arXiv:2602.02593v1 [cs.LG] 1 Feb 2026

# 有效前沿：神经缩放定律的统一

**Jiaxuan Zou** [*1] **Zixuan Gong** [*2] **Ye Su** [3 4] **Huayi Tang** [2] **Yong Liu** [2]


## 摘要

神经缩放定律刻画了测试损失随模型容量（$N$）、数据规模（$D$）和计算量（$C$）呈幂律改善的预测规律。然而，现有理论解释往往依赖于特定架构或复杂的核方法，缺乏直观的普适性。本文提出一个统一框架，将一般学习任务抽象为对长尾（齐夫分布）中模式的渐进覆盖。我们引入**有效前沿**（$k_\star$），即模式秩空间中的一个阈值，用以区分已学习知识与未学习尾部。我们证明，可约损失渐近地由尾部的概率质量决定，即一种依赖于资源的**前沿截断**。基于该框架，我们推导了 $N$、 $D$和 $C$的精确缩放定律，并分别将其归因于**容量**、**覆盖度**和**优化** 瓶颈。此外，我们通过**最大瓶颈** 原则统一这些机制，表明 Kaplan 和 Chinchilla 缩放定律并非相互矛盾，而是同一约束优化问题在不同活跃瓶颈下的均衡解。


## 1. 引言

神经网络缩放定律已成为现代机器学习的基础经验常数。从语言建模（Brown 等，2020；Kaplan 等，2020）和计算机视觉（Zhai 等，2022），到多模态生成（Henighan 等，2020）等领域，测试损失相对于关键资源——模型规模 $N$、数据集大小 $D$ 和训练计算量 $C$（Hestness 等人，2017）——始终遵循精确的幂律轨迹。这些缩放定律为基础模型训练中的资源分配提供了指引。

然而，尽管这些定律在不同架构和数据集上都表现出经验上的普适性，其理论原理仍不明朗。目前，这些定律通常被视为观察所得的常数，而非由数学必然性推导得出。这种有限的理解限制了它们的预测能力，最明显的是，我们仍无法明确调和看似不同的缩放机制，例如 Kaplan 等人（2020）（Kaplan）提出的参数中心缩放，与 Hoffmann 等人（2022）（Chinchilla）后来展示的数据中心缩放。这促使我们建立一个统一框架，从理论上推导不同资源约束下的缩放定律。

解释缩放定律的理论研究大多分散于各自独立的视角，每种视角都只描绘了整体图景的一部分。一类研究依赖可求解的替代模型，例如核方法（Jacot et al., 2020年；Bordelon et al., 2021年；Maloney et al., 2022年；Bahri et al., 2024年）。尽管这些方法在数学上严谨，却往往局限于特定的架构极限，缺乏普适性。另一种视角侧重于谱分析，将缩放与流形的内在维度联系起来（Sharma & Kaplan, 2020年；Spigler et al., 2020年）。然而，这些连续方法往往忽略了自然数据固有的离散、重尾特性（例如齐夫分布定律）（Pan 等，2025）。除了静态谱极限，另一个研究方向分析优化动力学（Bordelon & Pehlevan, 2022年；Arous et al., 2025年；Ren et al., 2025年；Luo et al., 2025年；Li et al., 2025年），但通常假设容量或数据无限，未能考虑静态资源限制（$N$， $D$）与动态学习过程（$C$）之间的相互作用。关键在于，目前尚缺乏一个统一框架，能够推导所有三种瓶颈（$N$， $D$， $C$）对应的缩放定律，并调和它们相互冲突的权衡关系。

在这项工作中，我们将这一挑战视为希尔伯特第六问题的一个类比：从微观统计力学推导宏观热力学（Deng et al., 2025）。我们抽象掉架构细节，将该任务建模为从长尾分布中逐步获取信息。我们提出了**有效前沿**($k_\star$)这一概念，它是模式

[*]贡献相同 [1]西安交通大学数学与统计系，中国陕西省西安市 710049 [2]中国人民大学高瓴人工智能学院人工智能系，北京 100872 [3]中国科学院大学，北京 101407 [4]中国科学院深圳先进技术研究院，深圳 518055。通讯作者：Yong Liu < liuyonggsai@ruc.edu.cn>。

预印本。*2026*年*2*月*4*日。

秩空间中的一道清晰边界，将已学习知识与尚未学习的尾部分隔开来（Pan 等，2025）。我们表明，缩放定律只是这一前沿在资源约束下向重尾（齐夫分布）数据分布中推进的体现。我们的主要贡献总结如下：

- **通过有效前沿实现统一：** 我们提出一个统一框架，将学习概念化为有效前沿 $k_\star$在秩空间中的逐步推进。具体而言，该框架揭示了与不同资源瓶颈相关的缩放定律，例如模型容量（$N$）、数据规模（$D$）和计算量（$C$），从根本上取决于前沿向数据分布重尾部分推进的特定速率（第 2∼3 节）。

- **缩放定律的推导：** 我们通过刻画不同的瓶颈机制，解析地推导出 $N$、 $D$ 和 $C$ 的精确缩放指数：由记忆容量（第 4 节）和统计覆盖率（第 5 节）决定的静态瓶颈，以及由梯度下降的优化动力学和隐式偏置所产生的动态约束（第 6 节）。这一推导构成了一个统一基础，将数据的统计属性与多种物理约束联系起来。

- **调和最优前沿：** 通过将联合损失表述为一个“最大瓶颈”问题， $\Delta L \asymp \max(\varepsilon_N,\ \varepsilon_D,\ \varepsilon_\tau)$，我们化解了 Kaplan 和 Chinchilla 缩放定律之间表面上的冲突（Kaplan 等，2020；Hoffmann 等，2022）。我们证明二者并不矛盾，而是在不同活跃瓶颈条件下，同一个约束优化问题的平衡解（第 7 节）。

## 2. 问题设定

符号说明。本文中，我们使用记号 $o$和 $\asymp$来表示当 $R \to \infty$时的渐近行为。具体而言，$g(R) = o(f(R))$ 表示 $\lim_{R\to\infty} g(R)/f(R) = 0$。当 $f(R) = 1$时，我们简写为 $g(R) = o_R(1)$，表示一个随 $R \to \infty$趋于零的量。此外，$g(R) \asymp f(R)$ 表示存在两个正常数$c_1$、 $c_2$，使得 $c_1 f(R) \le g(R) \le c_2 f(R)$；而 $f(R) \propto g(R)$ 表示存在正常数 $c$，使得$f(R) = cg(R)$。所有这些常数均与 $R$无关。

**加性分解**。 记 $L$ 为总测试损失，则可约损失 定义为 $\Delta L := L - E$，其中 $E$ 是数据分布的不可约熵（贝叶斯误差）。直观而言， $\Delta L$ 表示可通过学习资源弥合的性能差距。要推导通用缩放定律，一个关键且必要的步骤是对可约损失进行数学建模。一方面，我们旨在纳入影响可约损失的主要因素。另一方面，为便于理论分析，我们希望这些因素之间的关系尽可能简单。因此，本文提出，数据分布中可学习的结构可以分解为一组离散且可学习的单元，我们称之为原子模式。换言之，我们假设，任何现实世界预测任务的复杂性都源于可数无穷多个底层结构的组合。基于此，我们给出以下定义。

**假设 2.1**（加性模式模型）。 假设学习任务可分解为独立子问题。因此，标量可约损失 $\Delta L$ 恰好可表示为逐模式残差的期望：

$$\Delta L = \sum_{k=1}^{\infty} p_k q_k, \tag{1}$$

其中， $k$ 表示不同的原子模式。 $p_k \ge 0$ 与$\sum p_k = 1$ 共同表示模式 $k$的统计频率，而 $q_k \in [0,1]$ 是模式 $k$ 的归一化残差（超额风险）。这里， $q_k = 1$ 表示该模式完全未被学习，而 $q_k \to 0$ 则表示已掌握。

严格来说，在有限规模下可能存在较弱的模式间相关性。尽管如此，公式（1）仍可视为适用于常见情形的渐近极限，在这些情形中，此类相关性会被忽略。更重要的是，公式（1）将宏观损失线性化为相互独立的微观组成部分，使我们能够分别分析在不同资源约束下 $q_k$ 的行为。

**齐夫尾部**。 缩放定律的观测结果要求权重 $\{p_k\}_{k=1}^{\infty}$ 具有特定的结构形式。指数衰减会导致学习几乎立即饱和，这与实践中的观测结果并不一致。相反，损失的幂律缩放源于数据的基本幂律分布。这种关系由齐夫分布形式化。

**假设 2.2**（齐夫分布）。 假设模式已按重要性排序（即 $p_1 \ge p_2 \ge \dots$），并且它们遵循以下给出的齐夫分布

$$p_k = \frac{1}{Z(\alpha)} k^{-\alpha}, \tag{2}$$

其中， $\alpha > 1$ 是衰减指数， $Z$（$\alpha$）是定义为 $Z$（$\alpha$）$= \sum_{k=1}^{\infty} k^{-\alpha}$ 的归一化常数。

**尾部行为**。 指数 $\alpha$ 刻画了尾部的厚重程度。具体而言，在极限 $\alpha \to 1^+$下，分布趋近于极重尾状态，其中包含在无限多个稀有模式中的总概率质量仍然显著。因此，这种内在的统计结构构成了我们统一框架的基础，在该框架中，

观察到的缩放现象，可以比喻为通过系统地将覆盖前沿不断推进至尾部深处，从而“收割”这种分散概率质量的过程。

## 3. 有效前沿

在本节中，我们将假设2.1中的原子模式分解，桥接到秩空间中依赖于资源的几何图景。

令 $R$ 表示一个标量资源预算，代表模型容量 $N$、数据规模 $D$或计算量 $\tau$。预算 $R$ 下的训练结果由逐模式残差分布 $\{q_k(R)\}_{k\geq 1}$刻画，其中 $q_k(R) \in [0,1]$ 量化模式 $k$的归一化超额风险。我们将 $q_k(\cdot)$视为资源 $R$的函数。自然地，额外资源不应降低性能，这意味着 $q_k(R)$关于 $R$单调不增。

为了将这一分布转化为一维几何，我们引入贪心学习偏差，它表明频繁出现的模式会被优先学习（Gong等人，2025年）。

**假设 3.1**（贪心学习偏差）。 残差分布关于频率秩 $k$ 单调不减，即，对所有 $k \geq 1$， $q_k(R) \leq q_{k+1}(R)$。

对于固定容差 $\delta \in (0,\ 1/2)$，我们将已学习区域的转移边界定义为最后一个大部分已学习索引 $k_-(R) := \sup\{k : q_k \leq \delta\}$，以及第一个大部分未学习索引 $k_+(R) := \inf\{k : q_k \geq 1-\delta\}$。假设 3.1 保证 $k_-(R) \leq k_+(R)$，从而防止已学习模式与未学习模式交错出现。现在我们用这些边界来形式化有效前沿的概念。

**定义 3.2**（有效前沿）。如果 $k_\star(R)$ 在渐近意义下刻画了转移边界，我们就称其为尖锐有效前沿：

$$k_-(R) = \big(1 - o_R(1)\big)\, k_\star(R),$$
$$k_+(R) = \big(1 + o_R(1)\big)\, k_\star(R).$$

定义 3.2 意味着过渡窗口的相对宽度渐近趋于零，即，$(k_+ - k_-)/k_\star \to 0$。直观而言， $k_\star(R)$ 表示学习实际上停止的唯一宏观尺度（见图1）。这说明可以将残差分布 $q_k(R)$ 近似为宏观尺度上的阶跃函数： $q_k \approx 0$ 当 $k \lesssim k_\star$ 时，以及 $q_k \approx 1$ 当$k \gtrsim k_\star$ 时。

如图2所示，将学习过程视为秩空间中的离散截断后，可约损失（假设 2.1）便化为未学习尾部的概率质量：$\Delta L(R) \asymp \sum_{k>k_\star(R)} p_k$· 这种几何约化使我们只需分析模式分布的尾部行为，就能推导出缩放定律。在此基础上，我们提出普适缩放原理。

**定理3.3** (普适缩放原理)**.** 在假设 *2.1*∼*2.2* 下，如果某种资源 $R$ 诱导出一个有效前沿$k_\star(R)$ （定义 *3.2*），则可约损失按以下规律缩放：

$$\Delta L(R) \asymp k_\star(R)^{-(\alpha-1)}.$$

定理 3.3 将任何具体缩放定律的推导归结为一个纯粹的几何问题：确定特定瓶颈对应的**资源到前沿的映射** $R \mapsto k_\star(R)$（详细证明见附录 C）。我们确定了三种不同的情形：

• 模型缩放（模型容量 $N$）： 容量受限前沿 $k_\star(N)$。

• 数据缩放（数据规模 $D$）： 覆盖受限前沿$k_\star(D)$。

• 计算缩放（计算预算 $C$ 或优化步数 $\tau$）： 优化受限前沿 $k_\star(\tau)$，其中 $C \propto \tau$ 对于固定的模型配置成立。

## 4. 模型缩放：容量前沿

在本节中，我们推导模型缩放定律。我们考虑数据规模 $D$ 和计算量 $C$ 充足的情形，此时模型容量（$N$）成为主要瓶颈。在这一极限下，有效容量前沿 $k_\star(N)$ 和可约损失受到模型记忆容量的约束；记忆容量体现了模型从重尾分布中存储并解析复杂模式的基本能力。

为了刻画映射 $N \mapsto k_\star(N)$，我们以几何概念有效自由度为分析基础。从几何角度看，学习 $k$ 正交模式会施加 $k$ 个独立约束。要满足这些约束，需要足够的自由度，而参数预算 $N$定量地提供了这些自由度。近期的理论和实证研究证实了这一观点：深度网络完美拟合随机模式的容量随参数数量 $N$ 线性增长（Vershynin，2020年；Zhang 等人，2021年）。

基于这一理解，我们将模型规模与可学习前沿之间的关系形式化。

**假设4.1**（容量缩放）。 有效容量前沿随参数数量 $N$按 $k_\star(N) \asymp N^\gamma$，缩放，其中 $\gamma \in (0,1]$ 为架构效率因子。

注释 4.2。对于学习正交特征的理想线性模型，容量与参数数量严格成正比（$k_\star \asymp N$ 其中 $\gamma = 1$）。在深度非线性网络中，记忆容量理论支持这种关系；该理论证明，ReLU 网络能够完美插值的任意模式数量与其参数数量呈线性关系 $N$（Vershynin，2020年；Yun 等人，2019年）。尽管架构冗余、参数

