量化矩阵 $\tilde{A}$，扰动后的能量可表示为 $H_k(\tilde{A}) = (1+\alpha)H_k(A)$ 和 $T_k(\tilde{A}) = (1+\beta)T_k(A)$，其中 $\alpha, \beta \ge -1$。将这些结果代入公式（175）得：

$$\begin{aligned} F_{\tilde{A}}(k) &= \frac{(1+\alpha)H_k(A)}{(1+\alpha)H_k(A) + (1+\beta)T_k(A)} \\ &= \frac{H_k(A)}{H_k(A) + \frac{1+\beta}{1+\alpha}T_k(A)}. \end{aligned} \tag{176}$$

根据定理 4.5 和定理 4.6，均匀量化噪声会显著增加尾部能量，同时在头部能量中引入可忽略的相对误差，这意味着 $\beta > \alpha$。因此，缩放因子满足 $\frac{1+\beta}{1+\alpha} > 1$。由此可知，$F_{\tilde{A}}(k)$ 表达式中的分母相较于分子增大，从而得到：

$$F_{\tilde{A}}(k) < F_A(k), \quad \forall k. \tag{177}$$

具体而言，如果 $\alpha = -1$,

$$F_{\tilde{A}}(k) = \frac{(1+\alpha)H_k(A)}{(1+\alpha)H_k(A) + (1+\beta)T_k(A)} = 0, \tag{178}$$

这也会导致

$$F_{\tilde{A}}(k) < F_A(k), \quad \forall k. \tag{179}$$

注意，当 $\beta > \alpha$ 时，$\beta > -1$。最后，我们注意到，稳定秩 $S_r(A) = \frac{\|A\|_F^2}{\|A\|_2^2}$ 在数学上等价于 $k = 1$处能量集中的倒数：

$$S_r(A) = \frac{\sum_i \sigma_i^2(A)}{\sigma_1^2(A)} = \frac{1}{F_A(1)}. \tag{180}$$

在 $k = 1$处应用公式（177）中的不等式，我们得出结论：

$$S_r(\tilde{A}) = \frac{1}{F_{\tilde{A}}(1)} > \frac{1}{F_A(1)} = S_r(A). \tag{181}$$

□

E 更多实验

E.1 量化误差无偏性的实证验证

我们的谱分析的有效性很大程度上依赖于假设 4.2，该假设认为量化噪声是均值为零的中心化随机变量（$\mathbb{E}[e_{ij}] = 0$）。为从实证层面验证这一理论前提，我们分析了预训练大语言模型中量化误差的分布。

我们采用标准的仅解码器 Transformer 架构，遵循 **GPT-2-124M** 配置（$L = 12$层、 $H = 12$ 个注意力头数、$d_{model} = 768$）。模型使用 AdamW 优化器在 OpenWebText 数据集上训练，峰值学习率为 1.5×10^{-4}，全局批次大小为 2,560（通过在 8 个 GPU 上进行梯度累积实现）。详细的超参数汇总于表 3。本文中的所有 FP4 训练均采用严格的 W4A4G4 量化方案，其中权重、激活值和梯度都统一以 E2M1 NVFP4 格式表示。需要指出的是，由于英伟达闭源 FP4 训练软件栈具有专有性质，目前无法在开放研究中使用原生硬件支持的 FP4 训练。因此，按照算法研究中的标准做法，我们通过 BF16 高保真模拟开展 NVFP4 实验。该模拟能够准确模拟 E2M1 格式的量化噪声和动态范围约束，同时在标准的 BF16 兼容硬件上执行计算。为验证误差特性，我们从模型全部 24 个子层的多层感知机（MLP）权重中提取了量化误差矩阵 $E = Q(W) - W$。图 3 展示了各层的误差频率直方图，并叠加了拟合的正态分布曲线。

