Can We Derive Scaling Law From First Principles?

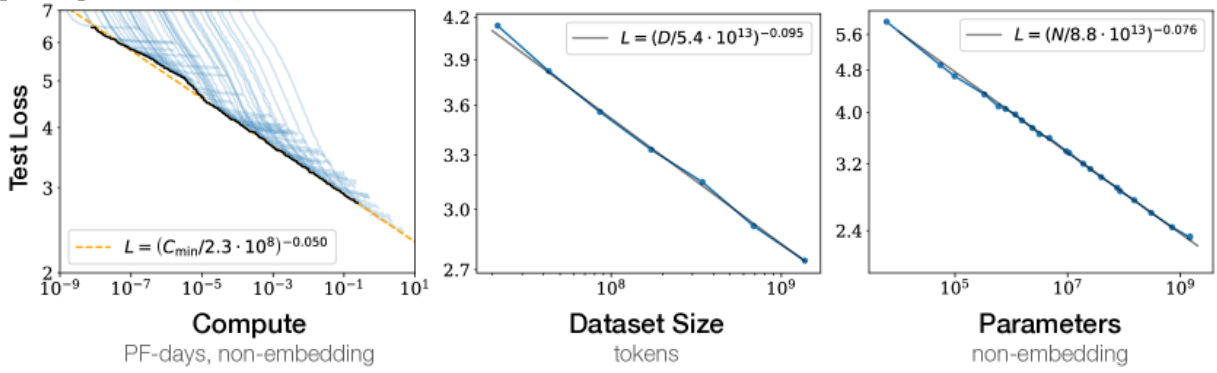
Jiaxuan Zou

2025 年 12 月 30 日

摘要

大规模语言模型预训练中，测试损失随模型规模 N、数据规模 D 与计算量 C 在一定区间内呈现稳定的幂律衰减；在给定算力预算下，最优资源分配也形成近似幂律的前沿。已有理论解释往往依赖 NTK 线性化或对谱分布（如核谱、Teacher 系数）的特定先验假设。本文提出一种不预设核谱幂律、也不依赖线性化的推导框架：将语言建模抽象为对可数模式集合的逐步掌握过程，并用语言数据的长尾统计结构与梯度下降的频率依赖收敛动力学，解析推出四类常见 scaling law 的指数形式。

核心设定是：模式按其重要性排序后满足 Zipf 型分布 $p_k \propto k^{-\alpha}$（$\alpha > 1$）。训练资源的有限性在模式秩空间诱导一个有效覆盖的临界秩 $k_\star$，可约损失由未覆盖尾部概率质量主导。进一步地，我们用一个最小化假设刻画优化动力学：模式 k 的残差强度 $q_k(\tau)$ 以速率 p_k^β 指数衰减，从而得到训练时间（或 FLOPs）scaling。将三类单变量瓶颈以 max 结构组合，即可直接推出两段式学习曲线与 optimal compute point 的幂律关系，并在相同的瓶颈模型下统一解释 Kaplan 与 Chinchilla 的 compute-optimal 分配律。



目录

1 引言

经验研究表明，语言模型预训练的测试损失在相当宽的范围内满足幂律拟合：当单独改变模型规模 N、数据规模 D 或计算预算（训练步数、token-steps 或累计 FLOPs）时，损失以近似固定指数衰减；当施加算力约束并在可行集合内优化资源分配时，最优损失同样以幂律方式随预算降低。该现象的数学机制需要同时解释两个事实：其一，长尾数据导致误差在后期由稀有结构主导；其二，优化过程对不同结构的收敛速度存在系统性差异。

> **Remark 1.1.** 研究如何从第一性原理出发推导 scaling law 本身就存在一个尺度的问题，类似于希尔伯特第六问题：从牛顿运动定律出发，通过严格数学推导得出纳维-斯托克斯方程等宏观流体力学方程。无论是 NTK、KRR 这种线性化近似还是 feature learning 等高度非线性的框架，这些都是非常 fine-grained 的工具，想要从这些小尺度的工具得到一个大尺度的 scaling law，难度是非常高的。更别说我们对梯度下降带来的 implicit bias 和 feature learning 也没有研究清楚。