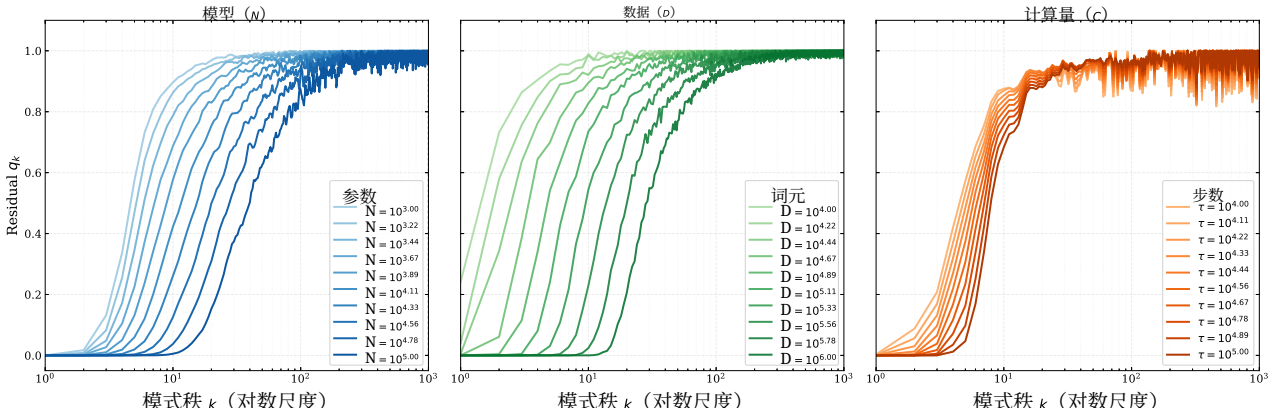



图*1.* **秩空间中的有效前沿。** 不同约束下，残差分布 $q_k$ 与模式秩 $k$（对数尺度）的关系：（左）模型容量 $N$，（中）数据集大小 $D$，（右）计算量 $\tau$（注意，对于固定的模型配置，$C \propto \tau$）。在所有情况下，增加资源都会将清晰的覆盖前沿 $k_\star$ 推向齐夫尾部更深处（$\alpha = 1.5$）。这验证了定义 3.2 中的几何抽象，即将学习视为一个渐进的覆盖过程。

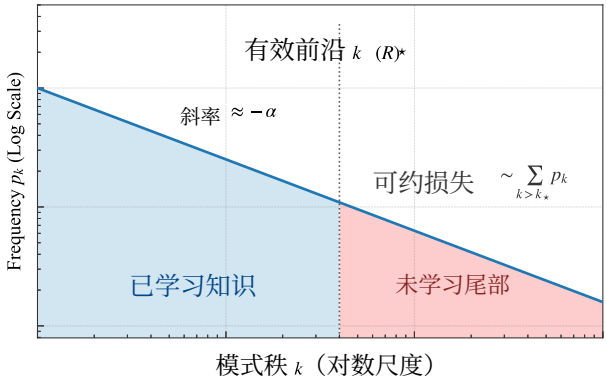



图*2.* 秩空间中的有效前沿 $k_\star(R)$。在齐夫频率 $p_k \propto k^{-\alpha}$下，资源 $R$（例如 $N$、$D$、$\tau$）会产生一个截断：已学习模式（$k \le k_\star$）与未学习尾部（$k > k_\star$）。可约损失主要由尾和 $\sum_{k>k_\star(R)} p_k$决定；随着资源增加，$k_\star(R)$ 向右移动，尾和随之减小。

共享或基于梯度的学习效率低下等实际因素可能导致次线性缩放（$\gamma < 1$）（Sharma & Kaplan，2020年）。尽管如此，基本依赖关系仍根植于上述认识之中，保留了幂律形式。

下面我们给出模型缩放定律。

**命题 4.3** (模型缩放定律)。 在假设 *2.1*~*2.2* 和假设 *4.1*下，可约损失随模型参数 $N$ 缩放如下：

$$\boxed{\Delta L(N) \asymp N^{-\gamma(\alpha-1)}.}$$

命题 4.3 给出了受模型容量约束的渐近缩放极限。重要的是，这一结果在结构上将经验指数分解为两个组成部分：(a) 项 $(\alpha - 1)$ 表示数据结构，仅由分布的重尾性决定；(b) 因子 $\gamma$ 则体现了架构效率，用于量化特定架构利用参数进行模式学习的能力。

## 5. 数据缩放：覆盖前沿

本节推导数据缩放定律。具体而言，我们考虑模型容量和计算量充足、有限数据覆盖成为主要瓶颈的情形。我们采用三步理论方法：*(i)*在模式层面定义由覆盖度诱导的概率残差代理；*(ii)* 利用加性分解将其提升为数据集层面的可约损失代理；以及*(iii)* 推导其在秩空间中诱导的有效前沿。

设训练集包含从底层模式分布 $\{p_k\}_{k\ge1}$中独立同分布抽取的 $D$ 个样本*i.i.d.*。对于频率为 $p_k$的每个模式 $k$，其出现次数服从二项分布 $X_k \sim \text{Binomial}(D, p_k)$。受以下观察启发：对于训练集中实际不存在的模式，没有任何学习规则能够可靠地降低其残差误差；我们将以下内容作为步骤*(i)*引入残差代理。

**定义 5.1** （残差代理）。我们将模式 $k$ 的覆盖度诱导残差代理定义为未观测概率，即，$q_k(D) \triangleq \Pr[X_k = 0] = (1 - p_k)^D$。

定义 5.1 隐含地假设，在无限算力区间中，出现一次就足以完成学习。我们在定义 5.3 中通过引入广义阈值 $m$扩展这一表述，要求至少出现 $m$ 次才能有效学习；在有限数据设置下，这自然与多轮数据复用相关联。（Yan 等人，2025）

接下来进行步骤（*ii*）和（*iii*），我们汇总这些模式级残差，推导数据集级缩放行为。

**定理 5.2** （数据缩放定律）。在假设

*2.1*∼*2.2* 和定义 *5.1* 下，有效覆盖度前沿满足

$$k_\star(D) \asymp D^{1/\alpha}.$$

因此，可约损失随数据规模 $D$ 缩放为

$$\boxed{\Delta L(D) \asymp D^{-\frac{\alpha-1}{\alpha}}.}$$

定理 5.2 给出了数据缩放行为，明确量化了尾部厚重程度 $\alpha$ 如何通过指数 $\alpha_D = (\alpha-1)/\alpha$ 支配样本效率。在重尾极限（$\alpha \to 1^+$）下，该指数趋于消失（$\alpha_D \to 0$），这反映了独立同分布采样在覆盖稀有模式长尾时固有的低效性。反之，更大的$\alpha$（更轻的尾部）会产生更大的 $\alpha_D$，从而加快额外数据转化为有效覆盖度与损失降低的速率（见图 3）。详细证明留待附录 D.1。

虽然定义 5.1 提供了一个简单的残差代理，但现实中的学习往往需要反复接触，才能将信号与噪声区分开来。为刻画这一点，我们引入一个广义代理，要求出现次数达到最低阈值。

**定义 5.3**（广义残差代理）。我们将模式 $k$ 的广义覆盖诱导残差代理定义为：模式出现次数 $X_k$ 低于阈值 $m > 1$ 的概率，即，$q_k(D) \triangleq \Pr[X_k < m]$。

**推论 5.4**。在假设 *2.1*∼*2.2* 和定义 *5.3* 下，有效覆盖前沿满足

$$k_\star(D) \asymp \left(\frac{D}{Zm}\right)^{1/\alpha}.$$

因此，可约损失随数据规模 $D$ 按如下规律缩放：

$$\boxed{\Delta L(D) \asymp D^{-\frac{\alpha-1}{\alpha}}.}$$

推论 5.4 展示了数据缩放定律的普适性，表明渐近指数不受最小出现阈值 $m$ 的影响。详细证明见附录 D.2。

## 6. 计算缩放：优化前沿

在本节中，我们推导计算扩展定律。虽然模型容量（$N$）和数据规模（$D$）施加了静态极限，但计算量（$C$）对学习过程引入了动态约束。即使容量和数据无限，若优化尚未对罕见的尾部模式收敛，可约损失仍不为零。为确立这一点，我们首先从优化动力学分析中推导梯度下降（GD）的精确缩放指数（第6.1节）。随后，我们通过自相似缩放核对残差演化进行建模来扩展这一结果，从而形式化一条独立于特定优化器的一般缩放定律（第6.2节）。

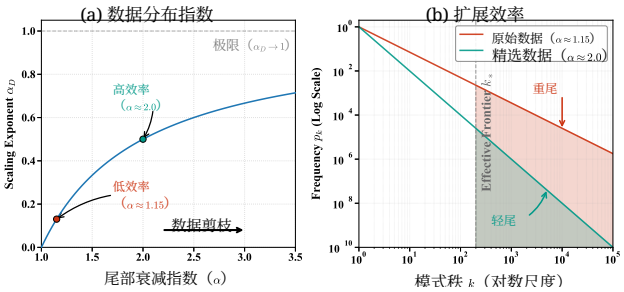



图 *3*。**尾部厚重程度决定数据缩放效率。（a）**数据缩放指数 $\alpha_D$ 由尾指数 $\alpha$ 解析确定（定理 5.2）。尾部较轻（$\alpha$ 较大，例如通过数据剪枝实现）会带来严格更高的样本效率。**（b）**从几何角度看，对于固定的有效前沿 $k_*$，重尾分布（红色）比经过筛选的分布（青绿色）保留了显著更多的未学习概率质量（阴影尾部），因此损失下降更慢。

### 6.1. 梯度下降的优化动力学与缩放定律

我们采用梯度下降作为典型示例，为计算缩放分析奠定基础。具体而言，我们识别出由优化过程诱导的动态有效前沿，并推导梯度下降机制下由此产生的计算扩展定律。在陈述正式定理之前，我们首先建立必要的假设和引理。我们考虑独立模式的解耦优化轨迹（假设2.1）。令 $L_k(\theta)$表示与模式 $k$相关联的损失分量，而 $\nabla L_k(\theta)$表示其关于参数 $\theta$的梯度。

**假设 6.1**（正交性与 Polyak-Lojasiewicz（PL）条件）。我们假设损失景观满足以下几何条件：

*(a)* $\epsilon$-正交性：对于任意一对不同的模式 $j \neq k$，其梯度的内积以一个较小的常数 $\epsilon \geq 0$ 为界：$|\langle \nabla L_j(\theta), \nabla L_k(\theta) \rangle| \leq \epsilon$。*(b)* 逐模式 *PL* 条件：对于每个模式 $k$，梯度范数以下界由其残差 $q_k$ 给出：$\|\nabla L_k(\theta)\|^2 \geq 2\lambda_k q_k(\theta)$，其中 $\lambda_k > 0$ 表示该模式特定的 PL 常数。*(c)* 逐模式光滑性：损失分量 $L_k$ 在轨迹区域上是 $\beta_k$-光滑的，即 $\|\nabla L_k(\theta) - \nabla L_k(\theta')\| \leq \beta_k \|\theta - \theta'\|$。

假设 6.1 将损失景观的几何性质形式化。条件 (a) 确保参数空间中的梯度近乎正交，这意味着不同模式的学习轨迹在 $\epsilon$-误差范围内实际上彼此解耦。条件 (b) 假设每个损失分量都满足 PL 不等式，这是非凸优化分析中的标准假设（Polyak，1963年；Lojasiewicz，1963年；Karimi 等人，2016年）。这里，$\lambda_k$ 表示

特定模式的 PL 常数，对应于局部二次区域内分块对角 Hessian 的最小特征值。

基于这些条件，我们在以下引理 6.2 中形式化逐模式残差 $q_k$ 的演化。

**引理 6.2**（迭代残差动力学）。设 $I_{t,\,k} \in\{0,1\}$ 为指示变量：若模式 $k$ 在步骤 $t$被采样，则$I_{t,\,k}=1$；否则$I_{t,\,k}=0$。在假设 *6.1* 下，残差 $q_k\,(t)$ 的演化为

$$q_k(t+1)=\begin{cases}(1-\eta\lambda_k)\,q_k(t) & \text{if } I_{t,k}=1,\\ q_k(t) & \text{if } I_{t,k}=0,\end{cases}$$

其中，$\eta$ 表示（常数）学习率，$\lambda_k$ 是特定模式的 *PL* 常数。

引理 6.2 表明，残差仅在被采样时才会几何收缩。我们将 $\lambda_k$ 称为有效误差校正系数，因为它决定了每一步的缩减率。

备注 6.3。更一般的学习率调度 $\eta=\eta(t)$ 会产生相同的“仅在被采样时收缩”结构，但收缩因子会随时间变化；近期研究探讨了明确纳入学习率调度的函数缩放定律。（Li et al., 2025年）

详细证明见附录 E.1。

此外，我们假设 $\lambda_k$ 在不同模式间并不均一，并基于深度线性网络的理论洞见，将这种非齐次性形式化；在深度线性网络中，收敛速率本质上随模式频率 $p_k$的幂次变化。

**假设 6.4**（非齐次性）。有效误差校正系数 $\lambda_k$ 遵循幂律 $\lambda_k \propto p_k^{\beta-1}$，其中 $\beta>0$ 为隐式偏置指数。

**理论依据**。假设 6.4 将频繁模式优先得到学习这一归纳偏置形式化，体现为非齐次动力学。我们在附录 H 中通过对深度线性网络（DLN）进行梯度流分析，推导出这一幂律。具体而言，我们证明偏置指数 $\beta$ 由网络深度 $L$ 和可学习性指数 $\zeta$（假设 H.1）决定，满足 $\beta=2+\zeta\,(2-2/L)$。重要的是，这一结果意味着 $\beta\geq 2$，从而使优化处于丰富或特征学习机制下；在此机制中，深度引入了非线性选择性，主动加速对占主导地位的高频特征的学习。我们的理论分析还基于特定的 DLN 架构进一步验证了引理 6.2，且无需依赖假设 6.1 中的 PL 条件。

**引理 6.5**（渐近残差动力学）。 在引理 *6.2* 和假设 *6.4* 下，当训练步数$\tau$ 较大且常数为 $c>0$ 时，残差 $q_k\,(\tau)$ 依概率收敛到以下指数形式：

$$q_k(\tau)\asymp \exp\big(-c\,\tau\,p_k^{\beta}\big). \tag{3}$$

**定理 6.6**（梯度下降的计算扩展定律）。 在假设 *2.1∼2.2* 和引理 *6.5* 下，动态有效前沿 $k_\star\,(\tau)$ 满足

$$k_\star(\tau)\asymp \tau^{\frac{1}{\alpha\beta}}.$$

因此，可约损失的缩放规律为

$$\boxed{\Delta L(\tau)\asymp \tau^{-\frac{\alpha-1}{\alpha\beta}}.}$$

定理 6.6 将引理 6.5 中微观的优化行为与宏观计算缩放定律联系起来。具体而言，它刻画了这样一种优化动力学：有效前沿 $k_\star(\tau)$ 逐渐推进至尾部。这一机制表明，可约损失主要由这一移动前沿之外的累积未学习质量决定。因此，渐近计算缩放取决于数据的尾部厚重程度 $\alpha$ 与模型的归纳偏置 $\beta$之间的相互作用。详细证明见附录 E.1。

