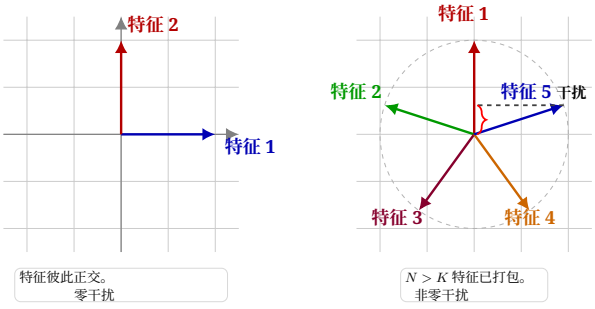
arXiv:2602.01045v1 [cs.LG] 1 Feb 2026

叠加统一了幂律训练动力学

Zixin Jessie Chen [1] **Hao Chen** [2] **Yizhou Liu** [3] **Jeff Gore** [1]

摘要

我们采用教师—学生框架，研究特征叠加在幂律训练动力学涌现中的作用。我们首先推导了无叠加情况下训练的解析理论，发现幂律训练指数取决于输入数据统计和通道重要性。值得注意的是，我们发现，叠加瓶颈会引发向普适幂律指数~ 1的转变，且该指数与数据和通道统计无关。与没有叠加时发生的纯序列学习相比，这种随时间进行的叠加训练可将训练速度最多提高十倍。我们发现，叠加能够带来快速训练，且幂律指数不依赖于数据；这一发现可能对广泛采用叠加的神经网络（包括生产级大语言模型）具有重要意义。



（a）无叠加　　（b）叠加

图*1.* **叠加的几何示意图。(a)** 没有叠加时，各特征之间不存在干扰。**（b）**叠加将特征压缩到更小的潜在规模中，从而引入特征之间的干扰（例如，特征5投影到特征1上）。

1. 引言

大语言模型的显著成功以神经网络缩放定律为支撑，其中模型性能随计算量、数据集规模和参数数量以幂律可预测地缩放（Kaplan 等人，2020年；Hoffmann 等人，2022年；Zhai 等人，2022年；Clark 等人，2022年）。这些定律不仅描述最终性能，也描述优化本身的轨迹，其中训练损失$\mathcal{L}(t)$通常按$\mathcal{L}(t) \propto t^{-\alpha}$在训练步数的多个数量级上衰减。尽管这些动力学普遍存在于各种模态（Henighan 等人，2020年）和架构（Gu &Dao，2024年），主导这一宏观行为的微观机制仍是激烈争论的主题。

标准学习理论通常将这些幂律归因于数据的谱性质（Advani 等人，2020年；Bahri 等人，2020年；Sharma 和 Kaplan，2020年）。在这一观点中，学习是一个序贯的谱滤波过程，模型按幅度降序拟合数据协方差的本征模（Saxe 等人，2013；Bordelon 等人，2020年；Canatar 等人，2021年）。尽管这种谱偏差成功解释了线性机制和核方法中的动力学（Rahaman 等人，2019；Cui 等人，2022年；Bordelon 等人，2025年a；b），但它通常假设模型维度与数据特征之间存在直接映射。然而，现代大语言模型运行在一种根本不同的机制中：特征的潜在表征彼此高度非正交，这一现象被称为叠加（Elhage 等人，2022年）。

大语言模型中的叠加。 大语言模型中嵌入空间的维度是一个关键瓶颈。模型必须为包含 32,000 至 256,000 个词元的词汇表（Achiam 等人，2024年；Mesnard 等人，2024年）以及数百万个抽象概念（Templeton 等人，2024年；Gao 等人，2024年）学习嵌入向量，却要将它们映射到仅有几千个维度的隐藏状态中。为突破这一限制，模型利用叠加，将特征存储在非正交方向上（Henighan 等人，2023年；Bricken 等人，2023年）。虽然这提高了容量，却引入了“干扰噪声”。值得注意的是，先前关于神经网络缩放定律的研究通常隐含地假设处于宽度足够的情形，

[1]麻省理工学院物理系，剑桥，马萨诸塞州 02139，美国 [2]普林斯顿大学电气与计算机工程系，普林斯顿，新泽西州 08544，美国 [3]麻省理工学院机械工程系，剑桥，马萨诸塞州 02139，美国。联系：Zixin Jessie Chen< jzxchen@mit.edu>。

预印本。*2026*年*2*月*3*日。

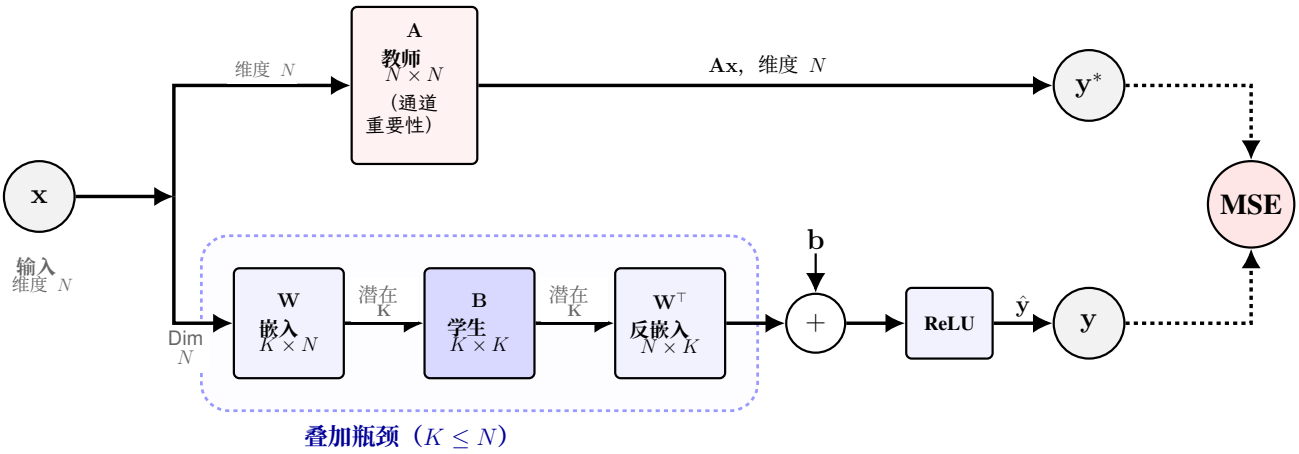


图*2*。 **叠加下的师生模型设置**，其中 $K \leq N$。学生通过嵌入层 $\mathbf{W}$ 和 $\mathbf{W}^\top$在压缩的潜在空间中学习。对输出应用偏置和 ReLU 非线性，以抑制干扰噪声。

即模型维度足以覆盖特征空间（Maloney 等人，2022年）。这一假设与生产模型的**强叠加** 情形并不相符（Liu 等人，2025年），这引出了一个问题：叠加固有的干扰噪声会如何改变训练动态?

为回答这个问题，我们必须超越黑箱观察。正如受控的 **玩具模型** 对于分离深度线性网络的作用机制至关重要（Saxe 等人，2019），我们提出一个易于分析的教师—学生框架，用于模拟大语言模型在叠加下的训练动态。我们的模型剥离了 Transformer 的架构复杂性，完全聚焦于特征结构与维度约束之间的相互作用。

我们的研究发现，训练机制存在令人惊讶的分化。在没有叠加的情况下，学生按序学习，其幂律训练动力学严格由输入数据统计和通道衰减决定。然而，当被迫采用叠加时，训练动力学便趋于统一。叠加的随机性会在不同通道之间“混合”特征，从而有效地使它们的学习行为趋于一致。这会带来一个**普适的幂律指数** $\alpha \approx 1$，不受具体输入数据和通道衰减的影响，同时相较于序贯情形有所加速。

1.1. 贡献

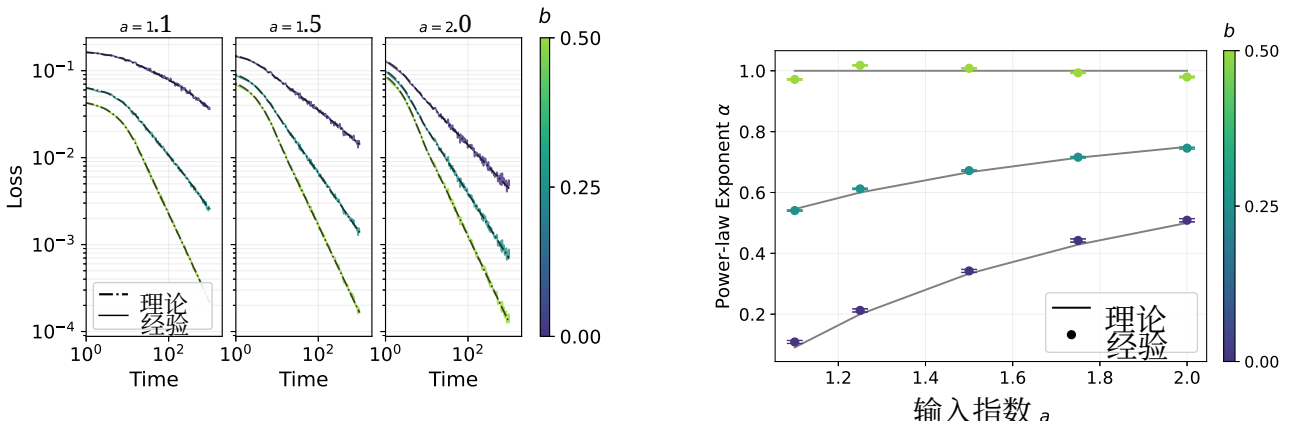
我们的贡献如下：

- 我们提出了一种教师-学生玩具模型，用以刻画特征分布、通道重要性与叠加之间的相互作用，反映生产规模大语言模型所面临的约束。
- 我们为幂律训练建立了一套解析理论动力学，并明确揭示训练指数与输入数据统计及通道重要性之间的联系，该理论适用于无叠加 情形。
- 我们通过实证证明，叠加机制会引发向普适、加速的幂律训练的转变，其指数为 $\alpha \approx 1$。
- 我们将叠加与训练动力学的标度律联系起来，并提出：与先前研究所假设的充分宽度机制相比，大语言模型特有的叠加机制会带来一致且加速的训练轨迹。