为避免在假设层面直接引入核谱幂律，本文采用模式化建模：将语言分布分解为可数的结构单元（模式），并把测试损失的可约部分写成模式残差的加权和。该视角的关键在于：即使不显式指定网络内部的特征谱，只要能够把资源约束对应到一个随资源增长的有效覆盖范围（临界秩 $k_\star$），并且模式权重满足长尾统计，则幂律可由尾部质量的渐近行为给出。本文进一步给出一种可计算的动力学近似，用于导出训练时间/算力 scaling 及其指数与数据统计参数之间的关系。

本文主要结论可概括为四类幂律：模型规模 scaling、数据规模 scaling、训练时间（或 FLOPs）scaling，以及 optimal compute point（或 compute-optimal 前沿）对应的 loss–compute 幂律。为便于读者定位逻辑结构，全文将按照以下顺序推进：先给出模式分解与 Zipf 假设；再给出优化动力学的频率依赖收敛模型；随后分别推导三类单变量 scaling；最后给出联合资源下的瓶颈组合律与两类常见算力约束下的 compute-optimal 解。

2 预备：模式空间与 Zipf 假设

本节定义模式空间与损失分解，并明确本文使用的统计假设与近似范围。推导目标是获得可约损失 ΔL 的主导阶（数量级与指数），因此下文主要关注渐近等价与上、下界夹逼，常数因子只在需要给出精确常数时讨论。

2.1 模式、重要性排序与可约损失分解

模式与索引 设语言分布中存在一族可数结构单元，记为模式集合 $\{1, 2, \dots\}$。模式的具体含义不限定为 token 或 n-gram，也可以是语法依赖、事实映射、长程关联等任何可重复、可统计识别的结构。本文只使用如下抽象属性：模式能够被排序，并且在测试损失上可近似加性分解。

重要性权重 令 p_k 表示模式 k 的重要性权重（可理解为其出现概率、或其对损失的边际贡献所诱导的权重）。不失一般性，按重要性降序排列：

$$p_1 \geq p_2 \geq \cdots \geq 0, \qquad \sum_{k=1}^{\infty} p_k = 1. \tag{1}$$

可约损失的模式分解 将总体测试损失 L 分解为不可约误差 E 与可约部分 $\Delta L := L - E$。在期望意义下，我们假设：

$$\Delta L \approx \sum_{k=1}^{\infty} p_k q_k, \tag{2}$$

其中 $q_k \in [0,1]$ 表示模式 k 的残差强度：$q_k = 0$ 表示该模式已被充分拟合（对 ΔL 的贡献可以忽略），$q_k = 1$ 表示该模式基本未被拟合。

式 (2) 的作用是把宏观量 ΔL 转换为两个对象的乘积求和：数据统计权重 p_k 与训练状态 q_k。本文所有 scaling law 均来自对该求和在不同资源约束下的渐近估计。

2.2 Zipf 统计假设

假设 2.1 (Zipfian Distribution)**.** 设模式集按重要性严格降序排列，其概率质量 p_k 服从 Zipf 分布：

$$p_k = \frac{1}{Z} k^{-\alpha}, \quad \alpha > 1, \tag{3}$$

其中 $Z = \zeta(\alpha)$ 为归一化常数。

Remark 2.1. 该假设可进一步推广至正则变差情形，即 $p_k = k^{-\alpha} L(k)$，其中 $L(\cdot)$ 为缓变函数。这一推广允许对数修正项的存在，但在渐近意义下不改变 Scaling Law 的幂律指数，因此后文推导主要基于标准形式 (3)。

基于核谱和 Teacher 系数满足正则变差的假设的结果我已经推导过了，无非只是要额外用一些 lemma（Potter bound 等）导出幂律形式，不新鲜，很 trivial，因为还是在条件处引入了带有幂律形式的东西 $k^{-\alpha}L(k)$。最后并不是很满意，也没有放在本文里。