### 6.2. 推广至一般计算缩放定律

定理 6.6 中建立的缩放定律基于梯度下降动力学中特有的指数残差衰减（引理 6.5）。鉴于经验缩放定律在各种优化器之间表现出显著的普适性，我们现在将公式 (3) 推广至一般的自相似缩放核。

**假设6.7**（自相似缩放核）。 假设残差演化遵循由单调缩放核 $g$ 控制的自相似形式： $[0,\infty)\to[0,1]$:

$$q_k(\tau)\asymp g\big(c\,\tau\,p_k^{\beta}\big),$$

其中 $c>0$为常数。核 $g$ 满足边界条件 $g(0)=1$ 以及可积性条件$\int_0^\infty u^{s-1}g(u)du<\infty$。

假设6.7将GD下的指数衰减扩展到了任意优化器。尽管具体的轮廓 $g(\cdot)$可能有所不同（例如指数型或多项式型），但其基本的自相似性质仍然存在：当以有效时间 $\tau p_k^{\beta}$为参数时，不同模式的学习曲线会坍缩到单一主曲线 $g(\cdot)$上。

接下来，我们将形式化一般计算缩放定律。

**定理6.8**（一般计算缩放定律）。 在假设*2.1∼2.2*和假设 *6.7*下，当训练步长 $\tau$较大时，可约损失的缩放关系为

$$\boxed{\Delta L(\tau)\asymp \mathcal{K}\cdot\tau^{-s},}$$

其中 $s=(\alpha-1)/\alpha\beta$，且前因子 $\mathcal{K}$ 由$\mathcal{K}=(\alpha\beta)^{-1}Z^{-1/\alpha}c^{-s}[\int_0^\infty u^{s-1}g(u)\,du]$给出。

定理6.8阐明了支撑计算缩放定律的最小条件，明确将一般指数 $s = (\alpha - 1)/\alpha\beta$ 与核依赖前置因子解耦。至关重要的是，我们发现缩放率内在取决于数据的尾部厚重程度 $\alpha$ 与归纳偏置 $\beta$，而 $g$ 的形状仅影响系数 $\mathcal{K}$。详细证明留待附录E.2。

## 7. 联合缩放与统一前沿

在本节中，我们提出一种基于渐近主导的**组合定律**，以解决 Kaplan 等人（2020）与 Hoffmann 等人（2022）的缩放定律之间的冲突。

**命题 7.1 (**最大瓶颈原理**)**。 联合可约损失在渐近意义上由容量（$N$）、数据规模（$D$）和优化步数（$\tau$）这几项约束中最严格的一项主导：

$$\Delta L(N, D, \tau) \asymp \max\left(\varepsilon_N(N),\, \varepsilon_D(D),\, \varepsilon_\tau(\tau)\right).$$

尽管经验定律往往呈现加和形式，但渐近缩放由主导项决定。这种结构意味着存在一个转折点 $\tau_\star$，此处动态优化瓶颈 $\varepsilon_\tau$ 与由模型规模或数据规模决定的静态极限相交（见附录 F.2）。

**统一最优缩放**。 我们可以将最优缩放定律推导为约束优化问题的解，而非经验拟合。最优前沿是给定计算预算 $C$时，使联合损失最小的配置（$N, D, \tau$）的集合：

$$\min_{N,D,\tau} \max\left(AN^{-\alpha_N}, BD^{-\alpha_D}, G\tau^{-\alpha_\tau}\right). \tag{4}$$

在乘法预算 $C$ 下最小化方程（4），需要平衡活跃瓶颈。该框架恢复了两种不同制度的平衡解（推导见附录 F.3）：

**制度 A：数据充裕（Kaplan 等人（2020））**。在数据充裕极限（$D \to \infty$）下，瓶颈位于容量与优化步数之间（$N$ 与 $\tau$）。平衡 $\varepsilon_N \asymp_{\alpha_\tau} \varepsilon_\tau$，得到参数中心缩放 $N_{\text{opt}} \propto C^{\frac{\alpha_\tau}{\alpha_N + \alpha_\tau}}$。关键的是，该指数取决于优化偏差 $\beta$。

**制度 B：数据受限（Hoffmann 等人（2022））**。在固定数据极限（$D \propto C/N$）下，瓶颈位于容量与数据覆盖之间（$N$ 与 $D$）。平衡 $\varepsilon_N \asymp \varepsilon_D$，得到平衡缩放 $N_{\text{opt}} \propto C^{\frac{\alpha_D}{\alpha_N + \alpha_D}}$，其指数由数据尾部厚度 $\alpha$决定。

总之，我们的框架表明，Kaplan定律和Chinchilla定律因此是方程（4）的有效解在不同的有效约束下。最优标度指数是一个拓扑不变量，由数据结构（$\alpha$）与优化动力学（$\beta$）之间的相互作用决定。

## 8. 实验

在本节中，我们使用附录A所述的受控合成环境，对我们的统一框架进行实证验证。我们的目标是验证三项核心主张：（1）资源约束在秩空间中表现为一条尖锐的有效前沿$k_\star$；（2）该前沿按照定理4.3、5.2和6.6中推导出的特定几何幂律推进；（3）由此得到的可约损失缩放指数严格由数据尾部厚重度（$\alpha$）与所推导的前沿缩放之间的相互作用决定。

### 8.1. 有效前沿的可视化

图1针对固定的数据分布（$\alpha = 1.5$），将逐模式残差 $q_k$ 相对于频率秩 $k$ 进行了可视化。结果表明，存在一个明显的相变，它通过一个狭窄区域将“已学习”模式（$q_k \approx 0$）与“未学习”尾部（$q_k \approx 1$）分隔开来。随着资源增加——无论是模型容量 $N$（左）、数据规模 $D$（中）还是计算量 $\tau$（右）——该前沿在保持形状的同时系统地向右平移。这些经验结果验证了我们将学习近似为阶跃函数的理论（定义 3.2），并证实各种物理瓶颈可以普遍抽象为有效前沿 $k_\star$ 向重尾分布内部的逐步推进。

### 8.2. 有效前沿的缩放

图 4 在双对数尺度上绘制了提取出的有效前沿 $k_\star$ 与资源（$N$、 $D$、 $\tau$）之间的关系。经验测量结果与我们框架推导出的几何幂律精确吻合：覆盖前沿（中间）按 $k_\star \propto D^{1/\alpha}$ 缩放（定理 5.2），证实样本效率完全由尾指数 $\alpha$决定；而容量前沿（左侧）则遵循 $k_\star \propto N^{\gamma}$（假设4.1）。此外，优化前沿（右侧）呈现出预测的 $k_\star \propto \tau^{1/\alpha\beta}$ 缩放规律（定理 6.6）；观测到的斜率表明隐式偏置为 $\beta \approx 2.03$，与理论预测 $\beta \approx 2$（假设 6.4）高度一致。这些结果验证了前沿的扩展受数据结构（$\alpha$）与优化动力学（$\beta$）之间精确相互作用的支配；我们还确认，这一结论在附录B所述的不同分布下均稳健成立。

### 8.3. 损失缩放定律的普适性

图5展示了不同尾指数 $\alpha \in \{1.3, \ldots, 2.1\}$ 下的加权可约损失 $\Delta L$ 与$N$、 $D$和 $\tau$的关系。

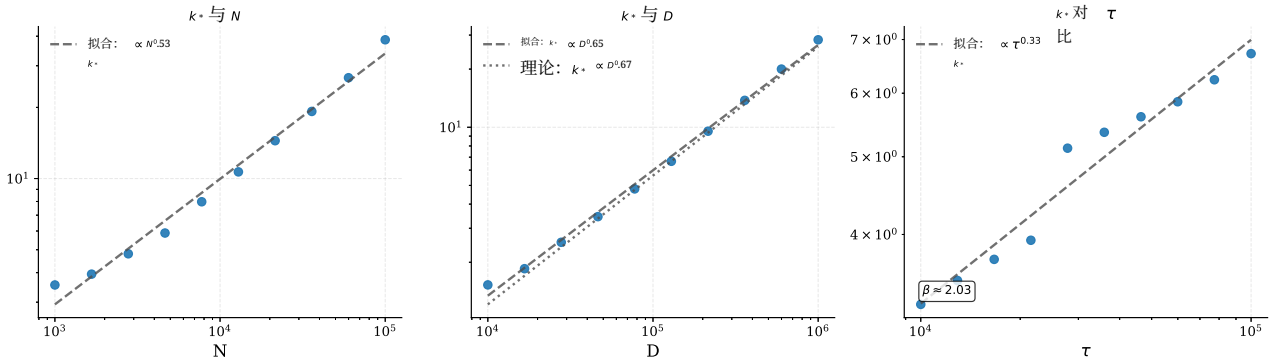



图 4. **前沿的几何缩放**。 我们提取资源的函数 $k_\star$（阈值为 $\delta = 0.5$）。（左）容量前沿 $k_\star \propto N^{0.53}$。（中）覆盖前沿 $k_\star \propto D$。关键的是，对于 $\alpha = 1.5$，理论预测 $D^{1/\alpha}$（细虚线）与经验拟合（粗虚线）完美吻合，证实了定理 5.2。（右）优化前沿 $k_\star \propto \tau$。由于固定模型的总计算量 $C \propto \tau$，这反映了计算扩展定律。斜率使我们能够估计优化偏差 $\beta \approx 2.03$。

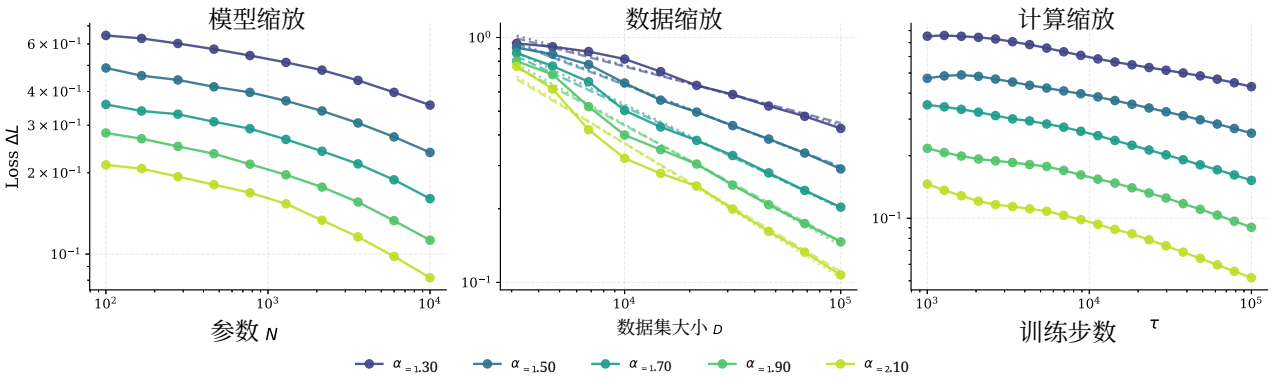



图5. **通用缩放定律**。 不同尾部厚重程度 $\alpha$的分布，其可约损失 $\Delta L$ 与资源的关系。（中）对于数据缩放，经验曲线（实线）与理论预测 $\Delta L \propto D^{-(\alpha-1)/\alpha}$（虚线）完美吻合，证实了 $\alpha$决定样本效率。（右）计算缩放定律（通过步数 $\tau$测量，其中固定模型的 $C \propto \tau$）表明，较轻的尾部（较大的 $\alpha$）会带来更快收敛，这与推导出的指数 $-\frac{\alpha-1}{\alpha\beta}$ 一致。

结果表明，所有机制下都呈现出稳健的幂律缩放。最重要的是，我们将经验缩放指数与理论推导进行比较：定理 5.2 预测 $\Delta L \propto D^{-\frac{\alpha-1}{\alpha}}$。在图5（中）中，虚线表示由真实值 $\alpha$ 确定的理论斜率。它们与经验数据（实线）高度吻合，这证实了数据缩放纯粹是覆盖齐夫分布的一种统计必然性。对于计算缩放（右），随着 $\alpha$ 增加，斜率逐渐变平。利用前沿分析（图 4）得出的$\beta \approx 2.03$，预测的计算缩放指数 $-\frac{\alpha-1}{\alpha\beta}$ 准确地重现了观测到的变化轨迹，从而验证了定理 6.6。

总而言之，这些实验证实，神经缩放定律并非不透明的经验值常数，而是有效前沿向重尾数据分布推进所产生的精确数学结果。

## 9. 结论

在这项工作中，我们提出了一个用于理解神经缩放定律的统一框架。我们将视角从架构微观细节转向**模式秩空间几何**。我们表明，缩放定律并非深度网络中神秘的涌现现象，而是在资源约束下从重尾数据中学习所必然产生的统计结果（Newman，2005）。我们的框架核心是**有效前沿**$k_*$这一概念，它将缩放过程分解为一个固定的数据几何指数和一个依赖于资源的前沿截断。这种抽象使我们能够解析地推导模型容量（$N$）、数据规模（$D$）和计算量（$C$）的缩放定律，同时成功地将 Kaplan（Kaplan 等，2020）和 Chinchilla（Hoffmann 等，2022）看似矛盾的情形统一解释为不同瓶颈构成下的平衡解。最后，我们的结果表明，缩放指数是数据结构与优化动力学相互作用的基本不变量，

为理论约束与经验性的深度学习表现之间搭建了一座有原则的中观桥梁。

**局限性与未来工作**。 虽然我们的框架统一了标准训练范式下的缩放定律，但目前仍将数据结构（$\alpha$）和优化偏差（$\beta$）视为静态约束。不过，这一局限也指明了一个颇具吸引力的未来方向：从观察转向打破 这些定律。我们的推导表明，缩放指数是 $\alpha$ 和 $\beta$的可变函数，这意味着主动干预，尤其是通过**数据剪枝** 来平滑重尾（提高有效 $\alpha$），以及通过**课程学习**重新加权梯度动力学（调节 $\beta$），都可能改善缩放率。未来的研究应探索这些机制，以设计新一代方案。

## 影响声明

本文旨在推动机器学习领域的发展。我们的工作可能带来许多社会影响，但我们认为其中没有任何一项需要在此特别强调。

## 参考文献


Arora, S.、Cohen, N. 和 Hazan, E. 深度网络的优化：过参数化带来的隐式加速。发表于国际机器学习大会，第 244–253 页。PMLR，2018年。

Arous, G. B., Erdogdu, M. A., Vural, N. M., and Wu, D. 高维下二次神经网络的学习：随机梯度下降动力学与缩放定律，2025年。URL[https://arxiv.org/abs/2508.03688](https://arxiv.org/abs/2508.03688)。