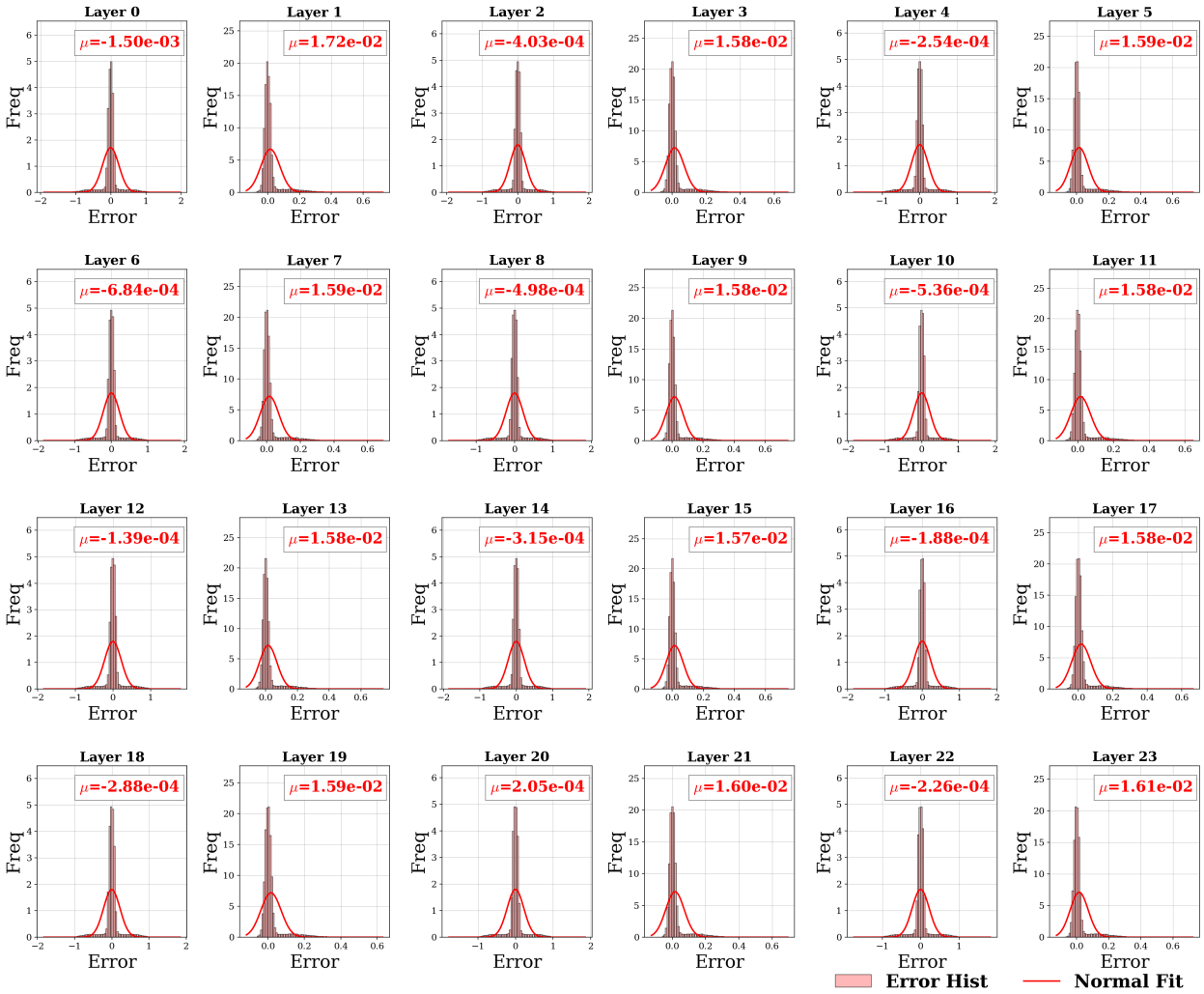


图 3：直方图展示了使用 NVFP4 量化的**GPT-2-124M**模型中多层感知机线性层的量化误差频率。红色曲线表示正态分布拟合。标注的 μ 数值表示每层误差的经验均值。

结果。 如图 3 所示，量化误差分布呈现出高度的零点对称性。尤为重要的是，所有层的经验均值(μ)都微不足道，范围从 10^{-4} 到 10^{-2} （例如，第 0 层：$\mu = -1.50 \times 10^{-3}$，第 10 层： $\mu = -5.36 \times 10^{-4}$），与权重的动态范围相比，在统计上并不显著。这些实证结果有力地支持了假设 4.2。虽然理论假设对噪声进行了简化，认为其服从均匀分布，