在 Zipf 假设下，尾部质量的渐近行为将成为后续推导的核心。直观上，α 越接近 1，尾部衰减越慢，稀有模式的累计权重越大，导致任何以尾部质量为主导的误差下降都更缓慢。后文会把该直观量化为具体的幂律指数。

3 优化动力学的频率依赖收敛模型

本节给出一个最小化的动力学模型，用于把训练时间（或累计计算）与残差序列 $\{q_k\}$ 联系起来。目标是得到一个可解析的 $q_k(\tau)$ 形式，使得代回 (2) 后可以对 $\Delta L(\tau)$ 做渐近估计。

Remark 3.1. 我觉得不仅是在具有长尾分布的模式空间中表现出模式学习的有序性这种 implicit bias，在前面的 remark 里提到，在 NN+GD 的设定下，很多时候都会出现这个 implicit bias。像是许志钦提出的频率原则等，都是这种 bias 的表现形式。

3.1 有效覆盖前沿的抽象假设

在有限资源下，训练往往只能显著降低部分模式的残差。我们用一个临界秩 $k_\star$ 对此进行抽象：秩较小的模式残差趋近于 0，秩较大的模式残差接近于 1。该抽象将复杂的训练轨迹压缩为秩空间

中的截断结构，后续推导将主要依赖尾部（$k > k_\star$）的质量。（一下子就把问题简化为了一个求和尾部的估计问题）

假设 3.1 (Ordered Learning & Effective Frontier). 设模式按重要性降序排列。在任意给定的有限训练资源（参数量 N、数据量 D 或计算量 τ）约束下，残差序列 $\{q_k\}_{k\geq 1}$ 满足：

1. **单调性：** $0 \leq q_1 \leq q_2 \leq \cdots \leq 1$。
2. **有效前沿：** 存在临界截断秩 $k_\star$（依赖于资源约束），使得当 $k \ll k_\star$ 时 q_k 很小，而当 $k \gg k_\star$ 时 q_k 接近 1。

该假设不指定 $k_\star$ 如何随资源变化，而只要求存在一个可用于夹逼尾部贡献的分离区间。后文在模型规模、数据规模与时间规模三种情形下分别给出 $k_\star$ 的数量级，从而把 ΔL 的估计归约为 Zipf 尾和或其连续近似积分。

3.2 训练时间下的残差动力学

我们在训练时间维度上引入三个假设：模式以其权重 p_k 被观测；每次观测对该模式残差进行近似乘法收缩；收缩强度允许依赖频率，并用最简幂律参数化。

假设 3.2 (模式的随机采样 (i.i.d.)). 训练过程可视为对模式集合的随机观测：在每一步（或每个 token-step）t，观测到的模式 $K_t \in \mathbb{N}$ 独立同分布，且

$$\Pr[K_t = k] = p_k.$$

记指示变量 $I_{t,k} := \mathbf{1}\{K_t = k\}$，则 $\mathbb{E}[I_{t,k}] = p_k$。

假设 3.3 (局部线性纠错 / 乘法收缩). 对每个模式 k，存在残差强度 $q_k(t) \in [0,1]$ 描述该模式未拟合程度。当第 t 步观测到模式 k 时，该模式残差发生一次近似乘法收缩：

$$q_k(t+1) = (1 - \eta\lambda_k)\, q_k(t) \qquad 若 I_{t,k} = 1,$$

若未观测到该模式则保持不变：

$$q_k(t+1) = q_k(t) \qquad 若 I_{t,k} = 0,$$

其中步长 $\eta > 0$，$\lambda_k > 0$ 为该模式的有效纠错系数。

假设 3.4 (频率调制的有效纠错系数). 有效纠错系数 λ_k 允许随模式频率变化。为得到可解析的宏观指数，采用幂律参数化：

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

下面给出残差随训练时间的显式近似。证明将把离散随机更新转化为平均意义下的指数衰减，并明确近似的来源（大数定律与 $\log(1-x)$ 展开）。

引理 3.1 (频率依赖的指数收敛). 在 假设 3.2 to 3.4 下，令训练步数（或 token-step）为 τ，并取 $q_k(0) = 1$。当 $\eta\lambda_k$ 足够小且 τ 足够大时，有

$$q_k(\tau) \;\approx\; \exp\big(-c\,\tau\,p_k^{\beta}\big), \qquad c := \eta\lambda_0. \tag{4}$$