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. 解释神经网络缩放定律。美国国家科学院院刊，121(27)，2024年6月。ISSN 1091-6490。doi: 10.1073/pnas.2311878121。URL [http://dx.doi.org/10.1073/pnas.2311878121](http://dx.doi.org/10.1073/pnas.2311878121)(http: //dx.doi.org/10.1073/pnas.2311878121)。

Bordelon, B.andPehlevan, C. 结构化特征上的随机梯度下降学习曲线，2022年。URL [https: //arxiv.org/abs/2106.02713](https: //arxiv.org/abs/2106.02713)(https: //arxiv.org/abs/2106.02713)。

Bordelon, B., Canatar, A., 和 Pehlevan, C. 核回归和宽神经网络中依赖谱的学习曲线，2021年。网址：https://arxiv.org/ abs/2002.02561。

Bordelon, B.、Atanasov, A. 和 Pehlevan, C.《神经缩放定律的动力学模型》，2024年。网址 [https://arxiv.org/abs/2402.01092](https://arxiv.org/abs/2402.01092)(https://arxiv.org/abs/2402.01092)。Bordelon, B.、Atanasov, A. 和 Pehlevan, C.《特征学习如何改进神经缩放定律》，2025年。网址[https://arxiv.org/abs/2409.17858](https://arxiv.org/abs/2409.17858)(https://arxiv.org/abs/2409.17858)。

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., 和 Amodei, D. 语言模型是小样本学习器，2020年。网址：https:// arxiv.org/abs/2005.14165。

Deng, Y.、Hani, Z. 和 Ma, X. 希尔伯特第六问题：通过玻尔兹曼动理论推导流体方程，2025年。链接：https://arxiv.org/abs/2503. 01800。

Gong, Z.、Li, S.、Liu, Y. 和 Teng, J. 解耦特征结构：一种可从数学上证明的 Transformer 两阶段训练动力学，2025年。链接：https: //arxiv.org/abs/2502.20681。

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. 自回归生成建模的缩放定律，2020年。链接：https://arxiv.org/abs/ 2010.14701。

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., 和 Zhou, Y. 深度学习的规模扩展在经验上是可预测的，2017年。URL[https: //arxiv.org/abs/1712.00409](https: //arxiv.org/abs/1712.00409)(https: //arxiv.org/abs/1712.00409)。

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., Rae, J. W., Vinyals, O., 和Sifre, L. 训练计算量最优的大语言模型，2022年。URL https://arxiv.org/ abs/2203.15556。

Hutter, M. 学习曲线理论，2021年。网址：https: //arxiv.org/abs/2102.04074。

Jacot, A., Gabriel, F. 和 Hongler, C. 神经正切核：神经网络中的收敛与泛化，2020年。网址：https://arxiv.org/abs/1806.07572。

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. 和 Amodei, D. 神经语言模型的缩放定律，2020年。网址：https://arxiv.org/abs/2001. 08361。

Karimi, H., Nutini, J., and Schmidt, M. 在 Polyak–Łojasiewicz 条件下梯度方法和近端梯度方法的线性收敛性。载于欧洲机器学习与数据库知识发现联合会议，第 795–811 页。施普林格，2016年。

Li, B., Chen, F., Huang, Z., Wang, L., and Wu, L. 核回归中的函数缩放定律：损失动态与学习率调度，2025年。网址：https://arxiv. org/abs/2509.19189。

Lin, L., Wu, J., Kakade, S. M., Bartlett, P. L., and Lee, J. D. 线性回归中的缩放定律：计算、参数与数据，2025年。网址：[https: //arxiv.org/abs/2406.08466](https: //arxiv.org/abs/2406.08466)。

Lojasiewicz, S. 实解析子集的一个拓扑性质。*CNRS*讨论会，《 ′偏微分方 ′程 ′》，117(87-89):2，1963年。

Luo, K., Wen, H., Hu, S., Sun, Z., Liu,Z.,Sun, M.,Lyu, K.和 Chen, W. 一种用于跨学习率调度进行损失曲线预测的多重幂律，2025年。URL https://arxiv.org/abs/2503.12811。

Maloney, A., Roberts, D.A.和Sully, J. 神经缩放定律的一个可解模型，2022年。URL[https://arxiv.org/abs/2210.16859](https://arxiv.org/abs/2210.16859)。

Michaud, E. J., Liu, Z.和Tegmark, M. 精密机器学习。*Entropy*，25(1):175，2023年1月。ISSN 1099- 4300。doi: 10.3390/e25010175。URL http://dx.doi.org/10.3390/e25010175。

Michaud, E. J., Liu, Z., Girit, U. 和 Tegmark, M. 神经网络缩放的量化模型，2024年。URL[https://arxiv.org/abs/2303.13506](https://arxiv.org/abs/2303.13506)。

Newman, M. 幂律、帕累托分布和齐夫定律。《当代物理学》，46(5):323–351，2005年9月。ISSN 1366-5812。doi: 10.1080/ 00107510500052444。URL http://dx.doi.org/ 10.1080/00107510500052444。

Pan, Z., Wang, S. 和 Li, J. 通过压缩理解大语言模型行为：数据生成、知识获取与缩放定律，2025年。网址：https://arxiv.org/abs/2504.09597。

Paquette, E., Paquette, C., Xiao, L. 和 Pennington, J. 4+3 计算最优神经网络缩放定律的阶段，2025年。网址[https: //arxiv.org/abs/2405.15074](https: //arxiv.org/abs/2405.15074)。

Polyak, B. T. 用于函数最小化的梯度方法。计算数学与数学物理杂志，3(4):643–653，1963年。

Ren, Y., Nichani, E., Wu, D. 和 Lee, J. D. 浅层神经网络随机梯度下降学习中的涌现与缩放定律，2025年。URL[https: //arxiv.org/abs/2504.19983](https: //arxiv.org/abs/2504.19983)。

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

Sharma, U. 和 Kaplan. J. 基于数据流形维数的神经网络缩放定律，2020年。网址：https: //arxiv.org/abs/2004.10802。

Spigler, S., Geiger, M., 和 Wyart, M. 核方法的渐近学习曲线：经验数据与教师—学生范式。统计力学杂志：理论与实验，2020年(12):124001，2020年12月。ISSN 1742-5468。doi: 10.1088/1742-5468/ abc61d。网址 http://dx.doi.org/10.1088/ 1742-5468/abc61d。
Vershynin, R. 阈值和 ReLU 激活函数神经网络的记忆容量。*arXiv* 预印本 *arXiv:2001.06938*，2020年。
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., 和 Fedus, W. 大语言模型的新兴能力，2022年。网址：https: //arxiv.org/abs/2206.07682。

Yan, T.、Wen, H.、Li, B.、Luo, K.、Chen, W. 与 Lyu, K. 更大的数据集可以重复更多次：线性回归中多轮次缩放的理论分析，2025年。URL https://arxiv.org/abs/2511.13421。

Yun, C.、Sra, S. 与 Jadbabaie, A. 小型 ReLU 网络是强大的记忆器：对记忆容量的紧致分析。神经信息处理系统进展，32，2019年。

Zhai, X., Kolesnikov, A., Houlsby, N. 和 Beyer, L. 视觉变换器的规模化，2022年。网址：https://arxiv.org/abs/2106.04560。

Zhang, C., Bengio, S.,Hardt, M., Recht, B. 和
Vinyals, O. 理解深度学习（仍然）需要重新思考泛化。
*ACM*通讯，64(3):107– 115，2021年。

## A. 实验细节

在本附录中，我们提供了第8节所述合成实验的详细设置 .

### A.1. 任务与数据分布

我们构建了一个受控分类任务，旨在分离重尾分布对学习动态的影响。

- **词汇表：** 任务涉及一个由 $K = 1000$ 个不同词元组成的词汇表。
- **分布：** 词元频率从齐普夫分布中采样：

$$p_k = \frac{k^{-\alpha}}{Z_\alpha}, \quad \text{where } \alpha = 1.5 \text{ and } Z_\alpha = \sum_{j=1}^{K} j^{-\alpha}.$$

- **目标函数：** 模型学习严格独立的恒等映射。输入 $x$ 是对应于词元 $k$ 的独热向量 $e_k \in \mathbb{R}^K$，目标 $y$ 则是相同的向量 $e_k$。这确保假设 2.1（加性模式模型）完全成立，因为模式之间不存在语义干扰或迁移。

### A.2. 模型架构

我们采用一个带有瓶颈的两层神经网络，其定义如下：

$$f_\theta(x) = W_2 \cdot \text{ReLU}(W_1 x + b),$$

其中：

- $W_1 \in \mathbb{R}^{N \times K}$ 将独热输入映射到隐藏维度。
- $b \in \mathbb{R}^N$ 是偏置项。
- $W_2 \in \mathbb{R}^{K \times N}$ 将隐藏表示映射回输出词汇空间。
- $N$ 表示**模型容量**，我们在模型缩放实验中对其进行调整。

**初始化：** 权重 $W_1$，$W_2$ 从正态分布 $\mathcal{N}(0,\ 0.1_2)$中初始化。偏置 $b$ 初始化为零。

### A.3. 训练与指标

**损失函数**。 我们最小化预测逻辑值与目标独热向量之间的均方误差（MSE）。总损失定义如下：

$$\mathcal{L}(\theta) = \mathbb{E}_{k \sim p}\left[\|f_\theta(e_k) - e_k\|_2^2\right].$$

为与我们对可约损失的理论定义（公式 1）保持一致，我们跟踪<style id='1'>加权可约损失：

$$\Delta L = \sum_{k=1}^{K} p_k \underbrace{(f_\theta(e_k)_k - 1)^2}_{q_k}.$$

其中，$q_k$ 衡量了预测正确类别逻辑值的失败程度（在初始化时归一化为 1，在掌握时归一化为 0）。

**优化**。**所有模** 型均使用随机梯度下降（SGD）进行训练，超参数如下：

- 学习率 $\eta$：0。1
- 动量：0。9
- 批量大小：64

### A.4. 缩放机制

我们进行了三组实验来验证缩放定律。图5中报告的指数是通过对数变换数据进行线性回归得出的（$\log \Delta L$ 与 $\log\ R$）。

1. **模型缩放（$N$）：**
   - 变量：隐藏维度 $N$ 的范围为 10 到 1000（按对数间隔取值）。
   - 固定资源：无限数据极限近似（通过重采样使数据集大小 $D$ 实际上无限），训练步数 $\tau = 10,\ 000$。

2. **数据缩放（$D$）：**
   - 变量：数据集大小 $D$ 从 1,000 变化到 $100,000$。 $D$ 决定训练开始前从 $p_k$ 抽取的样本数量。
   - 固定资源：模型容量 $N = 2000$ （过参数化机制， $N > K$），训练步数 $\tau = 10 \times \lceil D/B \rceil$ （10 个训练轮次）。

3. **计算缩放（$\tau$）：**
   - 变量：优化步数 $\tau$ 从 $10_3$ 变化到 $10_5$。
   - 固定资源：模型容量 $N = 2000$，数据集大小 $D = \infty$ （重采样批次）。

### A.5. 环境设置

所有实验均使用 Python 和 PyTorch 框架实现。为确保第8节报告的经验值结果可复现，我们在下文详细列出所用硬件的具体规格和软件版本。

**硬件配置。** 实验在一台配备以下 GPU 加速设备的单节点 Linux 工作站上进行：

- **GPU：** NVIDIA GeForce RTX 5090（32,607 MiB 显存）
- **驱动程序：** NVIDIA 驱动程序 580.95.05 / CUDA 13.0

**软件栈。** 表1列出了所使用的具体库版本。

*表1。* 实验中使用的软件和库版本。

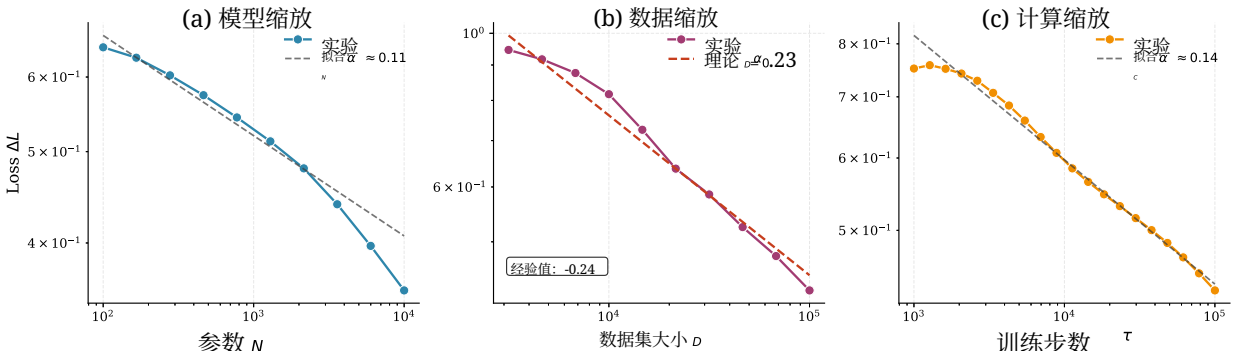
| 库 | 版本 | 库 | 版本 |
|---|---|---|---|
| Python | 3.11.8 | SciPy | 1.16.3 |
| PyTorch | 2.9.0+cu130 | Matplotlib | 3.9.0 |
| NumPy | 2.3.4 | Pandas | 2.2.2 |
| scikit-learn | 1.8.0 | | |

## B. 补充实验验证

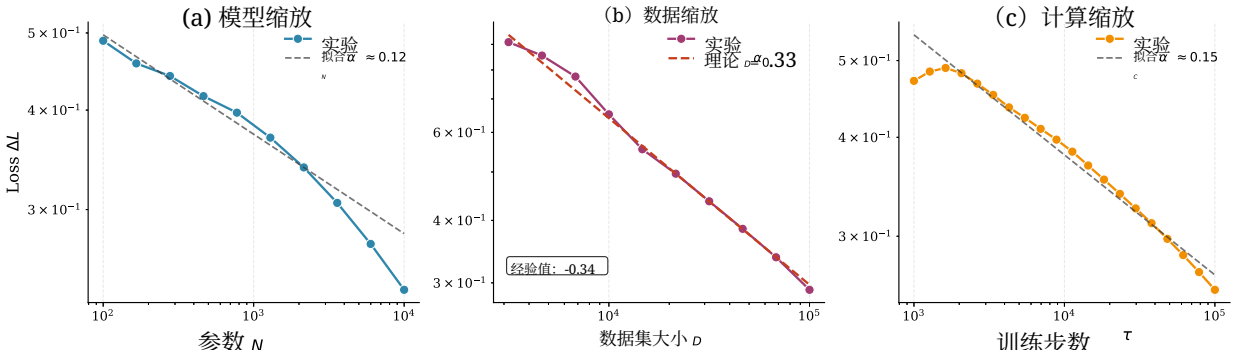
在本附录中，我们对第8节中给出的缩放指数进行详细的定量分析，并将验证范围扩展到更广泛的数据分布。这一分析进一步验证了理论推导的精确性，以及所提出框架的普适性。