组织结构。 本文其余部分安排如下。第2 节介绍玩具模型设置。第3 节推导无叠加模型的解析理论。第 4节展示普适加速的实证结果，并分析最优计算量缩放前沿。最后，第5节对叠加下的普适行为给出机制解释。我们在第 6节作总结。

1.2. 相关工作

神经网络缩放定律。幂律缩放已在预训练（Kaplan 等人，2020年；Hoffmann 等人，2022年）、迁移学习（Hernandez 等人，2021年）和涌现能力（Wei 等人，2022年）方面得到充分证实。这些定律的扩展研究涵盖了数据剪枝（Sorscher 等人，2022年）、混合专家架构（Clark 等人，2022年）和状态空间模型（Gu & Dao，2024年）。理论推导通常依赖核方法或无限宽极限（Bordelon 等人，2020年；Canatar 等人，2021年），并将缩放归因于数据流形的谱衰减（Sharma 和 Kaplan，2020年）或损失景观中奇点的消解（Wei 等人，2019年）。我们的工作则另辟蹊径，研究叠



(a) 理论预测与经验损失轨迹吻合　　*(b)* 指数与 $\alpha = (a + 2b - 1)/a$ 吻合

图 *3.* **没有叠加时，学习指数取决于输入数据和通道统计**。 我们验证了无叠加基线 ($N = K = 1024$) 的解析理论。**(a)** 在不同的输入衰减 a 和通道重要性 b 下，经验损失曲线（实线）均紧跟理论预测（虚线）。**(b)** 拟合得到的幂律指数与推导出的缩放定律 $\alpha = (a + 2b - 1)/a$ 精确吻合，证实学习完全由数据和通道统计决定。

讨叠加机制——即宽度成为瓶颈的情形。这种约束在基于核的方法理论中往往缺失，却是大语言模型的决定性特征（Liu 等人，2025年）。

学习理论。 深度学习理论中的一个核心争论，涉及神经切线核（NTK）机制（Jacot 等人，2018；Arora 等人，2019）与特征学习机制（Chizat 等人，2019；Yang& Hu，2022年）之间的区别。NTK机制预测由固定谱性质支配的动力学（Saxe 等人，2013；Rahaman 等人，2019；Bordelon 等人，2025年a；b），而大语言模型被认为运行于表征不断演化的特征机制中。我们的工作表明，叠加起到桥梁作用：尽管学生模型是线性的（与许多NTK分析一样），特征混合会诱发一种集体学习动态，从而打破NTK机制中标准的谱关联。

机制可解释性。机制可解释性旨在对神经网络进行逆向工程，将其拆解为可理解的组件（Olah 等人，2020年）。一个关键关注点是叠加假说，该假说认为，模型所表征的特征数量超过其神经元数量（Elhage 等人，2022年）。近期研究利用稀疏自编码器解缠这些混合表征（Bricken 等人，2023年；Huben 等人，2024年）。然而，大多数研究关注的是静态的、经过训练的网络。关于学习动力学的研究主要聚焦于算法任务中的“顿悟”（Power 等人，2022年；Nanda 等人，2023年）。我们通过将叠加作为一种积极塑造训练动态连续缩放定律的机制单独加以研究，为这一领域作出了贡献。

2. 玩具模型设置

为了探究特征叠加与训练动态之间的相互作用，我们提出了一个师生框架，用于模拟大语言模型所特有的特征结构和维度约束。

2.1. 任务定义：数据与通道

我们通过**稀疏输入分布**和教师模型来定义学习任务，以模拟自然语言数据的多尺度结构以及Transformer通道重要性。

输入特征。 我们考虑一个输入向量 $\mathbf{x} \in \mathbb{R}^N$，其中每个分量 x_i 表示一个特征激活。为捕捉自然数据的稀疏性， x_i 定义为伯努利分布与均匀分布的混合：

$$x_i = u_i v_i, \quad u_i \sim \text{Bernoulli}(p_i), \quad v_i \sim U(0,1). \tag{1}$$

这里， p_i 决定特征频率，并遵循幂律衰减 $p_i \propto i^{-a}$（其中 $a > 1$）。该分布确保低索引特征的出现频率显著高于高索引特征。我们使用激活密度 $E = \sum_i p_i = 1$ 对 p_i 进行归一化，以在不同 N下保持激活密度一致。

教师模型。 目标信号由一个固定的教师矩阵 $\mathbf{A} \in \mathbb{R}^{N\times N}$ 生成，该矩阵对 Transformer 块中观测到的**通道重要性**进行建模。我们对角元素施加幂律谱衰减：

$$A_{ii} = i^{-b}, \quad A_{ij} = 0 \text{ for } i \neq j, \tag{2}$$

其中， $b \geq 0$ 是衰减常数。这种结构迫使模型优先学习低索引特征，以最小化

损失，从而在优化景观中建立有序序列。

2.2. 学生架构与叠加

学生模型 $f_\theta(\mathbf{x})$ 尝试在瓶颈维度K的约束下，重建教师的输出 $\mathbf{y}^* = \mathbf{A}\mathbf{x}$。

叠加与干扰。 如图1所示，当 $K < N$ 时，学生必须利用叠加来表示 N 特征。虽然这提高了容量，却引入了非正交干扰噪声。为对此建模，学生通过固定的、列归一化的随机投影（嵌入）$\mathbf{W} \in \mathbb{R}^{K\times N}$将稀疏输入压缩为潜在状态$\mathbf{h} = \mathbf{W}\mathbf{x} \in \mathbb{R}^K$。学生矩阵$\mathbf{B} \in \mathbb{R}^{K\times K}$ 对这一潜在表示进行处理。

随后，通过转置$\mathbf{W}^\top$对信号进行解码。关键在于，为了管理叠加固有的干扰噪声，我们引入可学习偏置$\mathbf{b}$，并在输出端使用 ReLU 非线性：

$$\mathbf{y} = \mathrm{ReLU}(\mathbf{W}^\top\mathbf{B}\mathbf{W}\mathbf{x} + \mathbf{b}). \tag{3}$$

由于真实输入信号是非负的（$x_i \geq 0$），ReLU 函数便充当纠错机制，抑制由非正交基产生的负干扰分量。

图 2 展示了叠加条件下师生模型的图示。请注意，在不应用**叠加** 且 $K = N$时，嵌入层 $\mathbf{W}$ 和 $\mathbf{W}^\top$ 均取为单位矩阵。在无叠加的情况下，不应用偏置或 ReLU 非线性。

训练目标。 学生模型通过最小化相对于教师输出的均方误差（MSE）进行训练。目标函数为：

$$\mathcal{L} = \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[\|\mathbf{y}^* - \mathbf{y}\|_2^2\right]. \tag{4}$$

在所有经验实验中，我们均关注相同的特征维度 $N = 1024$，并为简洁起见省略损失除以N 的归一化操作。在此设置下，动力学由三个参数决定：特征衰减 a、通道重要性衰减 b和压缩比 N/K。

3. 理论

本节将推导无叠加情形（$N = K$，$\mathbf{W} = \mathbf{I}$）下训练动态的解析标度律。叠加情形涉及复杂的干扰项（留待附录 A.2讨论），而对角情形则有闭式解，可作为我们的基线。我们在图 3 中通过经验验证了这一预测，结果显示理论与实验高度吻合。

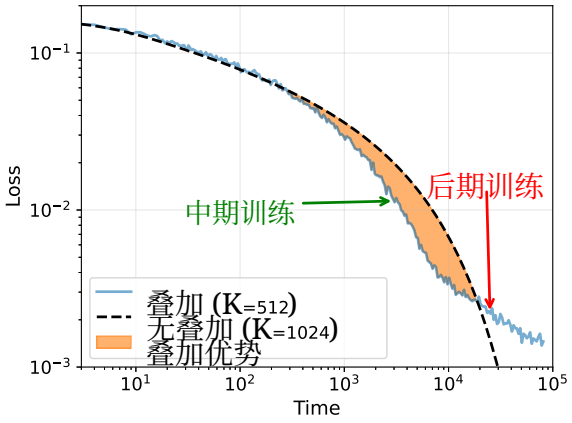


图 *4.* **叠加带来的训练中期加速。** 将叠加实验（$N = 1024$, $K = 512$）中的损失与无叠加理论（$N = K = 1024$）进行比较。尽管 K 中存在瓶颈，叠加仍使损失收敛在训练中期加速。

3.1. 无叠加时的动态

在不存在叠加的情况下，学生$\mathbf{B}$ 可以完全捕获教师$\mathbf{A}$。假设采用小初始化（$\mathbf{B}(0) \approx 0$），矩阵 $\mathbf{B}(t)$ 在整个训练过程中始终保持对角形式。损失可分解为各个独立特征模态上的和：

$$\mathcal{L}(t) \approx \frac{1}{2}\sum_{i=1}^{N}\lambda_i(s_i(t) - a_i)^2, \tag{5}$$