等价地，在连续时间极限下满足

$$\frac{d}{d\tau}q_k(\tau) = -c\,p_k^{\beta}\,q_k(\tau), \qquad q_k(0) = 1.$$

证明. 固定 k。由 假设 3.3，在 τ 步内该模式被观测到的次数为

$$n_k(\tau) := \sum_{t=0}^{\tau-1} I_{t,k}.$$

每当 $I_{t,k} = 1$，残差乘以 $(1-\eta\lambda_k)$；当 $I_{t,k} = 0$，残差不变。因此

$$q_k(\tau) = (1-\eta\lambda_k)^{n_k(\tau)} q_k(0) = (1-\eta\lambda_k)^{n_k(\tau)}.$$

由 假设 3.2，$\{I_{t,k}\}_{t\ge 0}$ 为 i.i.d. Bernoulli(p_k)，故

$$\frac{n_k(\tau)}{\tau} \to p_k \quad \text{几乎处处},$$

从而在大 τ 下可用 $n_k(\tau) \approx \tau p_k$ 近似（在概率意义或期望意义下均可成立）。

接下来处理指数近似。对 $0 < x < 1$，$\log(1-x) = -x + O(x^2)$。当 $\eta\lambda_k$ 足够小，

$$(1-\eta\lambda_k)^{n_k(\tau)} = \exp\Big(n_k(\tau)\log(1-\eta\lambda_k)\Big) \approx \exp\big(-\eta\lambda_k\, n_k(\tau)\big).$$

再用 $n_k(\tau) \approx \tau p_k$ 得

$$q_k(\tau) \approx \exp\big(-\eta\lambda_k\,\tau p_k\big).$$

最后由 假设 3.4，$\lambda_k = \lambda_0 p_k^{\beta-1}$，故

$$q_k(\tau) \approx \exp\big(-\eta\lambda_0\,\tau\,p_k^{\beta}\big) = \exp\big(-c\,\tau\,p_k^{\beta}\big),$$

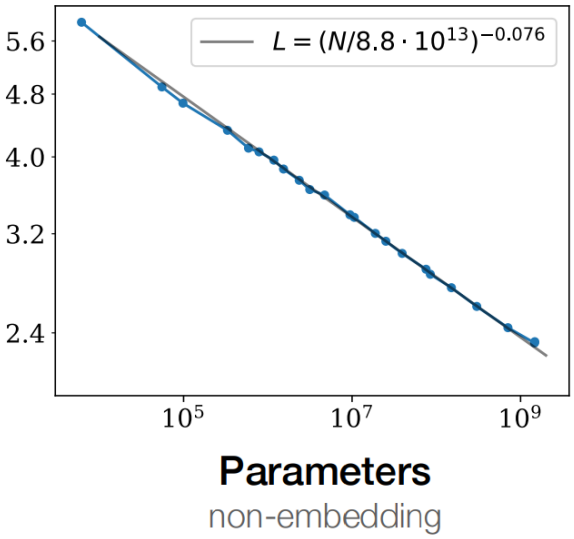
即 (4)。 □

Remark 3.2 (参数 β 的含义). 引理 3.1 给出学习速率的乘法分解：$c\,p_k^{\beta}$ 同时包含 (i) 模式被观测到的频率 $\propto p_k$，(ii) 单次观测的有效纠错强度 $\propto \lambda_k \propto p_k^{\beta-1}$。因此 β 定量刻画了优化过程对频率的敏感度：$\beta = 1$ 是“仅由观测频率决定”的基线；$\beta > 1$ 表示头部模式在每次被观测到时获得更强的有效更新，从而对头部模式呈现更强的隐式偏好并导致尾部学习滞后。