### B.1. 详细缩放轨迹

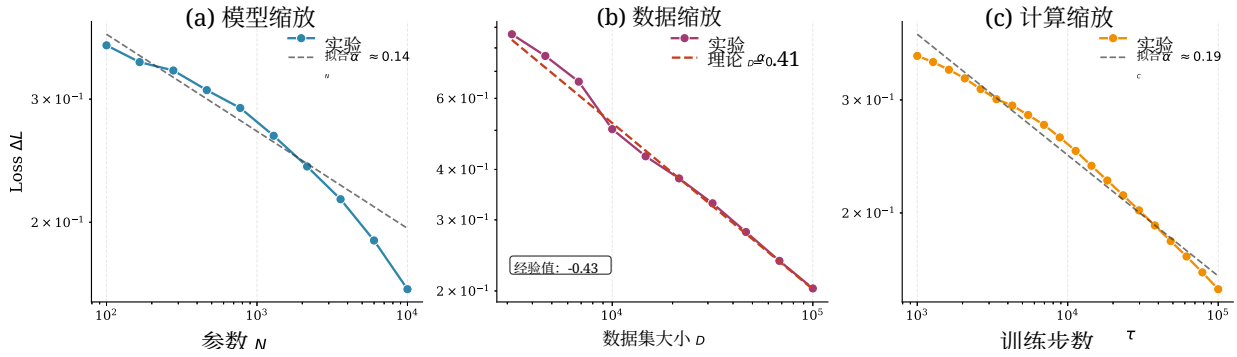
为展示研究结果的稳健性，我们绘制了尾指数为 $\alpha \in \{1.3, 1.5, 1.7, 1.9, 2.1\}$ 的齐夫分布在三个资源维数（模型容量 $N$、数据规模 $D$、计算量 $\tau$）下的宏观损失缩放情况。图6明确展示了这些轨迹。




*(a)* 尾指数 $\alpha = 1.3$

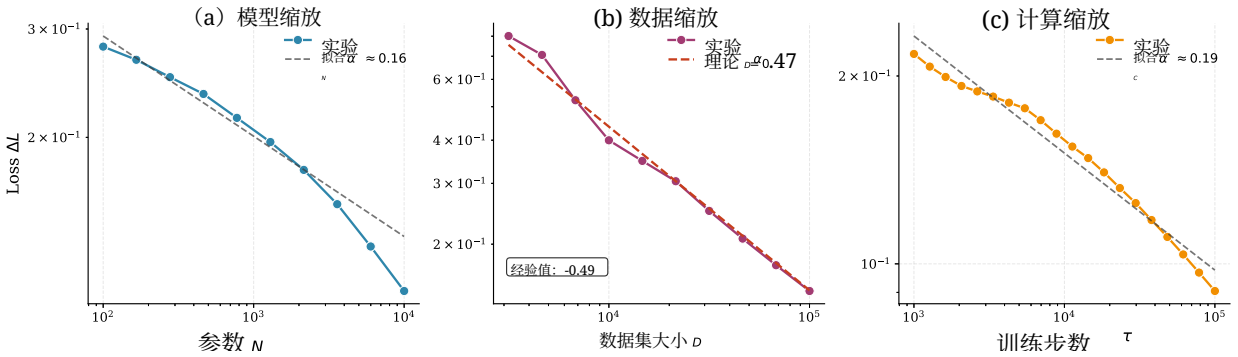



*(b)* 尾指数 $\alpha = 1.5$

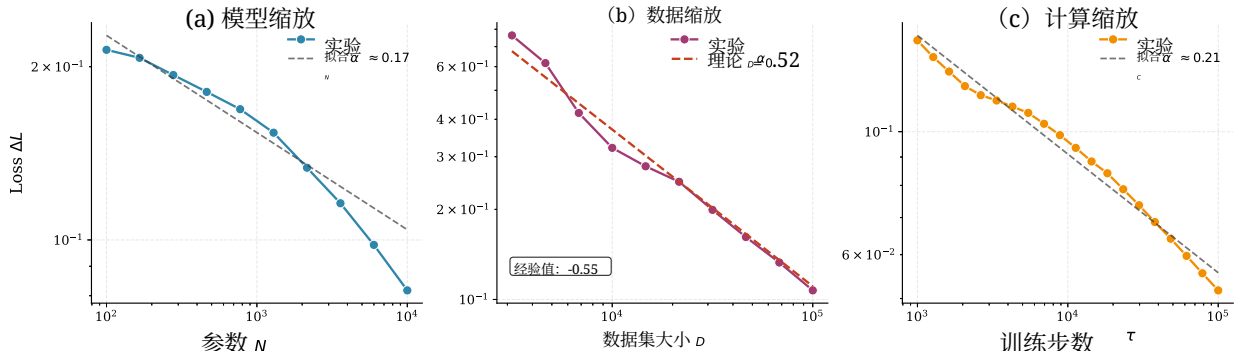



*(c)*尾指数 $\alpha = 1.7$

图*6.* **不同数据分布下的通用缩放定律（第1部分）**。我们绘制了可约损失 $\Delta L$ 与模型参数 $N$（左）、数据集大小 $D$（中）和训练步数 $\tau$（右）的关系，并指出，对于固定的模型配置，$C \propto \tau$。（续见下一页.....）




*(d)* 尾指数 $\alpha = 1.9$




*(e)*尾指数 $\alpha = 2.1$

图*6.* **不同数据分布下的通用缩放定律（第2部分）**。（续前页）。虚线表示双对数尺度上的最佳线性拟合。结果一致表明，随着尾部变轻（$\alpha$变大），数据缩放效率会提高（斜率变得更陡），这与我们的理论预测完全吻合。

### B.2. 数据缩放预测的精度

我们这项工作的一项核心贡献是对数据缩放定律（定理 5.2）的解析推导，该定律主张缩放指数 $\alpha_D$ 仅由数据分布的尾指数 $\alpha$决定：

$$\alpha_D^{\text{theory}} = \frac{\alpha - 1}{\alpha}. \tag{5}$$

在表2中，我们将这一理论预测与从图6（中间面板）提取的经验测量指数进行比较。我们观察到，理论与实验之间的平均绝对误差（平均绝对误差）低得惊人（$\approx 0.015$），从而证实有效覆盖前沿假设在广泛的尾部行为范围内严格成立。

表*2.* 理论与经验数据缩放指数的比较。这些经验值由图6(b)中的最佳拟合斜率得出。

| 尾指数 $\alpha$ | 预测 $\frac{\alpha-1}{\alpha}$ | 经验值 $\alpha_D$ | 错误 |
|---|---|---|---|
| 1.30 | 0.231 | 0.240 | 0.009 |
| 1.50 | 0.333 | 0.340 | 0.007 |
| 1.70 | 0.412 | 0.430 | 0.018 |
| 1.90 | 0.474 | 0.490 | 0.016 |
| 2.10 | 0.524 | 0.550 | 0.026 |

### B.3. 隐式偏置的一致性（$\beta$）

我们的框架引入隐式偏置指数 $\beta$，将其作为连接几何前沿 $k_\star$ 与计算扩展定律的拓扑不变量。我们使用从图6中提取的斜率进行交叉验证 $\beta$。

根据定理 6.6，计算缩放指数为 $\alpha_C = \frac{\alpha-1}{\alpha\beta}$。这使我们能够通过以下关系，根据实测计算缩放斜率 $\hat{\alpha}_C$（图6，右侧面板）推断优化隐式偏置 $\beta$：

$$\beta_{\text{inferred}} = \frac{\alpha - 1}{\alpha \cdot \hat{\alpha}_C}.$$

将其应用于我们在 $\alpha$整个范围内获得的实验结果：

- 对于 $\alpha = 1.3$($\hat{\alpha}_C \approx 0.14$)：$\beta \approx \frac{0.231}{0.14} \approx 1.65$。
- 对于 $\alpha = 1.5$($\hat{\alpha}_C \approx 0.15$)：$\beta \approx \frac{0.333}{0.15} \approx 2.22$。
- 对于 $\alpha = 1.7$($\hat{\alpha}_C \approx 0.19$)：$\beta \approx \frac{0.412}{0.19} \approx 2.17$。
- 对于 $\alpha = 1.9$($\hat{\alpha}_C \approx 0.19$)：$\beta \approx \frac{0.474}{0.19} \approx 2.49$。
- 对于 $\alpha = 2.1$($\hat{\alpha}_C \approx 0.21$)：$\beta \approx \frac{0.524}{0.21} \approx 2.49$。

推断出的数值始终集中在范围 $\beta \in [1.65, 2.50]$ 内，均值约为 2.2。这与几何测量（第 8.2 节中的 $\beta_{geo} \approx 2.03$）高度吻合，并支持由深度线性网络分析（假设 6.4）推导出的理论预测 $\beta \approx 2$。不同分布形状下 $\beta \approx 2$ 的一致性表明，二次频率偏置是深度神经网络上 SGD 动力学的一项基本且不变的性质。

## C. 定理 3.3 的证明

在本附录中，我们给出普适缩放原理（定理 3.3）的完整数学推导。其核心思路依赖于几何约化：尖锐有效前沿的存在，使我们能够将复杂的学习动态近似为秩空间中的阶跃函数，从而将损失计算简化为对齐夫分布进行尾和估计。

### C.1. 几何设置与辅助引理

有效前沿 $k_\star(R)$ 的定义（定义 3.2）意味着，从“已学习”（$q_k \approx 0$）到“未学习”（$q_k \approx 1$）的转变发生在一个窗口内，该窗口相对于 $k_\star$ 的尺度渐近可忽略。

我们首先通过定义**阶跃轮廓**来形式化这一点。该引理证明了我们在启发式论证中使用的 $0/1$ 近似是合理的。

**引理 C.1**（阶跃轮廓）。 在假设 *3.1*（贪心学习偏差）下，对于任意固定的 $\epsilon \in$（0，1），存在一个资源阈值 $R_0$（$\epsilon$），使得对于所有 $R \geq R_0$（$\epsilon$）：

$$q_k(R) \leq \delta \quad \textit{for } k \leq (1-\epsilon)\, k_\star(R) \quad \textit{(Mostly Learned)},$$
$$q_k(R) \geq 1-\delta \quad \textit{for } k \geq (1+\epsilon)\, k_\star(R) \quad \textit{(Mostly Unlearned)}.$$

证明。 根据定义 3.2，转移边界满足 $k_-(R) = (1 - o_R(1))k_\star(R)$ 和 $k_+(R) = (1 + o_R(1))k_\star(R)$。对于任意固定的 $\epsilon > 0$，当 $R \to \infty$ 时，相对误差 $o_R(1)$ 会变得小于 $\epsilon$。因此，对于充分大的 $R$：

$$k_-(R) \geq (1-\epsilon)k_\star(R) \quad \text{and} \quad k_+(R) \leq (1+\epsilon)k_\star(R).$$

根据 $k_-$ 的定义和秩单调性（假设 3.1），如果 $k \leq (1-\epsilon)k_\star(R)$，那么 $k \leq k_-(R)$，这意味着$q_k(R) \leq \delta$。类似地，如果 $k \geq (1+\epsilon)k_\star(R)$，那么 $k \geq k_+(R)$，这意味着 $q_k(R) \geq 1-\delta$。 □

接下来，我们证明**夹逼界**。这一结果至关重要，因为它将对未知残差 $\sum p_k q_k$ 求和的问题转化为对已知概率 $\sum p_k$求和的问题。

为使用紧凑记法，我们定义转移边界：

$$K_+(R) := (1+\epsilon)k_\star(R), \qquad K_-(R) := (1-\epsilon)k_\star(R).$$

**引理 C.2 (**夹逼界**)**。 根据假设 *3.1*，当 $R$足够大时，可约损失 $\Delta L(R)$ 由尾部质量界定：

$$(1-\delta)\sum_{k>K_+(R)} p_k \;\le\; \Delta L(R) \;\le\; \delta \sum_{k\le K_-(R)} p_k \;+\; \sum_{k>K_-(R)} p_k. \tag{6}$$

证明。 回顾 $\Delta L(R) = \sum_{k\ge 1} p_k q_k(R)$。根据引理 C.1，对于足够大的 $R$，残差分布在过渡窗口（$K_-, K_+$）之外被有效地确定。

**下界：**我们限 制损失求和于"未学习"尾部区域 $k > K_+(R)$。由于所有项均为非负：

$$\Delta L(R) = \sum_{k=1}^{\infty} p_k q_k(R) \ge \sum_{k>K_+(R)} p_k q_k(R).$$

利用引理 C.1，对于 $k > K_+$，我们有 $q_k(R) \ge 1-\delta$。因此：

$$\Delta L(R) \ge \sum_{k>K_+(R)} p_k(1-\delta) = (1-\delta)\sum_{k>K_+(R)} p_k.$$

**上界：**我们在"已学习"边界 $K_-(R)$ 处分割损失求和：

$$\Delta L(R) = \sum_{k\le K_-(R)} p_k q_k(R) + \sum_{k>K_-(R)} p_k q_k(R).$$

对于第一项（头部），引理 C.1 保证了 $q_k \le \delta$。对于第二项（尾部 +过渡），我们使用平凡上界 $q_k \le 1$。因此：

$$\Delta L(R) \le \sum_{k\le K_-(R)} p_k(\delta) + \sum_{k>K_-(R)} p_k(1) = \delta \sum_{k\le K_-(R)} p_k + \sum_{k>K_-(R)} p_k.$$

至此，公式（6）的证明完成。 □

**解读**。 引理 C.2 是关键的简化步骤：它表明，过渡窗口内学习曲线 $q_k(R)$ 的具体形状只会影响常数前因子。 $\Delta L(R)$ 的渐近缩放完全由未学习尾部 $\sum_{k\gtrsim k_\star} p_k$ 中的齐夫分布概率质量所决定。因此，推导缩放定律就成为一个纯粹的几何任务：确定每项资源 $R$ 如何将有效前沿 $k_\star(R)$推进到秩空间中。

### C.2. 定理 3.3（普适缩放原理）的证明

现在，我们将夹逼界与齐夫分布的具体性质相结合，以证明主要的缩放定理。

**定理 3.3（重述）**。 假设 $p_k = Z^{-1}k^{-\alpha}$ 且 $\alpha > 1$。如果资源 $R$ 诱导出有效前沿 $k_\star(R) \to \infty$，那么 $\Delta L(R) \asymp k_\star(R)^{-(\alpha-1)}$。