其中 $\lambda_i \propto i^{-a}$ 是数据协方差中占主导地位的对角方差。$a_i = i^{-b}$ 和 $s_i(t)$分别是教师系数和学生对角元素。

在学习率为 η的连续时间极限下，每个对角元素 $s_i(t)$的动力学遵循线性常微分方程：

$$\frac{ds_i}{dt} = \eta\lambda_i(a_i - s_i) \implies s_i(t) = a_i(1 - e^{-\eta\lambda_i t}). \tag{6}$$

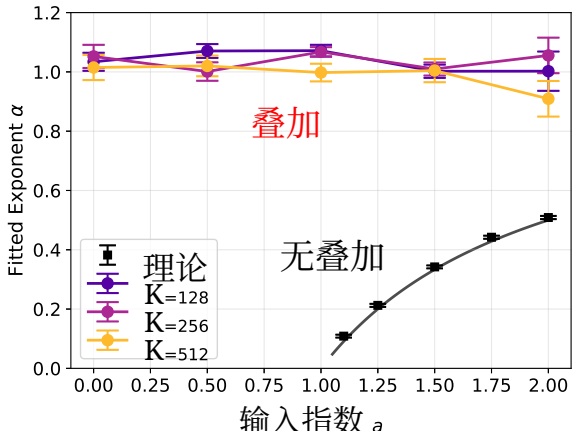
将该解代回损失函数可得：

$$\mathcal{L}(t) = \frac{1}{6}\sum_{i=1}^{N} i^{-(a+2b)}\exp\left(-\frac{2}{3}\eta t\cdot i^{-a}\right). \tag{7}$$

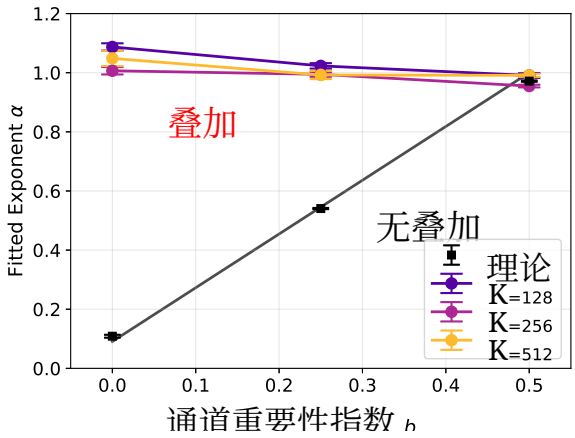
方程7 揭示了梯度下降的谱滤波本质：特征按其数据频率依次被学习 i^{-a}。在本文其余部分，我们将学习率与步数合并，并将时间定义为 $\eta\cdot t$，并**简单地用t**表示时间。

3.2. 幂律指数的推导

为提取渐近标度行为，我们将方程7 中的求和近似为积分。我们定义一个临界特征索引 $i_c(t) \propto t^{1/a}$，用以表示



(a) 指数 $\alpha \approx 1$ 与输入衰减 a无关



(b)指数 $\alpha \approx 1$与通道重要性无关 b

图 5。 **叠加机制下训练指数的普适性**。 我们绘制了不同学生规模下的幂律拟合指数 $\alpha\, K \in \{128, 256, 512\}$。与 α 会随数据和通道变化的序贯情形不同，叠加机制将指数固定为 $\boldsymbol{\alpha \approx 1}$，不受输入特征衰减 a 或通道重要性衰减 b的影响。

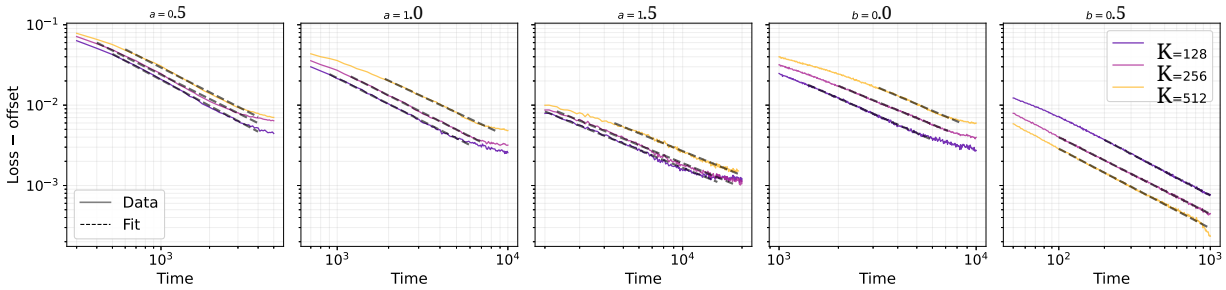


图 6。 **幂律动力学的稳健性**。 我们绘制了多种输入参数（$a \in [0.5, 1.5]$，$b \in [0, 0.5]$）下的样本损失曲线（实线）及其对应的幂律拟合（黑色虚线）。在所有情况下，训练中期动力学都能用指数恒为 $\alpha \approx 1$的幂律准确建模，这证实了幂律训练指数的普适性。

已学习特征与未学习特征之间的边界。假设特征维度 N很大，训练中期动力学（其中$1 \ll i_c \ll N$）由未学习特征的尾部主导。如附录A.1所述，这一积分得到幂律衰减：

$$\mathcal{L}(t) \propto t^{-\alpha}, \quad \text{where } \alpha = \frac{a + 2b - 1}{a}, \quad (a + 2b > 1). \tag{8}$$

这一结果表明，在序列学习机制中，训练速度与输入统计量（a）和通道重要性（b）严格相关。随着 a，b 增大，输入变得更加稀疏，通道衰减也更快，留给学生从教师那里学习的新信息减少，从而使学习轨迹加快。

4. 实验

我们现在研究叠加下的训练动态，此时由于特征混合，解析指数变得难以处理。我们对大小为 $N = 1024$的玩具模型进行实验，该模型通过随机梯度下降并结合在线数据生成进行训练，

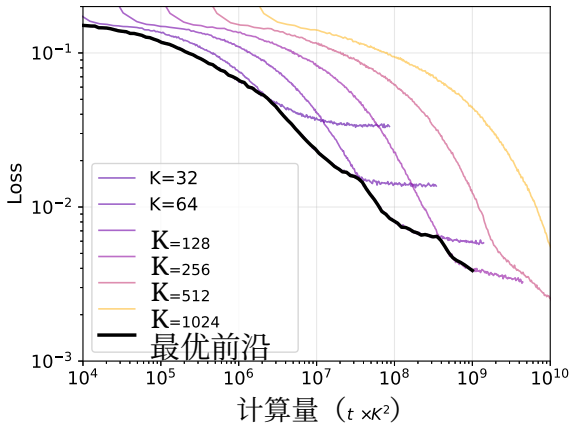
mimicking LLM预训练的无限数据机制 aining.

4.1. 方法论

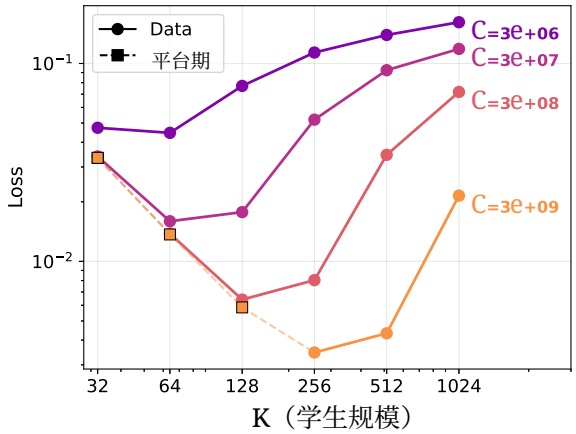
为了量化训练动力学，我们调整了Chinchilla缩放定律（Hoffmann 等人，2022年），以考虑不同的优化阶段。标准缩放定律将损失建模为时间相关项与参数相关项之和：

$$\mathcal{L}(t, K) \approx c_t t^{-\alpha} + c_k K^{-\beta} + \mathcal{L}_0. \tag{9}$$

然而，在我们的师生设置中，优化过程分为两个截然不同的阶段：早期至中期阶段由优化误差主导（受时间限制），后期阶段则由瓶颈容量主导（受宽度限制），如图 4 所示。在分析中，我们聚焦于训练中期，此时 $c_t t^{-\alpha} + \mathcal{L}_0$ 占主导地位，以提取有效训练指数 α。



(a) 最优计算前沿



(b) 等计算曲线可确定最优学生规模

图7. **模型规模缩放下最优计算前沿的示例**。我们分析固定输入分布（$a = 1.1$，$b = 0$）下模型规模与训练时长之间的权衡。**(a)** 原始训练曲线（彩色线）被一条黑色实线包络，该线定义了最优计算前沿。**(b)** 我们绘制了四个固定计算预算 C（对数均匀分布）下损失相对于学生规模 K 的变化。每条曲线中的不同极小值展示了 K 与 C之间的权衡，反映了大语言模型中观察到的行为。

4.2. 结果

我们改变输入衰减 $a \in [1.1, 2.0]$、通道重要性$b \in [0.0, 0.5]$和学生维度 $K \in [32, 1024]$。

理论验证（无叠加）。对于基准情形（$K = N$），我们确认了理论推导。如图3b所示，经验指数与预测的 $\alpha = (a + 2b - 1)/a$ 完全吻合。