由 (4) 可见：给定 τ，当 $c\tau p_k^{\beta} \gg 1$ 时 $q_k(\tau)$ 很小，当 $c\tau p_k^{\beta} \ll 1$ 时 $q_k(\tau) \approx 1$。因此可以把临界秩 $k_\star(\tau)$ 定义为满足 $c\tau p_{k_\star}^{\beta} \asymp 1$ 的解，并据此估计尾部贡献。这一过程将在 Section 6 中形式化为积分换元计算。

4 第一种 Scaling Law：模型规模 Scaling

本节在数据规模与训练时间足以使得数据覆盖与收敛误差次主导的情形下，推导 ΔL 随参数规模 N 的下降规律。这里的关键是把 N 映射为可有效拟合的模式数量（或临界秩）$k_\star(N)$。



4.1 容量前沿与尾部质量的归约

假设 4.1 (Capacity map). 存在单调递增函数 $M : \mathbb{N} \to \mathbb{N}$ 使得容量前沿满足

$$k_\star(N) \asymp M(N).$$

假设 4.1 的作用是把表征能力与结构可拟合性压缩进一个单调函数 $M(N)$。在许多参数化下，$M(N)$ 可进一步用幂律表示：

$$M(N) \propto N^\gamma, \qquad \gamma > 0. \tag{5}$$

下文先在一般 $M(N)$ 下推导，再在 (5) 下得到显式指数。

下面说明在有效前沿假设下，$\Delta L(N)$ 的数量级等价于 Zipf 权重的尾和。

命题 4.1 (容量前沿诱导的尾和表达). 在 假设 3.1 与 假设 4.1 下，存在 $k_\star(N) \asymp M(N)$ 使得对任意固定 $\varepsilon \in (0,1)$，

$$\lim_{N\to\infty} \sup_{k \le (1-\varepsilon)k_\star(N)} q_k(N) = 0, \qquad \lim_{N\to\infty} \inf_{k \ge (1+\varepsilon)k_\star(N)} q_k(N) = 1.$$

从而可约损失满足夹逼

$$\sum_{k>(1+\varepsilon)k_\star(N)} p_k \lesssim \Delta L(N) \lesssim \sum_{k>(1-\varepsilon)k_\star(N)} p_k. \tag{6}$$

特别地，

$$\Delta L(N) \asymp \sum_{k>k_\star(N)} p_k \asymp \sum_{k>M(N)} p_k.$$

证明. 由 (2),

$$\Delta L(N) \approx \sum_{k\geq 1} p_k q_k(N).$$

利用 $p_k \geq 0$ 与 q_k 的单调性, 分别对头部与尾部进行估计。对 $k \leq (1-\varepsilon)k_\star(N)$, 由假设有 $q_k(N) \to 0$, 因此这部分贡献可被任意小的常数控制。对 $k \geq (1+\varepsilon)k_\star(N)$, 有 $q_k(N) \to 1$, 因此尾部贡献与 $\sum_{k>(1+\varepsilon)k_\star} p_k$ 同阶。将两侧合并得到 (6)。最后用 $k_\star(N) \asymp M(N)$ 替换即可。 □

4.2 Zipf 尾和的渐近与模型 scaling 指数

定理 4.1 (模型规模 Scaling). 在 假设 2.1 下 ($p_k = \frac{1}{Z}k^{-\alpha}$, $\alpha > 1$), 有

$$\sum_{k>M} p_k = \frac{1}{Z}\sum_{k>M} k^{-\alpha} \sim \frac{1}{Z(\alpha-1)} M^{-(\alpha-1)}, \qquad M \to \infty. \tag{7}$$

结合 命题 4.1 得

$$\boxed{\Delta L(N) \propto M(N)^{-(\alpha-1)}}. \tag{8}$$

若进一步满足 (5), 则

$$\boxed{\Delta L(N) \propto N^{-\gamma(\alpha-1)}}. \tag{9}$$

证明. 给出两步: 先把离散尾和与积分夹逼, 再得到渐近等价。

对 $\alpha > 1$, 函数 $x \mapsto x^{-\alpha}$ 单调递减。由积分判别法,

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

计算积分得

$$\int_{M}^{\infty} x^{-\alpha}\,dx = \frac{1}{\alpha-1} M^{-(\alpha-1)}, \qquad \int_{M+1}^{\infty} x^{-\alpha}\,dx = \frac{1}{\alpha-1}(M+1)^{-(\alpha-1)}.$$