证明。 根据引理 C.2，我们有夹逼不等式。我们首先简化上界项。由于 $p_k$ 是一个具有无限支撑的概率分布，$\sum_{k>K_-} p_k > 0$。我们可以将常数项 $\delta$ 相对于尾和进行界定。具体而言，对于任意 $\alpha > 1$，尾和按多项式速率衰减，而 $\delta$ 是一个常数。然而，为了研究可约损失缩放（其趋于 0），我们关注变量分量。更严谨的代数运算表明，"学习质量"项可以忽略，或与之成比例。请注意， $\sum_{k\le K_-} p_k < 1$。因此，上界由尾和项主导：

$$\Delta L(R) \le \delta + \sum_{k>K_-(R)} p_k.$$

由于我们关注可约损失（趋于 0 的部分）的缩放指数，并假设 $\delta$ 可以任意小，或在更复杂的噪声模型中随损失缩放，因此我们关注尾和 $\sum_{k>K} p_k$。
只需估计齐夫尾和 $S(K) = \sum_{k>K} k^{-\alpha}$的渐近行为。

由于 $f(x) = x^{-\alpha}$ 对于 $x \geq 1$ 非负且单调递减，我们调用积分判别法的界：

$$\int_{K+1}^{\infty} x^{-\alpha}\, dx \;\leq\; \sum_{k=K+1}^{\infty} k^{-\alpha} \;\leq\; \int_{K}^{\infty} x^{-\alpha}\, dx.$$

计算 $\alpha > 1$ 的积分：

$$\int_{Y}^{\infty} x^{-\alpha}\, dx = \left[\frac{x^{-(\alpha-1)}}{1-\alpha}\right]_{Y}^{\infty} = \frac{1}{\alpha-1} Y^{-(\alpha-1)}.$$

**步骤 1：Lo 界估计**。 利用 $K = K_+ = \lceil (1+\epsilon) k_\star(R) \rceil$ 得到下夹逼界：

$$\begin{aligned}
\sum_{k>K_+} p_k &= Z^{-1} \sum_{k=K_++1}^{\infty} k^{-\alpha} \\
&\geq Z^{-1} \int_{K_++1}^{\infty} x^{-\alpha} dx \\
&= \frac{Z^{-1}}{\alpha-1} (K_+ + 1)^{-(\alpha-1)} \\
&\asymp \frac{Z^{-1}}{\alpha-1} \left((1+\epsilon) k_\star(R)\right)^{-(\alpha-1)} \\
&\asymp k_\star(R)^{-(\alpha-1)}.
\end{aligned}$$

将其代入夹逼下界：

$$\Delta L(R) \geq (1-\delta) \cdot (\text{const}) \cdot k_\star(R)^{-(\alpha-1)}.$$

**步骤 2：上界估计**。 利用 $K = K_- = \lfloor (1-\epsilon) k_\star(R) \rfloor$ 得到上夹逼界：

$$\begin{aligned}
\sum_{k>K_-} p_k &= Z^{-1} \sum_{k=K_-+1}^{\infty} k^{-\alpha} \\
&\leq Z^{-1} \int_{K_-}^{\infty} x^{-\alpha} dx \\
&= \frac{Z^{-1}}{\alpha-1} (K_-)^{-(\alpha-1)} \\
&\asymp \frac{Z^{-1}}{\alpha-1} \left((1-\epsilon) k_\star(R)\right)^{-(\alpha-1)} \\
&\asymp k_\star(R)^{-(\alpha-1)}.
\end{aligned}$$

将其代入夹逼上界（忽略不可约的 $\delta$ 下限，或将其视为总界的一部分）：

$$\Delta L(R) \lesssim \sum_{k>K_-} p_k \asymp k_\star(R)^{-(\alpha-1)}.$$

**结论**。 结合步骤 1 和步骤 2，我们得到了 $\Delta L(R)$ 的上下界，且两界均与$k_\star(R)^{-(\alpha-1)}$成正比。因此：

$$\Delta L(R) \asymp k_\star(R)^{-(\alpha-1)}.$$

□

## D. 数据缩放定律的推导

在本附录中，我们给出数据缩放定律（定理 5.2）的完整证明，并将其推广至$m$-次出现阈值（推论 5.4）。我们通过建立覆盖概率的严格界限，为正文中使用的有效前沿近似提供依据。

### D.1. 定理 5.2 的证明

我们的目标是证明，在齐夫分布 $p_k \propto k^{-\alpha}$ 下，可约损失按 $\Delta L(D) \asymp D^{-(\alpha-1)/\alpha}$ 缩放。证明分为三个步骤：(1) 为残差建立一个指数代理；(2) 证明这些求和的渐近等价性；(3) 通过有效前沿推导缩放定律。

**步骤 1：残差的指数界**。回顾一下，模式 $k$ 的覆盖诱导残差为 $q_k(D) = (1-p_k)^D$。我们首先证明该项的渐近行为为 $e^{-Dp_k}$。对于任意 $p \in [0,1)$，由标准不等式 $1+x \le e^x$ 可得上界：

$$(1-p)^D \le (e^{-p})^D = e^{-Dp}.$$

对于下界，我们考虑 $p \in [0, 1/2]$ 的情形。利用 $\log(1-p)$ 的泰勒展开，可得 $\log(1-p) \ge -p-p^2$。将其取指数后再提升至 $D$ 次幂，可得：

$$(1-p)^D \ge e^{-D(p+p^2)} = e^{-Dp}e^{-Dp^2}.$$

综合这些结果，我们得到残差的双侧控制：

$$e^{-Dp_k}e^{-Dp_k^2} \le (1-p_k)^D \le e^{-Dp_k}. \tag{7}$$

**步骤 2：求和层面等价**。我们现在证明，总损失 $\Delta L(D) = \sum_{k\ge1} p_k(1-p_k)^D$ 与代理和 $S(D) = \sum_{k\ge1} p_k e^{-Dp_k}$渐近等价。上界 $\Delta L(D) \le S(D)$ 直接由式 (7) 得出。对于下界，我们将指标集划分为“头部” $\mathcal{H} = \{k : p_k > D^{-1/2}\}$ 和“尾部” $\mathcal{T} = \{k : p_k \le D^{-1/2}\}$。

对于尾部 $\mathcal{T}$中的模式，我们有 $Dp_k^2 \le 1$。因此，式 (7) 中的阻尼因子以一个常数为界：$e^{-Dp_k^2} \ge e^{-1}$。这意味着对于稀有模式，精确的二项概率与指数代理在相差一个常数因子的意义下相匹配：

$$\sum_{k\in\mathcal{T}} p_k(1-p_k)^D \ge e^{-1}\sum_{k\in\mathcal{T}} p_k e^{-Dp_k}.$$

对于头部 $\mathcal{H}$中的模式，指数项以超指数速度衰减。具体而言，由于 $p_k > D^{-1/2}$，我们有 $Dp_k > \sqrt{D}$，从而 $\sum_{k\in\mathcal{H}} p_k e^{-Dp_k} \le e^{-\sqrt{D}}$，与损失的多项式衰减相比可以忽略不计。因此，总损失满足 $\Delta L(D) \asymp S(D)$，当 $D\to\infty$时。

**步骤 3：前沿推导与缩放**。在已确立 $\Delta L(D) \asymp \sum p_k e^{-Dp_k}$之后，我们确定有效前沿。当指数为一阶量级时，项 $e^{-Dp_k}$ 从约 0 过渡到 1：

$$Dp_{k_\star} \asymp 1.$$

代入齐夫定律 $p_k = Z^{-1}k^{-\alpha}$，我们求解前沿 $k_\star(D)$：

$$D\cdot Z^{-1}k_\star^{-\alpha} \asymp 1 \implies k_\star(D) \asymp D^{1/\alpha}.$$

最后，应用普适缩放原理（定理 3.3），损失由超出 $k_\star$的尾部质量主导：

$$\Delta L(D) \asymp k_\star(D)^{-(\alpha-1)} \asymp \left(D^{1/\alpha}\right)^{-(\alpha-1)} = D^{-\frac{\alpha-1}{\alpha}}.$$

定理 5.2 的证明到此结束。 □

### D.2. 推论 5.4 的证明

现在，我们将结果扩展到仅当模式出现至少 $m$ 次时才会被学会的情形。残差定义为$q_k^{(m)}(D) = \Pr[X_k < m]$，其中 $X_k \sim \mathrm{Bin}(D,\ p_k)$。我们证明，这会使有效前沿的位置改变$m$ 倍，但保留缩放指数。

令 $\lambda_k = Dp_k$ 为期望出现次数。二项分布的累积分布函数可以利用类泊松尾界进行界定。通过分析 $q_k^{(m)}$ 在 $\lambda_k$ 较小和较大时的行为，我们定义有效转变窗口。

**上界（软截断）**。对于 $p_k \le 1/2$ 的尾部区域，我们对观测到少于 $m$ 个样本的概率进行界定。利用二项分布的集中不等式（具体而言，对展开式中前 $m$ 项的和进行界定），可以证明，对于 $\lambda_k \ge 2m$：

$$q_k^{(m)}(D) \le e^{-\lambda_k} \sum_{j=0}^{m-1} \frac{(2\lambda_k)^j}{j!}.$$

这表明，当期望计数 $\lambda_k$ 显著超过 $m$ 时，残差呈指数衰减。

**下界与阈值化**。反过来，我们为"未学习"区域定义一个阈值。利用指数矩方法，对于任意 $t > 1$，$\Pr[X_k \ge m] \le \mathbb{E}[t^{X_k}]t^{-m}$。由此得到以下界：

$$\Pr[X_k \ge m] \le \left(\frac{e\lambda_k}{m}\right)^m.$$

若要将某种模式视为"基本尚未学会"（即 $\Pr[X_k \ge m] \le \delta$），只需满足：对于某个与 $m$ 成正比的常数 $c$（$m$），$\lambda_k \le c(m)$。具体而言，当期望出现次数达到阈值 $m$ 的量级时，模式便会从尚未学会转变为已学会：

$$\lambda_{k_\star} = Dp_{k_\star} \asymp m.$$

**缩放推导**。 将齐夫分布 $p_k \propto k^{-\alpha}$ 代入条件 $Dp_{k_\star} \asymp m$：

$$D \cdot k_\star^{-\alpha} \asymp m \implies k_\star(D) \asymp \left(\frac{D}{m}\right)^{1/\alpha}.$$

可约损失由这一新前沿 $k_\star$ 之后的尾部质量决定。应用定理 3.3：

$$\Delta L(D) \asymp k_\star(D)^{-(\alpha-1)} \asymp \left[\left(\frac{D}{m}\right)^{1/\alpha}\right]^{-(\alpha-1)} \propto D^{-\frac{\alpha-1}{\alpha}}.$$

因此，尽管要求出现 $m$ 次会延迟特定模式的学习（使前沿移动一个常数因子 $m^{1/\alpha}$），但相对于数据集大小 $D$ 的渐近幂律指数仍严格为 $\frac{\alpha-1}{\alpha}$。□

## E. 计算缩放的推导

本节将为第 6 节提出的计算缩放定律提供完整的数学框架。本附录分为两部分：

1. **微观推导（E.1 节）：** 我们通过分析 PL 条件和频谱偏置下逐模式残差的演化，推导出随机梯度下降（SGD）的具体缩放定律。
2. **普适性与不变性（E.2 节）：** 我们证明了一般缩放定理，并通过梅林变换表明，幂律指数是"有效前沿"的拓扑不变量。

### E.1. 第一部分：随机梯度下降的微观推导

我们在解耦下降与频率依赖的谱偏置的假设下，分析逐模式残差 $q_k(\tau)$ 的优化动力学。

E.1.1. 预备知识与符号约定

设 $L(\theta) = \sum_{k=1}^{\infty} p_k \ell_k(\theta)$。我们将逐模式残差定义为 $q_k^{(t)} := \ell_k(\theta_t) - \ell_k^\star$，与 $\Delta L = \sum p_k q_k$ 一致。

随机梯度下降更新为 $\theta_{t+1} = \theta_t - \eta g_t$。我们考虑单样本机制，其中索引 $K_t$ 以$P(K_t = k) = p_k$ 进行采样，这意味着 $\mathbb{E}[g_t] = \sum p_k \nabla \ell_k(\theta)$。

E.1.2. 结构性假设

为使分析易于处理，我们依赖于以下关于损失景观的几何假设：

**假设 E.1.** 1. **近似正交性：**不同原子模式的梯度近乎正交。对于$j \neq k$:

$$\langle \nabla \ell_j(\theta), \nabla \ell_k(\theta) \rangle \approx 0.$$

2. **逐模式 PL 条件：** 每个原子模式 $\ell_k$ 都在局部满足 Polyak-Łojasiewicz 不等式，其中 $\lambda_k$ 为条件数：

$$\|\nabla \ell_k(\theta)\|^2 \geq 2\lambda_k(\ell_k(\theta) - \ell_k^\star) = 2\lambda_k q_k(\theta).$$

E.1.3. 单步动力学分析

我们分析给定 $\theta_t$ 时，期望残差 $\mathbb{E}[q_k^{(t+1)}]$ 的演变。对采样的模式索引 $K_t$进行条件化后，有两种互不相交的情况：

**情况 I：采样到模式 $k$ （$K_t = k$）**
在这种情况下（概率为 $p_k$），更新使用目标模式的梯度，即 $g_t = \nabla \ell_k(\theta_t)$。假设 $\ell_k$ 是$\beta_L$-光滑的：

$$\ell_k(\theta_{t+1}) \leq \ell_k(\theta_t) - \eta \|\nabla \ell_k(\theta_t)\|^2 + \frac{\beta_L \eta^2}{2} \|\nabla \ell_k(\theta_t)\|^2.$$

当学习率足够小（$\eta\beta_L < 1$）时，一阶下降占主导。减去 $\ell_k^\star$ 并应用 PL 条件（$-\|\nabla \ell_k\|^2 \leq -2\lambda_k q_k$）：

$$q_k^{(t+1)} \leq q_k^{(t)} - \eta(1 - \frac{\eta\beta_L}{2}) \|\nabla \ell_k(\theta_t)\|^2 \leq (1 - 2\eta\lambda_k + \beta_L \eta^2) q_k^{(t)},$$

我们进一步简化了其形式：

$$q_k^{(t+1)} \leq (1 - 2\eta' \lambda_k) q_k^{(t)}.$$

其中 $\eta' = 2\eta \left(1 - \frac{\beta_L \eta}{2}\right)$

**情况 II：采样其他模式 $j$ （$K_t = j \neq k$）**
I在该事件中（概率为 $1 - p_k$），更新由 $g_t = \nabla \ell_j(\theta_t)$ 驱动。使用泰勒展开 ：

$$\ell_k(\theta_{t+1}) = \ell_k(\theta_t - \eta \nabla \ell_j) \approx \ell_k(\theta_t) - \eta \langle \nabla \ell_k, \nabla \ell_j \rangle.$$