叠加机制。当学生模型被迫进入叠加机制（$K < N$）时，动力学会发生显著转变。图 5总结了拟合指数。请注意，只有在 $a + 2b > 1$ 时才需要幂律理论近似。我们可以针对任意a，$b \geq 0$开展实验。

无论 a、b的具体取值，还是压缩比 K/N，训练中期幂律指数都会收敛至 $\boldsymbol{\alpha \approx 1}$。图 6展示了用于拟合训练指数的损失示例。

与序贯情形相比，这体现了普适加速。例如，当 $a = 1.1$和 $b = 0$时，序贯理论预测 $\alpha \approx 0.09$会缓慢衰减。叠加会使这一过程加速超过 $\mathbf{10}\times$。这种普适性表明，嵌入层中固有的随机性充当了一种机制，使各特征的有效学习率趋于一致，从而使优化动力学与数据的谱衰减解耦。

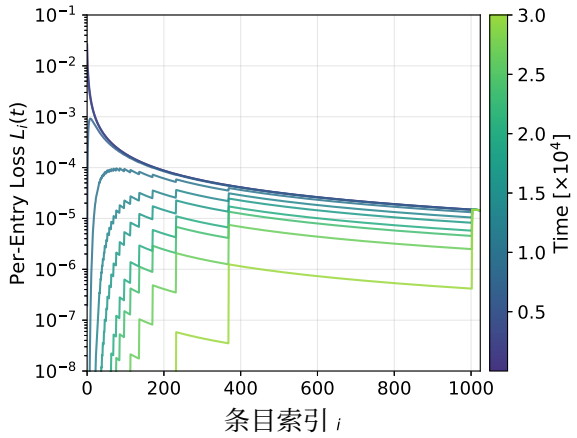
在附录 A.2中，我们分析了叠加模型在线性区间中的两种情形：最大正干扰和各向同性混合的极限。这些情形为实证结果中观察到的加速和均匀性提供了理论直觉。不过，我们强调，它们并不能解释 $\alpha \approx 1$ 指数，因为它只有在采用 ReLU 非线性和偏置时才会出现。

4.3. 最优计算量缩放

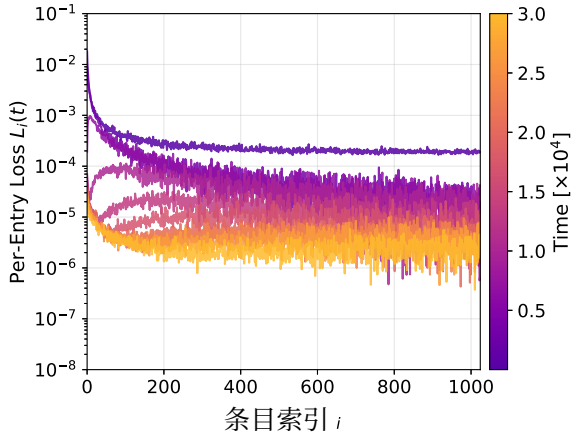
无论学生规模如何，叠加都能普遍加快中期训练动态，但模型的最终可达性能受瓶颈维度 K 的限制。为了理解模型规模与训练步数之间的权衡，我们分析了最优计算前沿。

在图 7a中，我们绘制了不同规模学生的损失曲线（$K \in [32, 1024]$），并以总计算量为横轴（近似为 $C = t \times K^2$）。我们观察到一条清晰的**最优包络**，以黑线标示，由任意给定计算预算下可达到的最低损失构成。与大规模 Transformer 中的观察结果一致，最优模型规模会随计算预算增加。如图14所示，K 大约会翻倍，而 C 增加十倍。我们进行了初步的幂律拟合，得到相对于 C 的规模缩放指数约为0.2702。

我们进一步使用等计算曲线（图 7b）量化这种关系。这些曲线记录了在固定计算量水平下，损失如何随学生规模变化。这些曲线表明，在固定计算预算下，最小损失与学生规模成幂律反比关系。我们还在附录 $c_k K^{-\beta}$中讨论了学生规模对损失缩放的实验结果（即宽度受限损失项 A.4）。这证实，尽管我们的玩具模型十分简单，却能重现生产级大语言模型中观察到的宏观缩放行为，同时还提供了一个便于研究底层机制的可处理试验平台。



*(a)*无叠加：顺序“行波”



*(b)*叠加：并行“全局衰减”

图 *8.* **叠加下普适损失动态的机制。** 我们可视化了逐特征损失 $L_i(t)$ 随时间的变化。**(a)**在无叠加时，模型按顺序学习特征，形成行波前沿，尾部特征在很长时间内仍未被学习。**(b)** 在叠加下，特征混合使有效梯度均衡，从而使所有特征——无论重要性如何——都被并行学习。

5. 讨论

我们的工作的核心发现是，叠加能够普适地加速训练动态。本节将探究驱动这一现象的微观机制。具体而言，我们将探讨两个相互交织的问题：（1）为什么指数在中期训练期间会增大；（2）为什么这种行为在不同数据分布中具有普适性?

答案在于模型如何处理输入数据和通道统计。我们对比基准情形的序列性与叠加所引发的并行性。

5.1. 机制

序列学习如同行波。 在没有叠加的情况下，输入概率 p_i 和通道重要性 A_{ii} 按设计严格遵循幂律衰减。这种结构迫使梯度下降动力学遵循有序序列：模型按重要性从高到低学习特征。

图 8a展示了这一现象。图中绘制了不存在叠加时各个条目随时间变化的损失贡献。这种动力学类似于“行进的波前”。索引较低的特征（重要性/频率较高）首先被学习，而索引较高的特征（重要性/频率较低）实际上保持冻结状态——在此前更具主导性的特征得到解决之前，它们的损失不会下降。因此，在任何时刻 t 的中期训练阶段，总损失都由分布尾部中数量庞大的未学习特征所累积的误差主导。因此，损失收敛的速率完全由数据的谱衰减决定。

叠加打乱了这一顺序。 叠加通过随机嵌入$\mathbf{W}$打乱了这一序列。压缩将来自不同位置的特征混合在一起，而逐特征归一化则确保它们以相近的幅度进行投影。这带来了两个根本性的变化：

1. **特征混合：** 不同的矩阵项被混合到潜在空间中，使学生无法将它们分离开来并按顺序学习。

2. **频率均衡：** 平均而言，学生矩阵 $\mathbf{B}$ 的每个维度都会接收到聚合信号，其中高频和低频特征兼而有之。

因此，学生以相近的速率同时学习所有矩阵项。图 8b展示了这一点。与没有叠加时的“行波”不同，叠加机制中的逐项损失始终聚集在相近水平，并同步衰减。“未学习的尾部特征”这一概念实际上消失了。

训练中期加速。 为什么叠加带来的这种均匀性会促成加速? 我们强调，这一优势仅存在于中期训练阶段。由于叠加学生（$K < N$）的容量严格低于教师（N），它最终必然会达到一个非零损失平台，而理想的无叠加学生（$K = N$）则会收敛至零。

然而，在达到这一饱和点之前，叠加展现出优势。如图4所示，叠加损失曲线在中期训练期间显著低于非叠加损失曲线。从数学上讲，这

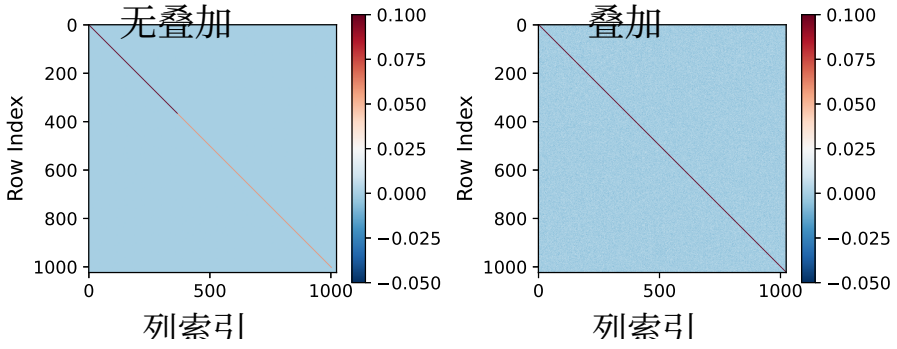


图 *9.* **学生模型学习模式的可视化：依次学习（无叠加）与并行学习（叠加）**。 学生矩阵$\mathbf{B}$ 在 $t = 10^3$时的快照。**左图：**没有叠加时，学生模型严格沿对角线学习，依次解决各项特征（仅学习前 ~ 400 项）。**右图：**采用叠加的学生模型将权重分布在整个矩阵中，并行学习全部 1024 项特征。

之所以发生，是因为顺序模型受到幂律重尾的惩罚——它对已学习特征的误差为零，但对未学习的大多数特征的误差最大。相比之下，叠加模型将误差预算均匀分配。在中期训练阶段，许多较小的平均误差之和（叠加）低于未学习谱尾之和（顺序）。随机性实际上起到了均衡收敛速率的作用，使模型能够绕过对谱的缓慢顺序遍历。

学生结构的可视化。我们在图中直观地展示了这种结构转变。9图中呈现了学生矩阵 $\mathbf{B}$ 在时间 $t = 10^3$（中期训练）的快照。由于教师矩阵是对角的，理想的学生矩阵应恢复对角元素。

没有叠加时，学生会严格沿着对角线逐项学习。在 $t = 10^3$时，它已成功学会 1024 个特征中的 ~ 400，其余特征仍保持初始化值。有了叠加后，学生矩阵会被全面激活。学到的权重分布在整个模型中，用于解码混合信号。模型实际上能够并行学会全部 1024 个条目，利用叠加中的干扰，比序列结构所允许的速度更快地降低全局错误率。