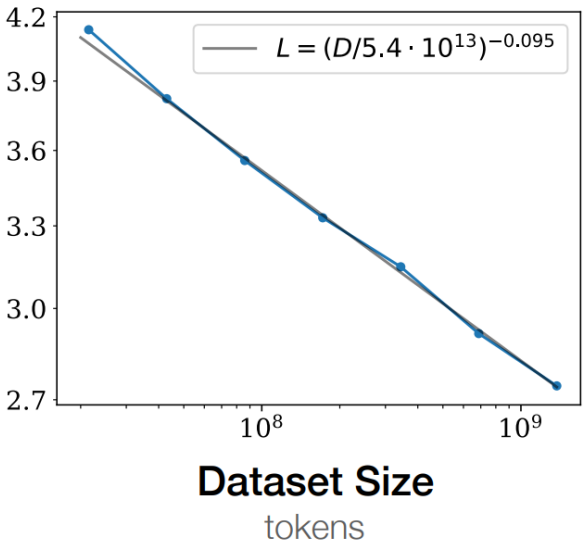
两侧相除并令 $M \to \infty$ 得 $\sum_{k>M} k^{-\alpha} \sim \frac{1}{\alpha-1} M^{-(\alpha-1)}$, 乘以 $1/Z$ 即得 (7)。

由 命题 4.1, $\Delta L(N) \asymp \sum_{k>M(N)} p_k$, 代入 (7) 得 (8)。若 $M(N) \propto N^\gamma$, 进一步代入得到 (9)。 □

由 (9) 可见, 模型 scaling 指数为 $\alpha_N = \gamma(\alpha-1)$。当数据分布固定 ($\alpha$ 固定) 时, α_N 与架构效率参数 γ 成正比; 当架构固定 (γ 固定) 时, α_N 由尾指数 α 决定, 且 $\alpha \to 1^+$ 时指数趋近 0, 对应尾部质量衰减变慢。

5 第二种 Scaling Law: 数据规模 Scaling

本节在模型容量与训练时间足够大的前提下, 研究仅由数据覆盖不足导致的可约误差随训练数据规模 D 的下降规律。为将问题与统计覆盖严格对齐, 我们在本节将可约误差的一个标准代理定义为训练样本的 missing mass 的期望, 并证明其在 Zipf 长尾下满足幂律。



5.1 missing mass 与指数夹逼

设训练集由 D 个 token 组成，并视为从真实分布独立抽样。对每个模式 k，令 $X_k \sim \mathrm{Binomial}(D, p_k)$ 表示其在训练集中出现次数。定义残差代理为

$$q_k(D) := \Pr[X_k = 0] = (1 - p_k)^D. \tag{10}$$

于是由 (2) 的加性形式出发，本节把数据覆盖不足导致的可约误差代理定义为

$$\Delta L(D) \;:=\; \mathbb{E}\Big[\sum_{k\geq 1} p_k\, \mathbf{1}\{X_k = 0\}\Big] \;=\; \sum_{k\geq 1} p_k\, \Pr[X_k = 0] \;=\; \sum_{k\geq 1} p_k (1 - p_k)^D. \tag{11}$$

（注：这里的 $\Delta L(D)$ 是 “coverage-limited regime” 的统计代理；其与真实测试损失的严格关系依赖于更细的泛化与估计误差建模，本节不涉及。）