表 3：此表总结了 E.1 节中用于验证量化误差无偏性的模型架构、优化超参数和量化设置。

模型与数据		训练超参数	
架构	GPT-2（仅解码器）	**优化器**	AdamW
参数	≈ 124M	**峰值学习率**	1.5×10^{-4}
层（L）	12	**权重衰减**	0.1
隐藏维度 (d)	768	**全局批次大小**	2,560
注意力头 (H)	12	**梯度累积**	40 步
上下文窗口	1024 个词元	**预热步数**	20
数据集	OpenWebText	**梯度裁剪**	1.0
硬件	8 × GPU	**训练轮数**	1
量化配置（NVFP4）			
格式	NVFP4（E2M1）- 4 位浮点数（2 个指数位，1 个尾数位）		
范围	权重、激活值、梯度（全 FP4 训练）		
方案	分块对称均匀量化		

但谱扰动界所要求的关键性质是噪声的无偏性（即一阶矩 $\mathbb{E}[E] = 0$）。观测到的分布证实，NVFP4 量化机制不会给权重矩阵引入系统性偏差，从而验证了在后续谱分析中采用中心化随机矩阵理论的合理性。

E.2 相对误差界的验证

在确认误差无偏性的基础上，本实验旨在严格验证定理 4.5。该定理提出了一种基本的谱不对称性：相对量化误差 ϵ_k 与奇异值 σ_k的幅值严格成反比。数学上，这一关系表示为 $\epsilon_k \propto 1/\sigma_k(A)$。本实验采用 E.1 节中所述并在表 3 中总结的相同实验设置。我们使用相同的预训练 **GPT-2-124M** 模型，并对所有线性层应用模拟的 NVFP4（W4A4G4）量化。对于每一层 l，我们对原始参数矩阵 $A^{(l)}$ 及其量化对应矩阵 $\tilde{A}^{(l)}$进行奇异值分解（SVD）。随后，我们针对每个秩索引 k计算逐分量相对误差 ϵ_k：

$$\epsilon_k = \frac{|\sigma_k(\tilde{A}^{(l)}) - \sigma_k(A^{(l)})|}{\sigma_k(A^{(l)})} \tag{182}$$

为检验定理 4.5 预测的线性性，我们对 ϵ_k 与奇异值倒数 $1/\sigma_k(A^{(l)})$ 进行了线性回归分析。

结果。 结果见图 4，涵盖第 0 层至第 23 层。可视化证据有力地印证了我们的理论推导。在大多数层中，我们观察到相对误差 ϵ_k 与奇异值 $1/\sigma_k$的倒数之间存在显著的线性关系。这证实了方程 (44) 的预测，即 $\epsilon_k(\theta) \propto \frac{1}{\sigma_k(A)}$。线性回归拟合得出的决定系数较高（$R^2$）。例如，较深的层，如第 9 层（$R^2 = 0.99$）、第 19 层（$R^2 = 0.97$）和第 23 层（$R^2 = 0.98$），都几乎完美地符合理论模型。尽管某些初始层（例如 $R^2 = 0.42$ 的第 0 层）由于初始化分布不同而呈现较高方差，但整体趋势仍保持一致。数据显示，主导奇异值（$1/\sigma_k$ 较小时）对应的相对误差可忽略不计，从而有效保留了头部信息。相反，当我们移至尾部（$1/\sigma_k$ 较大时），相对误差呈线性增长，导致显著的信号失真。这一机制解释了谱平坦化现象：量化噪声底会不成比例地破坏低能量分量，而这些分量对于细粒度语义差异至关重要；与此同时，粗粒度结构信息则相对完好地保留下来。这项实验为定理 4.5 提供了有力的实证支持，证实均匀量化并不会均匀地降低信息；相反，它会选择性地削弱对大语言模型性能至关重要的重尾谱结构。