6. 结论与展望

在这项工作中，我们确立了特征叠加与普适幂律训练动力学之间的联系。通过分析一个模仿大语言模型结构的教师—学生框架，我们识别出两种截然不同的学习模式。在没有叠加的情况下，训练由谱滤波过程主导，损失的收敛取决于按频率递减顺序对特征进行序列学习。

相比之下，我们发现，引入叠加会普遍改变动力学，无论输入数据和通道分布如何。叠加固有的随机性统一了不同特征的有效收敛速率，将依赖数据的缩放定律归并为一个**具有普适幂律指数** $\alpha \approx 1$的定律。

这一发现为模型宽度在训练动态中的作用提供了一个引人深思的视角。窄瓶颈并非仅仅是一种容量约束；当它与高维特征稀疏性相结合时，还能使学习动态趋于均衡。它以牺牲正交表示的精确性为代价，换取并行化误差缩减的速度，从而可能实现加速，这对高效训练大规模模型至关重要。

展望。 我们的发现为未来研究开辟了若干方向。首先，尽管我们的玩具模型捕捉到了叠加的线性和前馈特性，但将这一分析扩展到 **采用注意力机制的多层架构** 至关重要。我们推测，注意力头的混合作用可能会进一步增强我们观察到的均匀性。其次，**该** $1/t$ **缩放定律**的起源仍大多不为人所知。虽然附录A.2 中的线性理论为收敛速率的均匀性提供了直观解释，但对于非线性叠加模型而言，指数为何恰好稳定在 $\alpha \approx 1$——而不是其他数值——仍是一个尚待解决的难题。第三，我们关注的是优化误差占主导地位的中期训练阶段。对**后期训练动态**进行更细致的分析——此时模型达到瓶颈所导致的不可约近似误差——或许能帮助我们理解算法任务中观察到的“顿悟”现象。最后，我们的结果表明，**随机性普遍有益**于训练中期动力学。结构化或学习得到的嵌入能否在中期或后期训练阶段胜过随机投影，仍是优化高效模型时有待解答的问题。我们在附录A.3中介绍了对此问题的一些初步研究。

致谢

本研究获得了 SchmidtPolymath 奖的资助。ZJC 感谢库尔特·福里斯特基金会奖学金和亨利·肯德尔奖学金的支持。作者感谢麻省理工学院研究计算与数据办公室提供 Engaging 集群的使用权限。我们也感谢使用普林斯顿大学的 Della 集群；该集群由普林斯顿计算科学与工程研究所（PICSciE）和信息技术办公室的研究计算部门管理。

作者感谢美国国家科学基金会通过合作协议 PHY-2019786（美国国家科学基金会人工智能研究所：人工智能与基础交互）提供的支持。本研究由美国空军研究实验室和空军部人工智能加速器赞助，并依据合作协议编号 FA8750-19-2-1000 完成。本文中的部分计算在哈佛大学 FAS 科学部研究计算组支持的 FASRC 集群上运行。本研究使用了 DeltaAI 高级计算与数据资源。该资源由美国国家科学基金会（资助编号 OAC 2320345）和伊利诺伊州提供支持；相关计算资源通过高级网络基础设施协调生态系统：服务与支持（ACCESS）计划分配（分配编号 CIS240904）。该计划由美国国家科学基金会资助，资助编号为 # 2138259、#2138286、#2138307、#2137603 和 # 2138296；此外，本研究还通过国家人工智能研究资源（NAIRR）试点项目 NAIRR250043 获得支持。

影响声明

本文对叠加条件下神经网络的训练动力学进行了理论研究。作为基础研究，其主要影响在于增进科学界对过参数化模型和压缩模型（如大语言模型）为何能够高效扩展的理解。

能源与效率。 通过发现叠加会引发一种普适且加速的学习轨迹（t^{-1}），我们的工作为设计计算高效架构提供了理论依据。认识到压缩瓶颈能够加速中期训练的收敛，而非阻碍收敛，或许会推动开发更窄、更深的模型，以最大限度地发挥这种“随机性优势”，并有望降低基础模型预训练所产生的碳足迹和能源成本。

可解释性与可靠性。 我们的工作弥合了标度律（宏观行为）与机制可解释性（微观结构）之间的鸿沟。通过阐明干扰噪声如何主动塑造训练轨迹，我们有助于更好地理解黑盒模型的内部表征。这一理论基础对于开发更可靠的可解释性工具至关重要，而这些工具是人工智能系统在关键领域安全部署所必不可少的。

预计这项工作不会带来直接的负面社会影响，不过训练效率的提升自然会加快人工智能系统整体能力的发展。

参考文献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2024年。Advani, M. S., Saxe, A. M., 和 Sompolinsky, H. 神经网络中泛化误差的高维动力学。神经网络，132:428–446，2020年。ISSN 0893-6080。Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., 和 Wang, R. 关于无限宽神经网络的精确计算。Curran Associates Inc.，美国纽约州雷德胡克，2019。Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., 和 Ganguli, S. 深度学习的统计力学。年度凝聚态物理评论，11:501–528，2020年。Bordelon, B., Canatar, A., 和 Pehlevan, C. 核回归与宽神经网络中依赖于谱的学习曲线。载于第*37*届国际机器学习大会论文集，ICML’20。JMLR.org，2020年。Bordelon, B., Atanasov, A., 和 Pehlevan, C. 特征学习如何改进神经网络缩放定律。载于第十三届国际学习表征大会，2025a。Bordelon, B., Letey, M. I., 和 Pehlevan, C. 上下文内回归缩放定律理论：深度、宽度、上下文与时间。*arXiv* 预印本 *arXiv:2510.01098*，2025b。Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., 等。迈向单义性：通过字典学习分解语言模型。*Transformer Circuits* 讨论串，2023年。

Canatar, A.、Bordelon, B. 和 Pehlevan, C. 谱偏差与任务-模型对齐解释了核回归和无限宽神经网络中的泛化。《自然·通讯》，12(1)，2021年5月。

Chizat, L.、Oyallon, E. 和 Bach, F. 关于可微编程中的惰性训练。Curran Associates Inc.，美国纽约州雷德胡克，2019。

Clark, A.、De Las Casas, D.、Guy, A.、Mensch, A.、Paganini, M.、Hoffmann, J.、Damoc, B.、Hechtman, B.、Cai, T. 等。路由语言模型的统一缩放定律。载于第*39*届国际机器学习会议论文集，PMLR 第162卷机器学习研究论文集，第4057–4086页。PMLR，2022年7月17日至23日。

Cui, H.、Loureiro, B.、Krzakala, F. 和 Zdeborov´a, L. 核回归中的泛化误差率：从无噪声区间到有噪声区间的交叉。统计力学杂志：理论与实验，2022年(11):114004，2022年11月。

Elhage, N.、Hume, T.、Olsson, C.、Schiefer, N.、Henighan, T.、Kravec, S.、Hatfield-Dodds, Z.、Lasenby, R.、Drain, D.、Chen, C.、Grosse, R.、McCandlish, S.、Kaplan, J.、Amodei, D.、Wattenberg, M. 和 Olah, C. 叠加玩具模型。*arXiv* 预印本 *arXiv:2209.10652*，2022年。

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J. 和 Wu, J. 稀疏自编码器的缩放与评估。*arXiv* 预印本 *arXiv:2406.04093*，2024年。

Gu, A. 和 Dao, T. Mamba：使用选择性状态空间的线性时间序列建模。发表于第一届语言建模会议，2024年。

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D. 和 Mc-Candlish, S. 自回归生成建模的标度律。*arXiv* 预印本 *arXiv:2010.14701*，2020年。

Henighan, T., Carter, S., Hume, T., Elhage, N., Lasenby, R., Fort, S., Schiefer, N. 和 Olah, C. 叠加、记忆化与双重下降。*Transformer CircuitsThread*，2023年。

Hernandez, D., Kaplan, J., Henighan, T. 和 McCandlish, S. 迁移缩放定律。*arXiv* 预印本 *arXiv:2102.01293*，2021年。

Hoffmann, J.、Borgeaud, S.、Mensch, A.、Buchatskaya, E.、Cai, T.、Rutherford, E.、De Las Casas, D.、Hendricks, L. A.、Welbl, J.、Clark, A.、Hennigan, T.、Noland, E.、Millican, K.、van den Driessche, G.、Damoc, B.、Guy, A.、Osindero, S.、Simonyan, K.、Elsen, E.、Vinyals, O.、Rae, J. W. 和 Sifre, L. 训练计算量最优的大型语言模型。在第*36*届神经信息处理系统国际会议论文集中，NIPS '22，美国纽约州雷德胡克，2022年。Curran Associates Inc. 国际标准书号 9781713871088。

Huben, R., Cunningham, H., Smith, L. R., Ewart, A. 和 Sharkey, L. 稀疏自编码器在语言模型中发现高度可解释的特征。载于第十二届国际学习表征会议，2024年。

Jacot, A., Gabriel, F. 和 Hongler, C. 神经切线核：神经网络中的收敛与泛化。载于第*32*届神经信息处理系统国际会议论文集，NIPS'18，第8580–8589页，美国纽约州雷德胡克，2018年。Curran Associates Inc.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. 和 Amodei, D. 神经语言模型缩放定律。*arXiv*预印本 *arXiv:2001.08361*，2020年。