根据正交性假设，$\langle \nabla \ell_k, \nabla \ell_j \rangle \approx 0$。因此，学习模式 $j$ 意味着模式 $k$的更新可以忽略：

$$q_k^{(t+1)} \approx q_k^{(t)}.$$

E.1.4. 聚合动力学与频率偏差

综合这两种情况，第 $t+1$ 步的期望残差为：

$$\begin{aligned} \mathbb{E}[q_k^{(t+1)}] &= p_k \cdot (1 - 2\eta' \lambda_k) q_k^{(t)} + (1 - p_k) \cdot q_k^{(t)} \\ &= (1 - 2\eta' \lambda_k p_k) q_k^{(t)}. \end{aligned}$$

将此递推关系应用于 $\tau$ 步，并近似计算 $(1 - x)^\tau \approx e^{-\tau x}$：

$$q_k(\tau) \approx \exp\left(-2\eta' \lambda_k p_k \tau\right).$$

**假设：Spect ral Bias.** 我们采用有效收敛率取决于频率这一假设 :

$$\lambda_k = \lambda_0 p_k^{\beta-1}, \quad \text{where } \beta > 0.$$

将其代入公式 (E.1.4)，并定义集总常数 $c = 2\eta'\lambda_0$:

$$\boxed{q_k(\tau) \approx \exp\left(-c\tau p_k^{\beta}\right)}.$$

这证实了第 6 节讨论的随机梯度下降的特定核形式。

## E.2. 第二部分：普适缩放定理的证明

在第 6.2 节中，我们引入了普适计算缩放定律（定理 6.8）。在此，我们给出严格证明，说明缩放指数 $s = \frac{\alpha-1}{\alpha\beta}$ 是有效前沿的拓扑不变量，不依赖于优化核 $g(\cdot)$ 的具体选择，前提是它满足基本的可积性条件。

### E.2.1. 损失的积分近似

我们从广义分解拟设开始。令模式 $k$ 的残差按照单调缩放核 $g$演化:

$$q_k(\tau) \approx g\left(a\tau p_k^{\beta}\right),$$

其中 $p_k = Z^{-1}k^{-\alpha}$。总可约损失近似为秩空间 $k \in [1, \infty$上的积分 ):

$$\Delta L(\tau) \approx \int_1^{\infty} p_k\, g(a\tau p_k^{\beta})\, dk = \int_1^{\infty} \frac{1}{Z} k^{-\alpha} g\left(a\tau (Z^{-1}k^{-\alpha})^{\beta}\right) dk.$$

令 $A = aZ^{-\beta}$ 为集总速率常数。核的自变量为 $u(k) = A\tau k^{-\alpha\beta}$。

### E.2.2. 变量代换（几何不变性）

为将时间尺度 $\tau$ 与几何结构分离，我们进行变量代换 $u = A\tau k^{-\alpha\beta}$。解出 $k$，得:

$$k = (A\tau)^{\frac{1}{\alpha\beta}} u^{-\frac{1}{\alpha\beta}}.$$

微分变换为:

$$dk = -\frac{1}{\alpha\beta}(A\tau)^{\frac{1}{\alpha\beta}} u^{-\frac{1}{\alpha\beta}-1} du.$$

积分限变换如下:

- 下限 $k = 1 \implies u_{max} = A\tau$。
- 上限 $k \to \infty \implies u_{min} \to 0$。

将这些代入积分:

$$\Delta L(\tau) \approx \frac{1}{Z}\int_{A\tau}^{0} \underbrace{\left[(A\tau)^{\frac{1}{\alpha\beta}} u^{-\frac{1}{\alpha\beta}}\right]^{-\alpha}}_{k^{-\alpha}} \cdot g(u) \cdot \underbrace{\left(-\frac{1}{\alpha\beta}(A\tau)^{\frac{1}{\alpha\beta}} u^{-\frac{1}{\alpha\beta}-1}\right) du}_{dk}$$

$$= \frac{1}{\alpha\beta Z}\int_0^{A\tau} \left[(A\tau)^{-\frac{\alpha}{\alpha\beta}} u^{\frac{\alpha}{\alpha\beta}}\right] g(u) \left[(A\tau)^{\frac{1}{\alpha\beta}} u^{-\frac{1}{\alpha\beta}-1}\right] du.$$

现在，我们将时间相关项 $(A\tau)$ 和积分变量 $u$ 的幂次合并。

**1. 缩放指数（时间依赖性）：** $\tau$ 的总指数严格由齐夫尾部 $\alpha$ 和优化器偏差 $\beta$决定:

$$-\frac{\alpha}{\alpha\beta} + \frac{1}{\alpha\beta} = -\frac{\alpha-1}{\alpha\beta} := -s.$$

**2. 梅林核（系数）：** 积分变量 $u$ 的指数为:

$$\frac{\alpha}{\alpha\beta}-\frac{1}{\alpha\beta}-1=\frac{\alpha-1}{\alpha\beta}-1=s-1.$$

将这些代回后，表达式恰好可以分解为:

$$\Delta L(\tau)\approx\left[\frac{A^{-s}}{\alpha\beta Z}\right]\cdot\tau^{-s}\cdot\int_0^{A\tau}u^{s-1}g(u)du.$$

E.2.3. 收敛条件（取代“三大支柱”）

当 $\tau\to\infty$时，上限为 $A\tau\to\infty$。要使缩放定律成立（即损失有限且性质良好），积分项必须收敛至一个常数。该积分是核 $g$的**梅林变换**，记为 $\mathcal{M}[g](s)$。
条件 $\mathcal{M}[g](s)\ <\ \infty$ 对学习动态施加了两个物理约束，取代了文献中常用的启发式论证:

1. **尾部可积性($u\to 0$)：** 该区域对应于 $k\to\infty$ （稀有模式）。由于 $g(0)\ =1$ （未学习），被积函数的行为为 $u^{s-1}$。收敛要求 $s>0$，这意味着 $\alpha>1$。这从形式上证明，缩放定律仅适用于具有有限均值的重尾分布。

2. **头部衰减 ($u\to\infty$)：** 此情形对应于 $k\to 1$ （高频模式）。我们要求 $g(u)$ 在较大的 $u$ 下比 $u^{-s}$ 衰减得更快。从物理角度看，这意味着优化器必须能够有效地最小化高频模式的损失。

E.2.4. 最终结果

假设 这些条件成立，我们便可得到定理6中陈述的普适计算缩放定律。 8:

$$\Delta L(\tau)\sim\mathcal{K}\cdot\tau^{-s},$$

其中，前因子 $\mathcal{K}$ 通过梅林变换囊括了所有算法细节:

$$\mathcal{K}=\frac{1}{\alpha\beta}\left(\frac{1}{Z}\right)^{1/\alpha}a^{-s}\mathcal{M}[g](s).$$

证明至此完成。

## F. 联合缩放的证明与推导

在本附录中，我们为组合定律（命题 7.1）提供严谨的数学证明，并详细推导第 7 节讨论的最优缩放前沿。

### F.1. 最大瓶颈原理的证明

我们希望证明，联合可约损失 $\Delta L(N,\ D,\ \tau)$ 渐近上由各个单独资源瓶颈中的最大值所主导:

$$\Delta L(N,D,\tau)\asymp\max\left(\varepsilon_N(N),\ \varepsilon_D(D),\ \varepsilon_\tau(\tau)\right).$$

证明。 **下界：** 由于模型容量 $N$、数据集大小 $D$ 和训练计算量 $\tau$ 对模式秩空间施加了彼此不同且必不可少的约束，联合损失不可能低于任何单一瓶颈所设定的极限。形式化地说:

$$\Delta L\geq\varepsilon_N(N),\quad\Delta L\geq\varepsilon_D(D),\quad\Delta L\geq\varepsilon_\tau(\tau).$$

由此可直接得出:

$$\Delta L\geq\max\left(\varepsilon_N(N),\ \varepsilon_D(D),\ \varepsilon_\tau(\tau)\right).$$

**上界：** 考虑非负数 $x, y, z$ 的标准不等式：$\max(x,y,z) \le x+y+z \le 3\max(x,y,z)$。我们假设存在一个构造性上界，使总残差至多为各机制产生的残差之和（例如，如果某个模式不满足任一标准，即容量不足、覆盖度不足或优化不足，则认为该模式未被学习）：

$$\Delta L \lesssim \varepsilon_N(N) + \varepsilon_D(D) + \varepsilon_\tau(\tau).$$

应用最大值—求和不等式：

$$\varepsilon_N + \varepsilon_D + \varepsilon_\tau \le 3\max\left(\varepsilon_N, \varepsilon_D, \varepsilon_\tau\right).$$

由于常数因子（如 3）不会影响渐近缩放定律（幂律指数），因此我们得出：

$$\Delta L \asymp \max\left(\varepsilon_N(N),\ \varepsilon_D(D),\ \varepsilon_\tau(\tau)\right).$$

□

### F.2. 最优计算点（转折点分析）

对于由参数 $N$ 和数据集 $D$ 定义的固定模型配置，学习曲线是训练步数 $\tau$ 的函数。我们将固定资源施加的极限定义为**静态瓶颈**：

$$\varepsilon_{\text{stat}}(N, D) = \max\left(\varepsilon_N(N),\ \varepsilon_D(D)\right).$$

**定义 F.1**（最优计算点）。最优计算点 $\tau_\star$ 定义为动态收敛瓶颈与静态瓶颈相交时的训练步数。这代表了从优化受限区间向容量/数据受限区间的转变。

从数学上讲，这发生在：

$$\varepsilon_\tau(\tau_\star) \asymp \varepsilon_{\text{stat}}(N, D).$$

代入正文中推导出的幂律形式 $\varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau}$：

$$\begin{aligned} G\tau_\star^{-\alpha_\tau} &\asymp \varepsilon_{\text{stat}} \\ \tau_\star^{-\alpha_\tau} &\asymp \frac{\varepsilon_{\text{stat}}}{G} \\ \tau_\star &\asymp \left(\frac{\varepsilon_{\text{stat}}}{G}\right)^{-1/\alpha_\tau}. \end{aligned}$$

达到该点所需的计算预算（浮点运算次数）由 $C_\star \approx 6N\tau_\star$ 给出。

### F.3. 最优缩放前沿的推导

我们现在通过求解联合损失函数 $\Delta L$ 的约束优化问题来推导最优缩放定律。我们分析两种不同的制度，它们分别对应于 Kaplan 等人（2020）和 Hoffmann 等人（2022）所做的假设。

#### F.3.1. 制度 A：收敛受限（Kaplan 缩放）

我们假设处于“数据充裕”极限，即数据集大小 $D$ 实际上是无限的（或可无惩罚地重复使用（Yan 等人，2025）），因此 $\varepsilon_D$ 可以忽略不计。该优化问题需要在模型规模 $N$ 与训练步数 $\tau$之间权衡。

1. **设定：** 在固定计算预算 $C$（浮点运算次数）的约束下，将损失降至最低。
   - 目标函数：$\min_{N,\ \tau} \max\left(AN^{-\alpha_N},\ \ G\tau^{-\alpha_\tau}\right)$
   - 约束：$C \approx 6N\tau \implies \tau \approx \frac{C}{6N}$。—
2. **代入：** 将 $\tau$ 代入目标函数后，可得一个仅关于 $N$ 的函数：

$$\Psi(N) = \max\left(AN^{-\alpha_N},\ G\left(\frac{C}{6N}\right)^{-\alpha_\tau}\right).$$

将常数吸收到 $G'$ 中后，可简化为 ：

$$\Psi(N) \asymp \max\left(AN^{-\alpha_N},\ G'C^{-\alpha_\tau}N^{\alpha_\tau}\right).$$

3. 均衡： </style}项 $N^{-\alpha_N}$关于 $N$严格递减，而 $N^{\alpha_\tau}$严格递增。最大值函数恰好在两个瓶颈达到平衡的交点处取最小值：

$$AN^{-\alpha_N} \asymp G'C^{-\alpha_\tau}N^{\alpha_\tau}.$$

4. **解：** 求解 $N$：

$$\begin{aligned} N^{-\alpha_N} \cdot N^{-\alpha_\tau} &\asymp \frac{G'}{A}C^{-\alpha_\tau} \\ N^{-(\alpha_N+\alpha_\tau)} &\asymp \text{const} \cdot C^{-\alpha_\tau} \\ N &\propto C^{\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}}. \end{aligned}$$

这得出了 Kaplan 等人（2020）提出的缩放定律形式。
注释 F.2（损失缩放）。 将 $N_{opt}$ 代回主导项 $N^{-\alpha_N}$，可得最优损失缩放：

$$\Delta L_{opt}(C) \asymp \left(C^{\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}}\right)^{-\alpha_N} = C^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}}.$$

F.3.2. 区间 B：数据受限（Chinchilla 缩放）

我们假设处于“数据受限”区间，在该区间中，我们针对留出分布优化，以获得最低损失，而唯一词元是限制因素。此处假设 $\tau$ 足以覆盖 $D$ （即一个训练轮次），因此需要在模型规模 $N$和数据集大小 $D$之间权衡。

1. **设定：** 在固定计算预算 $C$ （浮点运算次数）下最小化损失。
   - 目标函数：$\min_{N,\ D}\max\left(AN^{-\alpha_N},\ \ BD^{-\alpha_D}\right)$
   - 约束： $C \approx 6ND \implies D \approx \frac{C}{6N}$。—

2. **代入：** 将 $D$ 代入目标函数：

$$\Phi(N) = \max\left(AN^{-\alpha_N},\ B\left(\frac{C}{6N}\right)^{-\alpha_D}\right).$$

   将常数并入 $B'$：

$$\Phi(N) \asymp \max\left(AN^{-\alpha_N},\ B'C^{-\alpha_D}N^{\alpha_D}\right).$$

3. **均衡：** 最小值出现在容量瓶颈与覆盖瓶颈的交点处：

$$AN^{-\alpha_N} \asymp B'C^{-\alpha_D}N^{\alpha_D}.$$

4. **解答：** 求解 $N$：

$$\begin{aligned} N^{-\alpha_N} \cdot N^{-\alpha_D} &\asymp \frac{B'}{A}C^{-\alpha_D} \\ N^{-(\alpha_N+\alpha_D)} &\asymp \text{const} \cdot C^{-\alpha_D} \\ N &\propto C^{\frac{\alpha_D}{\alpha_N+\alpha_D}}. \end{aligned}$$

这得出了 Hoffmann 等人（2022）提出的计算最优缩放定律。

## G. 相关工作

神经缩放定律描述了随着模型规模、数据集大小和计算量的增长，测试损失呈现稳定的幂律改进（Hestness 等人，2017；Kaplan 等人（2020））。虽然测试损失通常会平滑改善，但 Wei 等人（2022）观察到，特定的下游能力往往会随着规模扩大呈现出突然而显著的涌现式相变。一个核心的实际挑战在于如何定义计算最优前沿：Kaplan 等人（2020）主张优先进行参数缩放，而 Hoffmann 等人（2022）（Chinchilla）后来证明，若要达到最优效果，就必须按比例扩大数据规模。尽管这些经验规律提供了重要指导，但要在统一框架下解释这些看似相互矛盾的权衡，其背后的理论机制仍未得到充分探索。