E.3 稳定秩增长的普适性

定理 4.7 预测，均匀量化会使谱分布天然变平，从而严格提高参数矩阵的 StableRank（$S_r(A) > S_r(A)$）。为展示这一现象的普遍性，我们

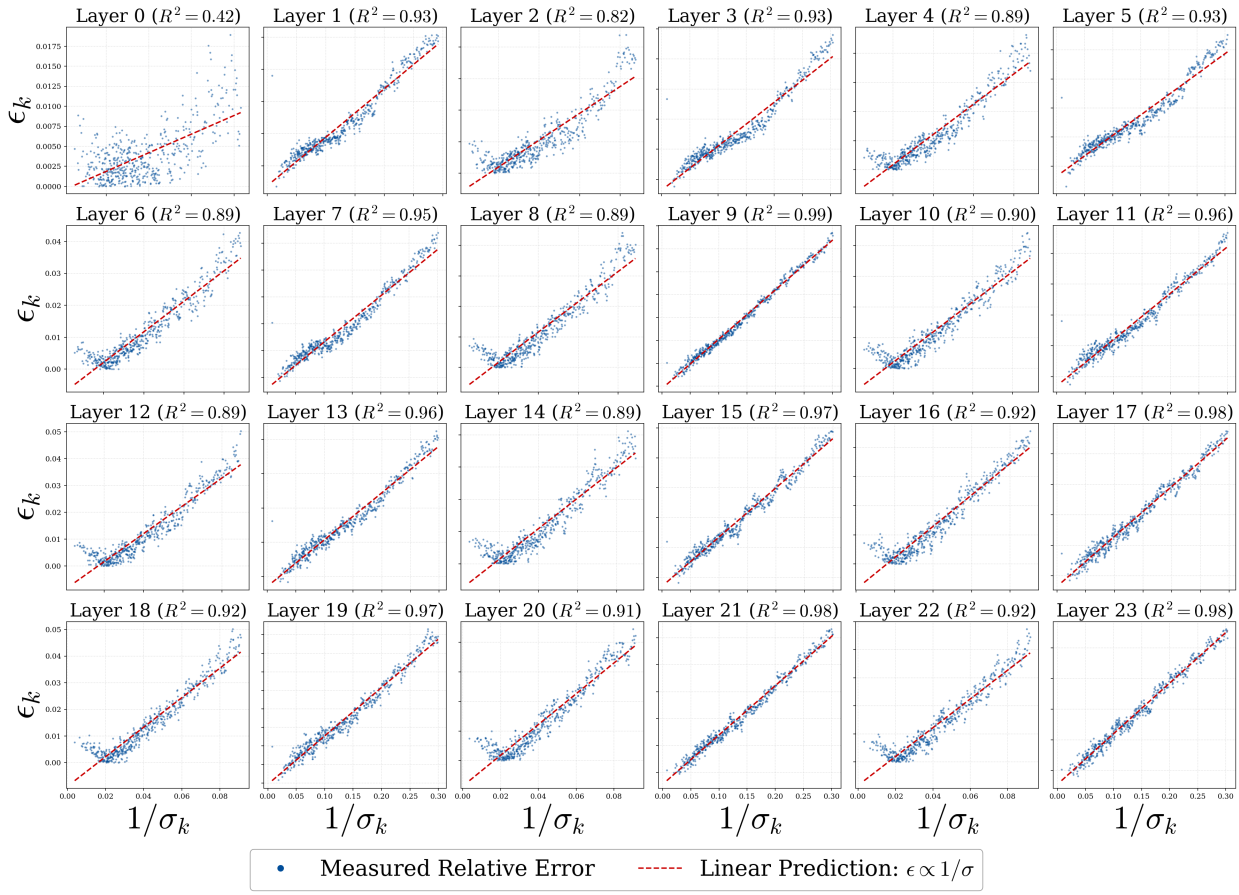


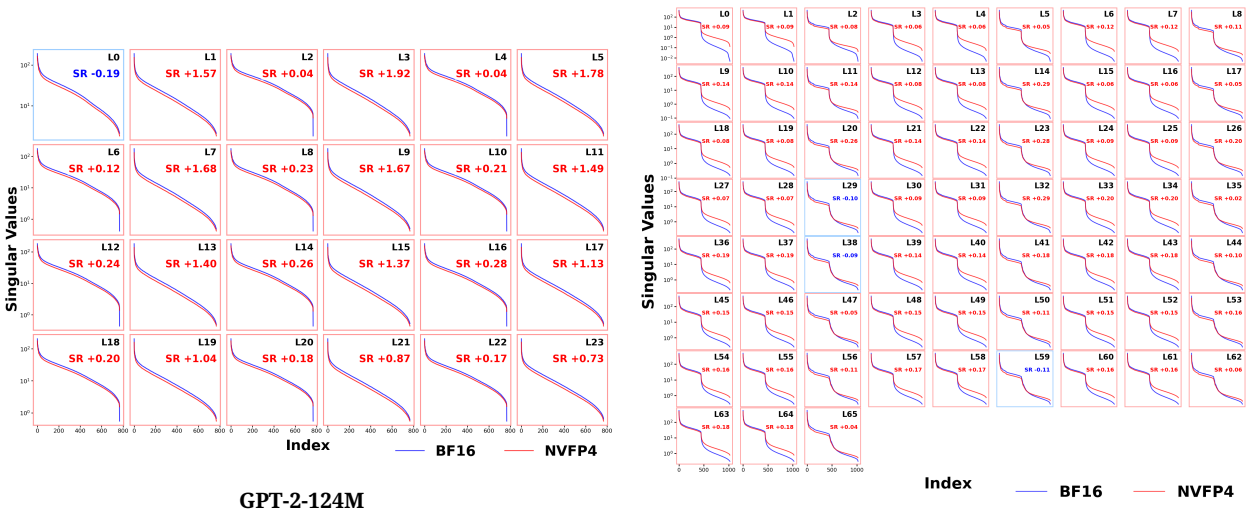
图 4：散点图展示了**GPT-2-124M** 模型各层中相对量化误差 ϵ_k（y 轴）与奇异值 $1/\sigma_k$ 的倒数（x 轴）之间的关系。红色虚线表示线性回归拟合。每层都标注了决定系数（R^2），以显示拟合优度。