Liu, Y., Liu, Z. 和 Gore, J. 叠加带来稳健的神经网络缩放。发表于第三十九届神经信息处理系统年会，2025年。

Maloney, A., Roberts, D. A. 和 Sully, J. 神经网络缩放定律的一个可解模型。*arXiv*预印本 *arXiv:2210.16859*，2022年。

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L. 等。Gemma：基于 Gemini 研究与技术的开放模型。*arXiv*预印本 *arXiv:2403.08295*，2024年。

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. 通过机制可解释性研究顿悟的进展度量。载于第十一届学习表征国际会议，2023年。

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. 深入了解：电路简介。*Distill*，2020年。

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. 顿悟：在小型算法数据集上超越过拟合，实现泛化。*arXiv*预印本 *arXiv:2201.02177*，2022年。

Rahaman, N., Baratin, A., Arpit,D., Draxler,F.,Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. 关于神经网络的谱偏差。*arXiv*预印本 *arXiv:1806.08734*，2019。

Saxe, A. M.、McClelland, J. L. 和 Ganguli, S. 深度线性神经网络学习非线性动力学的精确解。*arXiv* 预印本 *arXiv:1312.6120*，2013。

Saxe, A. M.、McClelland, J. L. 和 Ganguli, S. 深度神经网络中语义发展的数学理论。美国国家科学院院刊，116(23)：11537–11546，2019年5月。ISSN 1091-6490。

Sharma, U. 和Kaplan, J. 基于数据流形维度的神经网络缩放定律。*arXiv* 预印本 *arXiv:2004.10802*，2020年。

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S. 和 Mor- cos, A. S. 超越神经网络缩放定律：通过数据剪枝击败幂律缩放。载于第*36*届神经信息处理系统国际会议论文集，NIPS '22，美国纽约州雷德胡克，2022年。Curran Associates Inc. 国际标准书号 9781713871088。

Templeton, A., Conerly, T., Marcus, J., Lindsay, J., Bricken, T., Chen, B., Pearce, A. 等。单义性缩放：从 Claude 3 Sonnet 中提取可解释特征。*Transformer CircuitsThread*，2024年。Wei, C., Lee, J. D., Liu, Q., 和 Ma, T. 正则化至关重要：神经网络相对于其诱导核的泛化与优化。Curran Associates Inc.，美国纽约州雷德胡克，2019。Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Met- zler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., 和 Fedus, W. 大语言模型的涌现能力。机器学习研究汇刊，2022年。ISSN 2835-8856。网址 https://openreview.net/forum?id=yzkSU5zdwD。Survey Certification。Yang, G. 和 Hu, E. J. 无限宽神经网络中的特征学习。*arXiv*预印本 *arXiv:2011.14522*，2022 年。Zhai, X., Kolesnikov, A., Houlsby, N., 和 Beyer, L. 视觉Transformer的缩放。在*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第12104–12113页，2022年6月。

A. 附录

A.1. 一般分布的幂律理论

本节将详细推导一般输入分布下的训练动力学指数。我们将讨论范围扩展到正文分析的幂律情形之外，并纳入指数衰减和代数衰减。我们在图10的相图中总结了这些情形。

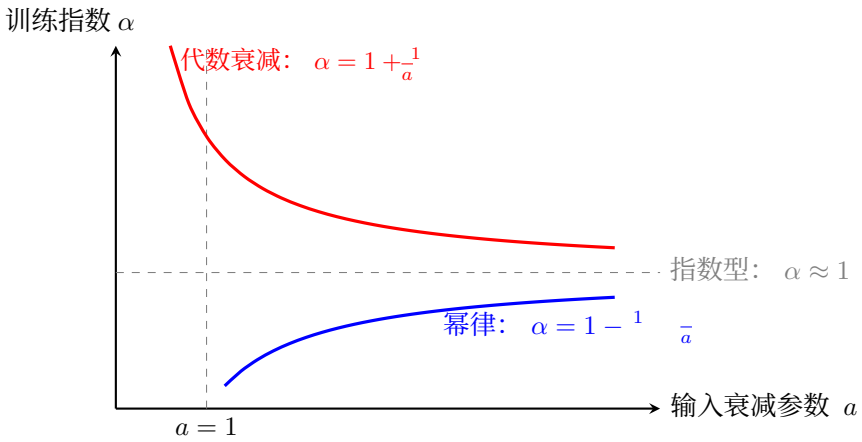


图 *10.* **训练动态的相图。** 绘制训练幂律指数 α 与输入分布衰减参数 a的关系。**红色：** 在有限边界处呈代数衰减的输入收敛最快（$\alpha > 1$）。**蓝色：** 具有重幂律尾部的输入收敛最慢（$\alpha < 1$）。**虚线：** 指数衰减输入（以及叠加模型）以 $\alpha \approx 1$收敛。

A.1.1. 连续近似

假设通道重要性近似均匀（在极限 $b \to 0$下），当 $N \gg 1$趋于极限时，损失函数可近似为连续积分。令 i 为离散特征索引， z 为其连续对应量。损失为：

$$\mathcal{L}(t) \propto \sum_{i=1}^{N} f(i) e^{-f(i)t/\tau} \approx \int_0^\infty f(z) e^{-f(z)t/\tau} dz, \tag{10}$$

其中， $f(z)$ 是单调递减的数据频率。我们定义一个**临界特征索引** $z_c(t)$，使得该指数为单位量级：

$$f(z_c) \cdot \frac{t}{\tau} = 1 \implies z_c(t) = f^{-1}\left(\frac{\tau}{t}\right). \tag{11}$$

在时间 t，具有 $z < z_c$ 的特征已被学习（$e^{-f(z)t/\tau} \approx 0$），而具有 $z > z_c$ 的特征则会对损失产生贡献（$e^{-f(z)t/\tau} \approx 1$）。损失由尾积分主导：

$$\mathcal{L}(t) \sim \int_{z_c(t)}^\infty f(z) dz. \tag{12}$$

请注意，要使这一近似成立， $1 \ll z_c \ll N$，这对应于动力学中的训练中期。

A.1.2. 幂律衰减

令 $f(z) = Cz^{-a}$，其中 $a > 1$。临界指标的标度关系为：

$$z_c^{-a} \sim \frac{\tau}{t} \implies z_c(t) \sim \left(\frac{t}{\tau}\right)^{1/a}. \tag{13}$$

损失积分变为：

$$\mathcal{L}(t) \sim \int_{z_c}^\infty z^{-a} dz = \left[\frac{z^{1-a}}{1-a}\right]_{z_c}^\infty \propto z_c^{1-a}. \tag{14}$$

代入 $z_c(t)$:

$$\mathcal{L}(t) \propto \left(t^{1/a}\right)^{1-a} = t^{-(1-1/a)}. \tag{15}$$

请注意，对于具有 $b > 0$的非平凡通道衰减，我们可以代入 $z^{-a} \to z^{-(a+2b)}$。临界特征索引 z_c（t）仅取决于输入数据，并保持不变。在此代换下，我们恢复了式8所示的损失幂律。

A.1.3. 指数衰减

设 $f(z) = Ce^{-\kappa z^a}$，且 κ，$a > 0$。我们进行变量代换 $u = f(z)$。则 $z = f^{-1}(u) = \kappa^{-1/a}(\log(C/u))^{1/a}$。微分为:

$$dz = \frac{d}{du}\left(\kappa^{-1/a}(\log(C/u))^{1/a}\right) du \sim -\frac{1}{au}(\log(C/u))^{1/a-1}du. \tag{16}$$

将其代入积分 $\int f(z)e^{-f(z)t/\tau}dz$:

$$\mathcal{L}(t) \sim \int_0^C ue^{-(t/\tau)u} \cdot \frac{1}{u}\left(\log\frac{C}{u}\right)^{1/a-1} du. \tag{17}$$

我们对 $s = (t/\tau)u$进行重标度，因此 $du = (\tau/t)ds$。在极限 $t \gg 1$下，该积分由较小的 u （有限的 s）主导，因此 $\log(C/u) = \log(Ct/\tau s) \approx \log(t)$。将对数项提出:

$$\mathcal{L}(t) \sim (\log t)^{1/a-1} \int_0^\infty e^{-s}ds \cdot \frac{\tau}{t}. \tag{18}$$

因此，缩放关系为:

$$\mathcal{L}(t) \propto t^{-1}(\log t)^{\frac{1}{a}-1}. \tag{19}$$

主导项为 t^{-1}，表示轻尾分布的衰减率。

A.1.4. 代数衰减

设 $f(z) \sim (z_* - z)^a$ 位于有限边界 z_*附近。 z 从左侧趋近于 z_*： $z \to z_*^-$。临界指数条件（$z_* - z_c$）$^a \sim \tau/t$ 意味着间隙为 $\Delta z = z_* - z_c \sim t^{-1/a}$。损失积分在间隙 Δz 上进行:

$$\mathcal{L}(t) \sim \int_{z_c}^{z_*} (z_* - z)^a dz \propto (z_* - z_c)^{a+1} \sim (t^{-1/a})^{a+1} = t^{-(1+1/a)}. \tag{20}$$