为解释这些经验现象，理论解释主要依赖可求解的替代模型和谱分析。Bahri et al.（2024年）、Maloney et al.（2022年）和 Bordelon et al.（2021年）将缩放指数与核和随机特征极限下的谱衰减联系起来，并指出一种“谱原理”：模型会依次学习目标函数的不同频率模态。同样，Spigler et al.（2020年）将幂律指数直接与目标函数在核特征基上的投影系数的衰减率联系起来。Hutter（2021年）和 Pan et al.（2025年）探讨了最简理论模型，其中幂律学习曲线直接源自数据分布。Paquette 等人（2025年）和 Lin 等人（2025年）进一步分析了线性设置中的计算最优性，阐明隐式正则化和协方差谱如何驱动幂律行为。另一种思路是，Sharma & Kaplan（2020年）将参数缩放与数据流形的内在维度联系起来。我们仍需要一种广义理论，能够刻画学习任务的重尾、离散特性，同时不受特定模型类别的限制。

除了静态谱极限之外，近期研究还直接从优化动力学和离散结构推导出缩放规律。对教师–学生模型上的随机梯度下降（Ren et al., 2025；Arous et al., 2025）以及结构化特征（Bordelon & Pehlevan, 2022；Bordelon et al., 2024）的分析表明，优化偏置和特征学习（Bordelon et al., 2025）如何随时间塑造学习曲线。在高精度机制下，Michaud 等人（2023）指出，神经网络通过自动发现模块化结构实现缩放，这一特性不同于经典逼近。与我们的概念框架最为接近的是，Michaud 等人（2024）提出了一种“量化模型”，其中离散技能按频率学习。然而，这些离散视角尚未从形式上统一容量、覆盖度和优化瓶颈。

## H. 引理 6.2 和假设 6.4 的理论依据

在本节中，我们为引理 6.2 中残差 $q_k$ 的动态变化以及假设 6.4 中校正系数 $\lambda_k$ 的幂律提供理论依据。具体而言，我们利用深度线性网络（DLN）的梯度流框架推导这些性质。

### H.1. 问题设定

**模型架构与有效权重**。我们考虑一个 $L$ 层的深度线性神经网络。为便于理论分析，我们假设该网络在数据的特征基上运行，此时输入特征彼此解耦，与假设 2.1 类似。令 $x_k$ 表示第 $k$ 个模式的输入。网络的预测 $\hat{z}_k$ 通过依次将 $x_k$输入 $L$ 层权重而得到：

$$\hat{z}_k = w_{L,k}(t) \cdot w_{L-1,k}(t) \cdots w_{1,k}(t) \cdot x_k,$$

其中 $w_{l,\ k}(t)$ 表示时刻 $t$模式 $k$ 对应的第 $l$ 层权重。我们将有效权重 $u_k(t)$定义为所有层权重的乘积：

$$u_k(t) \triangleq \prod_{l=1}^{L} w_{l,k}(t).$$

因此，网络函数可简化为有效参数空间中的线性缩放：$\hat{z}_k = u_k(t) \cdot x_k$。

注。 通过深度 $L$ 引入的过参数化，会使单个权重 $w_{l,k}$对应的优化动力学变为非凸，尽管我们考虑的是从 $x_k$ 到 $\hat{z}_k$的线性映射。这种深度线性网络（DLN）形式是研究特征学习的经典模型（Saxe 等，2013），它既捕捉了复杂模型（如 Transformer）中观察到的阶段式学习转变和深度诱导的加速效应，同时又保持了解析可处理性。

**损失函数**。 假设目标输出 $z_k^*$ 由底层的最优特征强度 $u_k^*$生成，使得$z_k^* = u_k^* \cdot x_k$。采用均方误差（MSE）损失后，通过对数据分布取期望，推导出第 $k$个模式对应的目标函数 $L_k$：

$$\begin{aligned} L_k &= \mathbb{E}_x\left[\frac{1}{2}(\hat{z}_k - z_k^*)^2\right] \\ &= \mathbb{E}_x\left[\frac{1}{2}(u_k \cdot x_k - u_k^* \cdot x_k)^2\right] \\ &= \frac{1}{2}(u_k - u_k^*)^2 \cdot \mathbb{E}_x[x_k^2] \\ &= \frac{1}{2}p_k(u_k - u_k^*)^2. \end{aligned}$$

不失一般性，我们假设：当采样到模式 $k$ 时，其输入特征的幅值为单位值（$x_k^2 = 1$）；否则为 $x_k = 0$。因此， $\mathbb{E}[x_k^2]$ 严格对应于第 $k$ 个模式的数据频率，即， $p_k = \mathbb{E}[x_k^2]$。

总损失对所有模式求和：

$$L = \frac{1}{2}\sum_k p_k(u_k - u_k^*)^2 \tag{8}$$

现在，我们将这种损失结构（特定的均方误差损失与 $L$ 层线性网络）明确映射到假设 2.1 中的原子模式分解。具体而言，与公式（1）对照可知：公式（8）中的平方误差项 $_{1\,2}(u_k - u_k^*)^2$刻画了模式 $k$的未学习程度，而归一化平方误差可视为正文中的归一化残差 $q_k$ ，

$$q_k(t) \triangleq \frac{(u_k(t) - u_k^*)^2}{(u_k^*)^2} \propto (u_k(t) - u_k^*)^2.$$

在这种映射下，初始化时的 $q_k = 1$ 为 $u_k(0) \approx 0$，而收敛时的 $q_k \to 0$ 为 $u_k(t) \to u_k^*$，这与正文中的损失形式完全一致。因此，接下来我们着重分析误差项的收敛动力学（$u_k(t) - u_k^*$）$^2$, , 它反映了残差 $q_k$的衰减。

**假设 H.1**（**可学习性条件**）。 给定常数 $\zeta > 0$，我们假设最优特征强度 $u_k^*$ 遵循关于数据频率的幂律：

$$u_k^* \propto p_k^{\zeta}.$$

**对假设 H.1 的理解**。 参数 $\zeta$ 控制着最优特征强度相对于输入分布的丰富程度。在自然语言中，输入谱 $p_k$ 遵循重尾的齐夫分布。假设 H.1 认为，模式的语义重要性与其频率自然保持一致，其中 $\zeta > 0$。换言之，频繁出现的模式（例如核心句法结构或常见主题词）通常具有更大的真实权重（$u_k^*$）。然而，如果 $\zeta \leq 0$，这意味着模型需要从稀有模式（较小的 $p_k$）中学习较大的目标权重（$u_k^*$）。但是，梯度下降更新步与数据频率成正比（公式（9）），由于极小的 $p_k$，有效学习率趋近于零。这使得较大的目标 $u_k^*$ 实际上难以学习，因为优化会停滞在持续时间较长的平台期。

### H.2. 梯度流动力学

我们分析连续时间梯度下降（梯度流， $\eta \to 0$）下的学习动态。应用链式法则，有效权重 $u_k$ 的时间演化为：

$$\begin{aligned} \frac{du_k}{dt} &= \sum_{l=1}^{L} \frac{\partial u_k}{\partial w_{l,k}} \frac{dw_{l,k}}{dt} \\ &= \sum_{l=1}^{L} \frac{\partial u_k}{\partial w_{l,k}} \left(-\eta \frac{\partial \mathcal{L}}{\partial u_k} \frac{\partial u_k}{\partial w_{l,k}}\right) \\ &= -\eta \frac{\partial \mathcal{L}}{\partial u_k} \sum_{l=1}^{L} \left(\frac{\partial u_k}{\partial w_{l,k}}\right)^2. \end{aligned}$$

在小幅平衡初始化下，深度线性网络中的梯度流保持不变量 $\frac{d}{dt}(w_{l,\,k}^2 - w_{j,\,k}^2) \to 0$，这意味着各层的权重大小相等：$|w_{l,\,k}| \approx |u_k|^{1/L}$（Arora 等，2018年）。利用关系式 $\frac{\partial u_k}{\partial w_{l\,,\,k}} = \frac{u_k}{w_{l\,,\,k}}$，我们可以简化预条件器项：

$$\sum_{l=1}^{L}\left(\frac{\partial u_k}{\partial w_{l,k}}\right)^2 = \sum_{l=1}^{L}\left(\frac{u_k}{w_{l,k}}\right)^2 = \sum_{l=1}^{L}\frac{u_k^2}{|u_k|^{2/L}} = L\cdot|u_k|^{2-\frac{2}{L}}.$$

将梯度 $\frac{\partial \mathcal{L}}{\partial u_k} = p_k(u_k - u_k^*)$ 代入动力学方程，可得支配 $u_k$ 的非线性常微分方程（ODE），

$$\frac{du_k}{dt} = \eta L p_k u_k^{2-\frac{2}{L}}(u_k^* - u_k). \tag{9}$$

### H.3. 收敛分析与有效速率

现在，我们推导浅层网络（$L=2$）和一般深度网络的渐近收敛速率。

**情形 1：浅层网络**（$L=2$）。不失一般性，假设 $u_k^* > 0$ 且 $u_k(0) > 0$，则方程（9）可简化为标准的逻辑斯蒂微分方程:

$$\frac{du_k}{dt} = 2\eta p_k u_k(u_k^* - u_k).$$

解析解为

$$u_k(t) = \frac{u_k^*\cdot Ae^{2\eta\Lambda_k t}}{1+Ae^{2\eta\Lambda_k t}} = \frac{u_k^*\cdot e^{2\eta\Lambda_k t}}{e^{2\eta\Lambda_k t}+A^{-1}},$$

其中 $\Lambda_k = p_k u_k^*$ 和 $A = \frac{u_k(0)}{u_k^* - u_k(0)}$ 由初始化决定。在收敛阶段（$t$ 较大时），误差项呈指数衰减:

$$u_k^* - u_k(t) = \frac{u_k^* A^{-1}}{e^{2\eta\Lambda_k t}+A^{-1}} \approx u_k^* A^{-1} e^{-2\eta\Lambda_k t} \propto e^{-2\eta\Lambda_k t},$$

因此

$$(u_k(t) - u_k^*)^2 \propto \exp\left(-4\eta\underbrace{p_k u_k^*}_{\Lambda_k} t\right). \tag{10}$$

**情况 2：一般深度网络**（$L>2$）。对于一般深度，精确解是隐式的。对方程（9）分离变量可得:

$$t = \frac{1}{\eta L p_k}\int_{u_k(0)}^{u_k(t)}\frac{dz}{z^{2-\frac{2}{L}}(u_k^* - z)}.$$

代入 $y = z/u_k^*$后，积分变为:

$$t = \frac{1}{\eta L p_k (u_k^*)^{2-\frac{2}{L}}}\int_{y_0}^{y(t)}\frac{dy}{y^{2-\frac{2}{L}}(1-y)}.$$

该积分涉及高斯超几何函数。然而，我们主要关注收敛区间（$y\to 1$）内的渐近行为。在 $y=1$ 附近展开被积函数 $\frac{1}{y^{2-2/L}(1-y)}$，主导项为 $\frac{1}{1-y}$。进行积分:

$$\int\frac{dy}{1-y} = -\ln(1-y) = -\ln\frac{u_k^* - u_k}{u_k^*}.$$

因此，时间演化渐近地由以下关系支配:

$$t \approx \frac{1}{\eta L p_k (u_k^*)^{2-\frac{2}{L}}}\left[-\ln(u_k^* - u_k) + C\right].$$

重新整理误差项，得到衰减定律：

$$(u_k(t) - u_k^*)^2 \propto \exp\left(-2\eta L \underbrace{\left[p_k (u_k^*)^{2-\frac{2}{L}}\right]}_{\Lambda_k} t\right). \tag{11}$$

这证实，对于任意深度 $L$，渐近收敛速率由 $\Lambda_k = p_k (u_k^*)^{2-\frac{2}{L}}$ 决定，这与针对 $L = 2$情形推导出的解析解一致。

### H.4. 引理 6.2 和假设 6.4 的论证

最后，我们将推导出的收敛结果与引理 6.2 和假设 6.4 联系起来。回顾一下，归一化残差 $q_k(t)$ 表示平方误差分量。根据公式（11）中的参数误差衰减：

$$q_k(t) \propto (u_k(t) - u_k^*)^2 \propto \exp\left(-2\eta L \Lambda_k t\right).$$

用步长 $\Delta t = 1$ 对这一连续衰减过程进行离散化后，更新规则变为：

$$q_k(t+1) = q_k(t) \cdot e^{-2\eta L \Lambda_k}.$$

对较小的学习率 $\eta$ 应用一阶泰勒展开 $e^{-x} \approx 1 - x$，可得

$$q_k(t+1) \approx (1 - \eta \cdot [2L\Lambda_k]) q_k(t), \tag{12}$$

这进一步验证了引理 6.2。

比较公式（12）与引理 6.2 中的 $q_k(t+1) = (1 - \eta\lambda_k) q_k(t)$，我们确定逐步修正系数 a s:

$$\lambda_k \equiv 2L\Lambda_k.$$

结合 $\Lambda_k = p_k (u_k^*)^{2-\frac{2}{L}}$ 和假设 H.1：

$$\begin{aligned} \lambda_k &\propto L p_k (u_k^*)^{2-\frac{2}{L}} \\ &\propto L p_k (p_k^{\zeta})^{2-\frac{2}{L}} \\ &= L p_k^{1+\zeta(2-\frac{2}{L})}, \end{aligned}$$

这为 随机梯度下降动力学下 $L$ 层深度线性网络中的假设 6.4 提供了理论依据 .

在假设 6.4 中，我们设 定 $\lambda_k \propto p_k^{\beta-1}$。通过匹配指数，我们从理论上确定缩放系数 $\beta$ :

$$\beta - 1 = 1 + \zeta\left(2 - \frac{2}{L}\right) \implies \beta = 2 + \zeta\left(2 - \frac{2}{L}\right). \tag{13}$$

由于 $L \geq 1$ 和 $\zeta > 0$，缩放指数满足 $\beta \geq 2$，这使优化处于富区间（其特征为 $\beta > 1$）。

推导出的表达式 $\beta = 2 + \zeta(2 - 2/L)$ 从数学上形式化了数据结构与模型架构之间的相互作用。具体而言，与深度相关的项 $2 - 2/L$ 引入了一种非线性选择机制：对于目标幅度较大的主导特征 $u_k^*$，网络会带来超线性加速；反之，对于幅度较小的弱特征或噪声特征 $u_k^*$，学习动态会受到严重抑制。这一机制体现了深度网络的隐式偏置，推动优化优先处理高频强特征，而非弱模式。