表4：E.3节的模型架构与训练超参数。此表详细列出了用于验证定理 4.7 普遍性的各类大语言模型。所有模型均采用 AdamW 优化器、余弦学习率调度和 W4A4G4（NVFP4）量化。

模型系列	参数量	层	隐藏维度	注意力头数	KV头数	词汇表大小	峰值学习率	上下文窗口
GPT-2	124M	12	768	12	12	50,257	1.5×10^{-4}	1024
TinyLlama	1.1B	22	2048	32	8	32,000	4.0×10^{-3}	1024
Phi-1.5	1.3B	24	2048	32	32	51,200	4.0×10^{-4}	1024
Pythia	1.4B	24	2048	16	16	50,304	4.0×10^{-4}	1024
DeepSeek-Coder	1.3B	24	2048	32	32	50,304	4.0×10^{-4}	1024
Qwen2.5	1.5B	28	1536	12	2	151,936	4.0×10^{-4}	1024

我们在多种大语言模型上进行了完整的 FP4 训练（模拟 W4A4G4）。实验涵盖 **GPT-2-124M**以及五种最先进的架构：**TinyLlama-1.1B**、**Phi-1.5-1.3B**、**Pythia-1.4B**、**DeepSeek-Coder-1.3B**和**Qwen2.5-1.5B**。这些模型涵盖了多种设计选择，包括不同的深度（12 至 28 层）、嵌入维度（768 至 2048）、注意力头数（12 至 32）和词汇表大小（32k 至 152k）。详细的架构规格和训练超参数汇总于表4。所有模型均使用 NVFP4 量化，在 OpenWebText 数据集上通过 8 个 GPU 进行训练。我们比较了各模型所有层中量化前权重（BF16）的奇异值谱与其量化后对应权重（NVFP4）的奇异值谱。结果见图5、图6和图7。

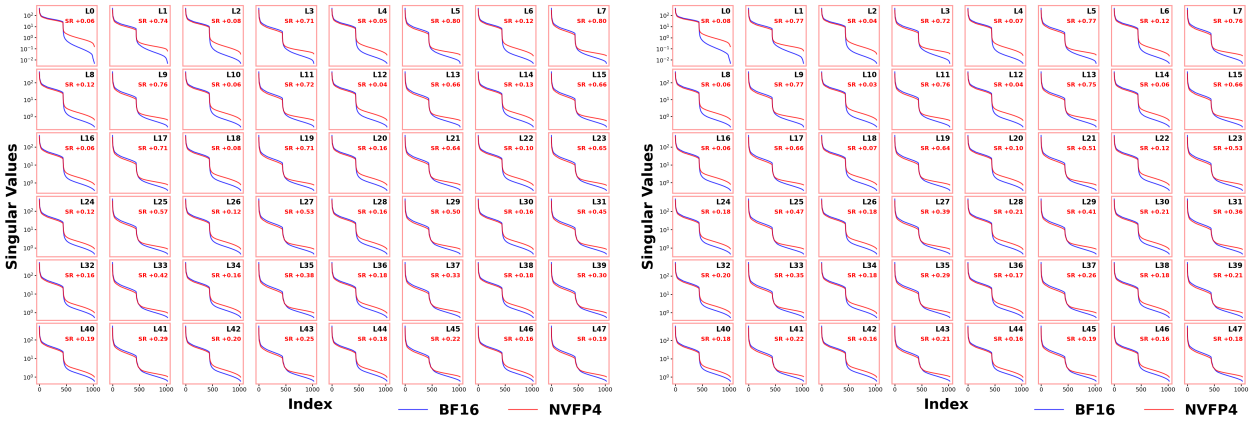
结果。 如图 5-7 所示，每种经过测试的架构都呈现出典型的谱平坦化模式。头部奇异值保持相对稳定，而尾部奇异值则因量化