A.1.5. 推导总结

1. Po 幂律衰减：对于 $f(z) \sim z^{-a}$ （其中 $a > 1$），数据的重尾会减缓学习 g.

$$\mathcal{L}(t) \sim t^{-(1-1/a)}. \tag{21}$$

如图 10 （蓝色曲线）所示， α 只有在 $a \to \infty$ 时才趋近于 1（陡峭衰减）。

2. 指数衰减：对于 $f(z) \sim e^{-\kappa z^a}$，分布衰减得足够快，因此动力学由 $\mathcal{O}(1)$ 时间常数主导。

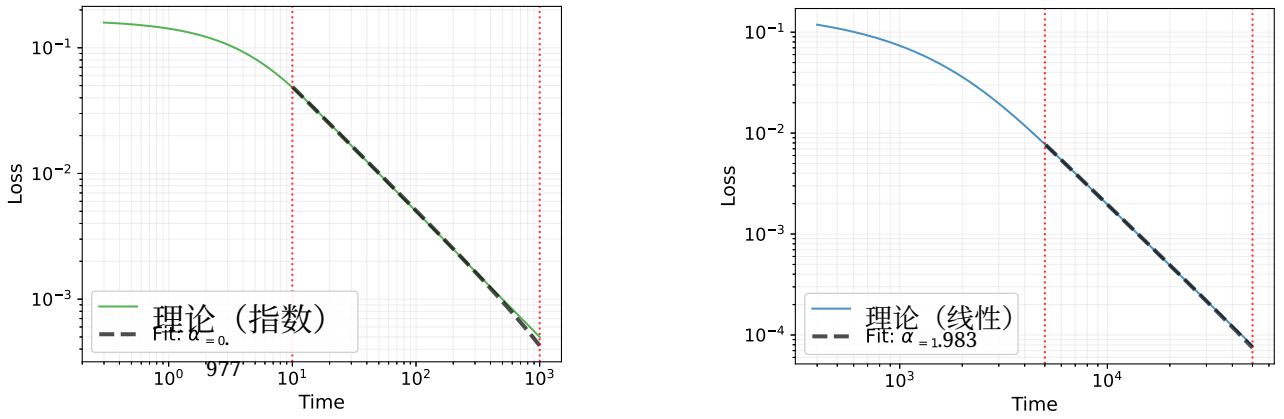
$$\mathcal{L}(t) \sim t^{-1}(\log t)^{\delta} \implies \alpha \approx 1. \tag{22}$$

这代表了我们相图中的分界线。

3. 代数衰减：对于支撑集在 z_* 处截止的有限支撑分布，按 $(z_* - z)^a$ 衰减时，“困难”样本的稀缺性会消失。

$$\mathcal{L}(t) \sim t^{-(1+1/a)}. \tag{23}$$

这带来了快速收敛（$\alpha > 1$），如图中红色所示。10。



*(a)*指数衰减（$p_i \propto e^{-0.5i}$）　　*(b)* 线性代数衰减（$p_i \propto N - i$）

图*11*。 **一般分布的标度律验证**。 实线曲线表示精确的离散理论损失，虚线表示幂律拟合。 **（a）** 指数输入分布得到 $\alpha \approx 1$。 **（b）** 线性代数衰减输入分布得到 $\alpha \approx 2$。二者都与我们的连续近似理论预测相符。

A.1.6. 幂律验证

我们针对非幂律分布，使用理论损失（基于公式 7 的修改）验证幂律近似。具体而言，我们测试了指数衰减和线性代数衰减。我们定义归一化特征频率 $f(i) = p_i$，并将其归一化，使得 $\sum_i p_i = 1$。

- 指数衰减： $p_i \propto e^{-0.5i}$。这里， $a = 1$ 属于指数衰减类，幂律近似预测$\mathcal{L}(t) \propto t^{-1}$，即 $\alpha = 1$。
- 线性代数衰减： $p_i \propto N - i$。这里， $a = 1$ 属于代数衰减类，幂律近似预测 $\mathcal{L} \propto t^{-(1+1/1)} = t^{-2}$，即 $\alpha = 2$。

在图 11中，我们对精确的离散求和理论损失进行幂律训练指数拟合。对于指数情形（图11a），我们得到拟合指数 $\alpha = 0.9769 \pm 0.0012$，与近似预测值 $\alpha = 1$高度一致。对于线性衰减情形（图 11b），我们得到 $\alpha = 1.9834 \pm 0.0007$，与理论预测值$\alpha = 2$高度吻合。这些结果证实，我们的连续积分近似（第 A.1节）能够准确捕捉不同普适类中的离散训练动力学。

A.2. 线性叠加模型的理论损失

我们推导了线性区间内叠加模型的解析损失动力学。通过考察两种极限情形，我们分析了嵌入矩阵$\mathbf{W}$ 的结构如何影响有效学习动力学：**(1)** $\mathbf{W}$ 是一个使正干扰最大化的设计簇；**(2)** $\mathbf{W}$ 由 $N \gg K$各向同性极限下的随机向量组成。具体而言，极限情形 **(1)** 阐明了加速机制，而极限情形 **(2)** 解释了普适性的出现。正文中介绍的经验叠加实验位于这两个极端之间，同时兼具相长干涉的速度和各向同性混合的均匀性。

A.2.1. 一般设置

我们考虑一个线性学生 $\mathbf{y} = \mathbf{W}^\top \mathbf{B}\mathbf{W}\mathbf{x}$，它试图模仿一个线性教师 $\mathbf{y}^* = \mathbf{A}\mathbf{x}$。损失为：

$$\mathcal{L} = \frac{1}{2}\mathrm{Tr}\left[(\mathbf{W}^\top \mathbf{B}\mathbf{W} - \mathbf{A})^\top (\mathbf{W}^\top \mathbf{B}\mathbf{W} - \mathbf{A})\mathbf{\Sigma}\right], \tag{24}$$

其中 $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \approx \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ 是数据协方差，主要由对角方差构成。定义投影协方差 $\mathbf{C} \equiv \mathbf{W}\mathbf{\Sigma}\mathbf{W}^\top$ 和相关矩阵 $\mathbf{P} \equiv \mathbf{W}\mathbf{W}^\top$后，损失可简化为：

$$\mathcal{L} = \frac{1}{2}\left[\mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}\mathbf{A}) + \mathrm{Tr}(\mathbf{B}^\top \mathbf{P}\mathbf{B}\mathbf{C}) - 2\mathrm{Tr}(\mathbf{S}\mathbf{B})\right], \tag{25}$$

其中 $\mathbf{S} = \mathbf{W\Sigma A W}^\top$ 表示投影后的教师信号。学生矩阵 $\mathbf{B}$ 的梯度流动力学为：

$$\frac{d\mathbf{B}}{dt} = -\eta \frac{\partial \mathcal{L}}{\partial \mathbf{B}} = -\eta(\mathbf{PBC} + \mathbf{CBP} - \mathbf{S} - \mathbf{S}^\top)/2. \tag{26}$$

对于对称矩阵，我们分析主导阶动力学：$\mathbf{B} \approx -\eta(\mathbf{PBC} - \mathbf{S})$。

A.2.2. 案例 1：聚类叠加

为分析加速机制，我们构造一种结构化嵌入。我们将 N 特征划分为 K 簇$\{G_1, \ldots, G_K\}$。我们假设采用理想化的“硬”分配，即每个特征 $i \in G_j$ 都被完美映射到维度 j：

$$W_{ji} = \begin{cases} 1 & \text{if } i \in G_j \\ 0 & \text{otherwise} \end{cases}. \tag{27}$$

在此极限下，矩阵$\mathbf{P}$和$\mathbf{C}$变为对角矩阵：

$$\mathbf{P} = \text{diag}(n_1, \ldots, n_K), \quad \mathbf{C} = \text{diag}\left(\sum_{i \in G_1} \sigma_i^2, \ldots, \sum_{i \in G_K} \sigma_i^2 \right), \tag{28}$$

其中，$n_j = |G_j|$ 为簇大小。教师投影$\mathbf{S}$ 也是对角矩阵，其对角元素为 $S_{jj} = \sum_{i \in G_j} A_{ii}\sigma_i^2$。在这一对称的对角极限下，ODE 中$\mathbf{B}$的对角元素 b_j 彼此解耦：

$$\dot{b}_j \approx -\eta(n_j C_{jj} b_j - S_{jj}). \tag{29}$$

该线性常微分方程的解为：

$$b_j(t) = \frac{S_{jj}}{n_j C_{jj}} \left(1 - e^{-\eta(n_j C_{jj})t}\right). \tag{30}$$

该解描述了指数弛豫。关键在于，簇 j 的有效收敛速率为 $\lambda_{eff} \propto n_j C_{jj} = n_j \sum_{i \in G_j} \sigma_i^2$。这明确表明，通过对簇中所有特征的方差（重要性）求和，聚类能够加速学习。

解释： 此情形代表叠加的一种极端极限，其特征是最大正干扰。这类似于 Elhage 等人（2022年）讨论的“特权基”。通过将特征完美对齐到同一子空间中，模型实现了两种乘法式加速效应：(i) 信号能量的求和（$\sum \sigma^2$）和 (ii) 来自 n_j 个特征的梯度累积。这带来了二次加速尺度 $\sim n_j^2 \bar{\sigma}^2$，代表训练速度的理论上限，但代价是较高的碰撞误差。

A.2.3. 案例 2：各向同性极限

现在我们考虑各向同性极限，其中 $\mathbf{W} \in \mathbb{R}^{K \times N}$ 是一个满足 $N \gg K$的随机矩阵。我们假设嵌入经过列归一化，使得 N 特征向量的长度为 1 $\|\mathbf{w}_i\|^2 = 1$。这对应于元素 $W_{ji} \sim \mathcal{N}(0, 1/K)$。由于随机向量的各向同性，相关矩阵 $\mathbf{P}$ 变为由框架势确定的缩放单位矩阵：

$$\mathbb{E}[\mathbf{W}\mathbf{W}^\top] = \sum_{i=1}^{N} \mathbb{E}[\mathbf{w}_i \mathbf{w}_i^\top] = \frac{N}{K}\mathbf{I}_K. \tag{31}$$

接下来，我们考虑投影协方差 $\mathbf{C} = \mathbf{W\Sigma W}^\top$。对于 $\mathbf{\Sigma} = \text{diag}(\sigma_i^2)$，该矩阵对按压缩比缩放的输入方差取平均：

$$\mathbb{E}[\mathbf{C}] = \sum_{i=1}^{N} \sigma_i^2 \mathbb{E}[\mathbf{w}_i \mathbf{w}_i^\top] = \frac{1}{K}\left(\sum_{i=1}^{N} \sigma_i^2 \right) \mathbf{I} = \frac{N}{K}\bar{\sigma}^2 \mathbf{I}. \tag{32}$$

将这些代入动力学方程（式 26）：

$$\frac{d\mathbf{B}}{dt} \approx -\eta \left[\left(\frac{N}{K}\mathbf{I}\right) \mathbf{B} \left(\frac{N}{K}\bar{\sigma}^2 \mathbf{I}\right) - \mathbf{S} \right] = -\eta \left(\frac{N}{K}\right)^2 \bar{\sigma}^2 \mathbf{B} + \eta \mathbf{S}. \tag{33}$$

解释： 这代表叠加的最大均匀/各向同性极限。这个结果解释了普适性。在非叠加情形中，每个特征 i 都以自身的速率 $\lambda_i \propto i^{-\alpha}$学习，导致尾部出现瓶颈。采用随机叠加时，随机性相当于对所有特征进行“均匀聚类”。在统计意义上，它实现了与均匀聚类情形相同的二次标度（N/K）^{2}。这种谱统一确保学生矩阵 $\mathbf{B}$ 以单一的有效速率均匀收敛，从而消除谱瓶颈并导出普适指数。

A.3. 随机性对于中期训练是否接近最优?

在我们的主要分析中，我们对嵌入层 $\mathbf{W}$ 和 $\mathbf{W}^\top$采用了固定的随机投影。一个自然的问题是：在中期训练中观察到的普适加速究竟是固定随机性的产物，还是在嵌入经过优化后仍然存在?

为了解决这一问题，我们修改了玩具模型设置，使编码器$\mathbf{W}$ 和解码器 $\mathbf{W}^\top$ 都成为完全可学习的参数。这使参数数量从 K^2 （仅学生）增加到 $K^2 + 2NK$。我们在输入衰减 $a = 1.1$、通道重要性 $b = 0.0$和特征维度 $N = 1024$的设置下重复叠加实验，并遍历学生规模 $K \in \{128, 256, 512\}$。我们在与固定嵌入实验相同的中期训练窗口（$t \sim 10^3$ 至 10^4）内拟合幂律训练指数。

结果。 如图 12a所示，我们发现幂律训练动力学对这一架构变化仍然稳健。可学习情形的拟合指数为：

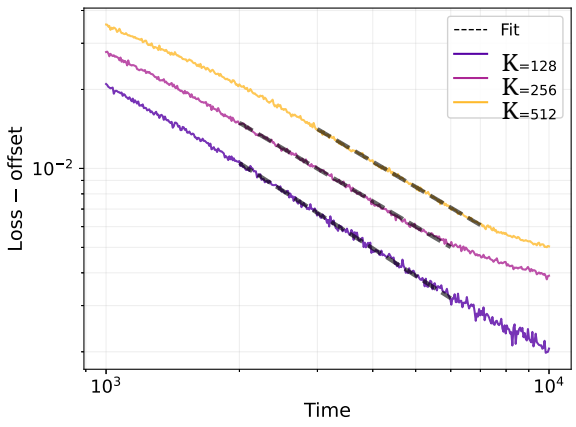
$$\alpha_{128} = 1.0812 \pm 0.0215, \quad \alpha_{256} = 0.9933 \pm 0.0149, \quad \alpha_{512} = 0.9946 \pm 0.0206.$$

这些数值与固定随机嵌入下观察到的 $\alpha \approx 1$ 区间无法区分。这表明，对于训练中期动力学，随机初始化所提供的“混合”足以产生普适加速效应；对基方向进行显式梯度优化并不会进一步加速损失收敛的速率（即幂 =律训练指数）。

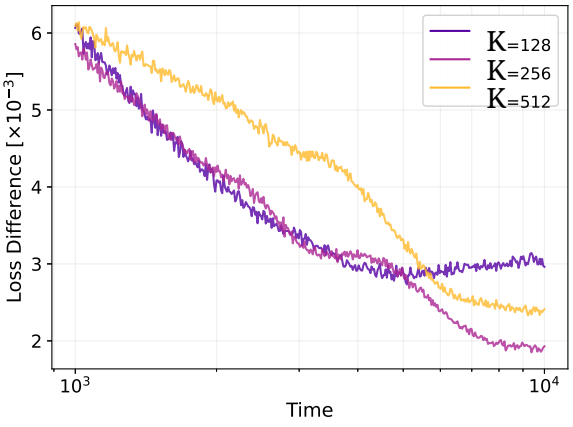
性能差距。尽管指数保持不变，可学习嵌入确实带来了性能优势。在图12b中，我们绘制了损失差异 $\Delta\mathcal{L} = \mathcal{L}_{\text{fixed}} - \mathcal{L}_{\text{learnable}}$。正差异表明，可学习模型的绝对损失更低。不过，这一改进的幅度为 $\mathcal{O}(10^{-3})$，属于次要效应

与绝对损失幅度 $\mathcal{O}(10^{-2})$相比。

我们得出结论：对于中期训练缩放定律而言，随机性接近最优。不过需要强调的是，这一结论仅适用于中期训练阶段；在无限训练时间下，可学习模型凭借更大的容量以及使瓶颈与数据的主成分对齐的能力，自然会收敛到更低的最终损失平台。



(a) 损失动态（可学习$\mathbf{W}$）



(b) 损失差异（$\mathcal{L}_{\text{fixed}} - \mathcal{L}_{\text{learnable}}$）

图*12*。**固定嵌入与可学习嵌入的比较。(a)** 可学习$\mathbf{W}$ 模型的训练曲线显示出与固定随机模型相同的 $\alpha \approx 1$ 幂律衰减。**(b)** 固定模型与可学习模型之间的损失差异。尽管可学习嵌入的损失略低（差异为正），但改进幅度比总损失小一个数量级，这表明随机性是训练指数的主要驱动因素。

A.4. 宽度受限损失缩放

为了完整起见，我们分析 分析宽度受限损失项 $\mathcal{L}_{t\to\infty}(K) \sim K^{-\beta} + \mathcal{L}$, 的缩放，它决定了
不可约误差下限，在 训练结束时由模型容量所致。
我们研究最终饱和损失与学生维度 K 之间的关系，采用固定输入和通道统计（$a = 1.1, b = 0.0$）的叠加设置。我们将瓶颈大小 K 从 32 扫描到 512。为分离容量受限项，我们计算最终训练平台期（晚期动力学）的平均损失，这不同于正文中分析的训练中期幂律区间。

如图图 13所示，最终损失与宽度之间呈现出明显的幂律缩放关系。对数据进行拟合后，得到的指数为 $\beta = 1.30 \pm 0.01$。这一结果与近期关于叠加的研究（Liu 等人，2025年）报告的宽度缩放定律高度一致，证实除普适的学习速率（α）之外，模型的容量也大致与瓶颈维度成反比 K。

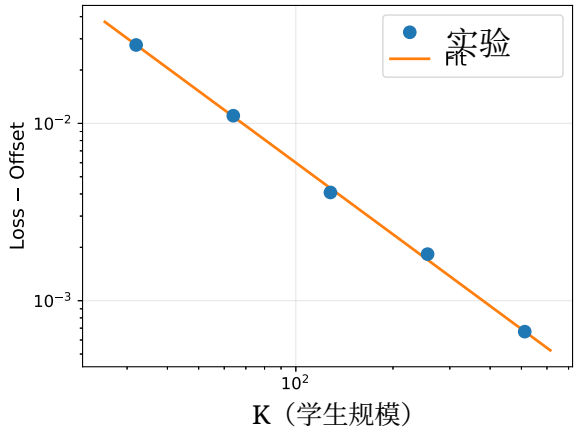


图 *13*。**最终平台期损失随学生宽度 K的变化**。损失按幂律衰减 $K^{-\beta}$，其中 $\beta \approx 1.3$，与叠加机制下的容量缩放定律一致。

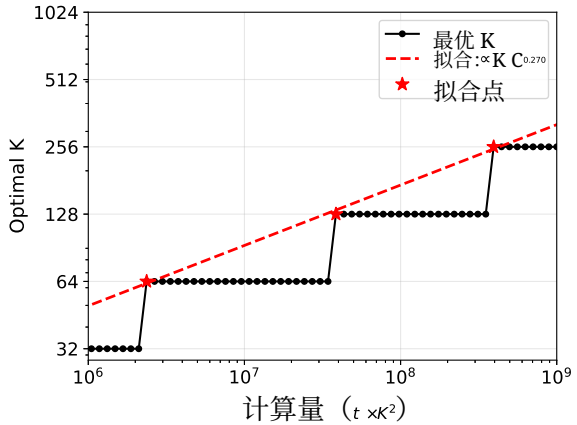


图 *14*。**最优模型规模随计算量缩放**。从前沿提取的最优学生规模 K 随计算预算缩放，从而可以预测最优资源分配。粗略估计下，相对于计算量的规模缩放指数为≈ 0.2702。