GPT-2-124M　　TinyLlama-1.1B

图5：对**GPT-2-124M** 和 **TinyLlama-1.1B**进行的谱分析，其中量化使大多数层的稳定秩（SR+）出现一致的正向偏移。

噪声底而显著升高。从定量角度看，这导致绝大多数层的稳定秩严格增加（图中以 SR+ 表示）。从**GPT-2-124M**开始，我们观察到了基础性的谱退化。在规模更大、深度更深的模型中，这一趋势进一步增强。在 **Qwen2.5-1.5B**中，其采用深而窄的结构（28 层，1536 维），中间层的稳定秩增幅尤为明显。同样，**TinyLlama-1.1B**和**Phi-1.5-1.3B**也表现出一致的谱退化。这种跨模型的一致性验证了定理 4.7 是支配低精度嵌入的基本规律，并表明，在重尾分布中，均匀量化通过尾部截断造成的表征能力损失是不可避免的后果。



Phi-1.5-1.3B　　Pythia-1.4B

图6：对**Phi-1.5-1.3B** 和**Pythia-1.4B**进行的谱分析，其中量化使大多数层的稳定秩（SR+）出现一致的正向偏移。

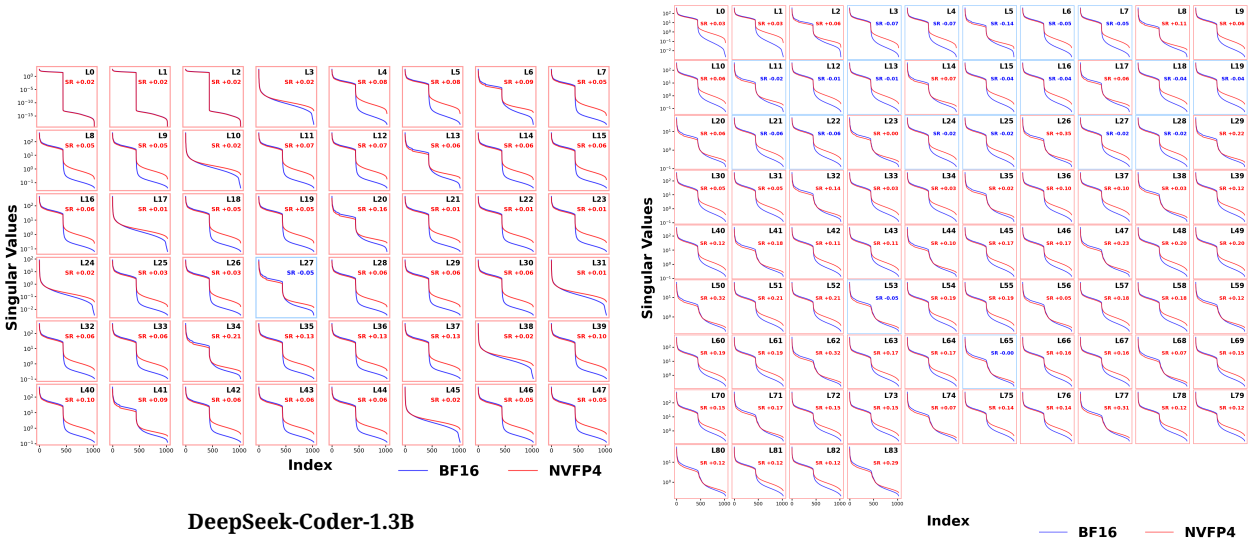


图7：**DeepSeek-Coder-1.3B** 和**Qwen2.5-1.5B**的谱分析，其中量化会使大多数层中的StableRank（SR+）一致地正向偏移。