为把 $(1 - p_k)^D$ 与指数形式联系起来，需要一个可用于求和的一致夹逼。

引理 5.1 (未观测概率的指数夹逼). 对任意 $p \in [0, 1)$ 与整数 $D \geq 1$，有

$$(1 - p)^D \;\leq\; e^{-Dp}. \tag{12}$$

并且当 $p \in [0, 1/2]$ 时还有下界

$$e^{-D(p+p^2)} \;\leq\; (1 - p)^D \;\leq\; e^{-Dp}. \tag{13}$$

证明. 由 $\log(1 - p) \leq -p$ 得 $(1 - p)^D = \exp(D \log(1 - p)) \leq e^{-Dp}$，即 (12)。当 $p \in [0, 1/2]$ 时，函数 $h(p) := \log(1 - p) + p + p^2$ 满足 $h(0) = 0$ 且 $h'(p) = -\frac{1}{1-p} + 1 + 2p = \frac{p(1-2p)}{1-p} \geq 0$，故 $h(p) \geq 0$，即 $\log(1 - p) \geq -p - p^2$。指数化并乘以 D 得 (13)。 □

接下来说明：在长尾情形下，(11) 的主阶可由

$$S(D) \;:=\; \sum_{k\geq 1} p_k\, e^{-Dp_k} \tag{14}$$

刻画；换言之，将 $(1 - p_k)^D$ 替换为 e^{-Dp_k} 不影响幂律主阶。

命题 5.1 (从 $(1-p_k)^D$ 到 e^{-Dp_k} 的主阶等价). 设 $\{p_k\}_{k\ge1}$ 单调不增且 $\sum_{k\ge1} p_k = 1$。令 $\Delta L(D)$ 与 $S(D)$ 分别由 (11) 与 (14) 定义。则对任意 $D \ge 4$，存在与 D 无关的常数 $c_1, c_2 > 0$ 使得

$$c_1\, S(D) \;\le\; \Delta L(D) \;\le\; c_2\, S(D). \tag{15}$$

特别地，$\Delta L(D) \asymp S(D)$。

证明. 由 引理 5.1 的上界 (12) 立得 $\Delta L(D) \le S(D)$，故可取 $c_2 = 1$。

为证下界，取阈值 $\theta := D^{-1/2} \le 1/2$（当 $D \ge 4$）。令

$$K(D) := \max\{k \ge 1 :\; p_k \ge \theta\}, \quad \text{若集合为空则取} K(D) = 0.$$

我们做一个截断，将 (11) 分解为头部与尾部：

$$\Delta L(D) = \sum_{k \le K(D)} p_k (1-p_k)^D \;+\; \sum_{k > K(D)} p_k (1-p_k)^D =: H(D) + T(D).$$

对尾部 $k > K(D)$，有 $p_k \le \theta \le 1/2$，于是由 (13) 得

$$(1-p_k)^D \;\ge\; e^{-D(p_k + p_k^2)} = e^{-Dp_k}\, e^{-Dp_k^2}.$$

又因 $p_k \le D^{-1/2}$，故 $Dp_k^2 \le 1$，从而 $e^{-Dp_k^2} \ge e^{-1}$。因此

$$T(D) = \sum_{k > K(D)} p_k (1-p_k)^D \;\ge\; e^{-1} \sum_{k > K(D)} p_k e^{-Dp_k}.$$

另一方面，头部 $H(D)$ 非负。于是

$$\Delta L(D) \;\ge\; T(D) \;\ge\; e^{-1} \sum_{k > K(D)} p_k e^{-Dp_k}.$$

最后注意到对头部 $k \le K(D)$，有 $p_k \ge D^{-1/2}$，故

$$\sum_{k \le K(D)} p_k e^{-Dp_k} \;\le\; \sum_{k \le K(D)} p_k e^{-D^{1/2}} \;\le\; e^{-D^{1/2}}.$$

因此

$$S(D) = \sum_{k \le K(D)} p_k e^{-Dp_k} + \sum_{k > K(D)} p_k e^{-Dp_k} \le e^{-D^{1/2}} + \sum_{k > K(D)} p_k e^{-Dp_k} \le (1 + e^{-D^{1/2}}) \sum_{k > K(D)} p_k e^{-Dp_k}.$$

合并两式得

$$\Delta L(D) \;\ge\; e^{-1}(1 + e^{-D^{1/2}})^{-1} S(D) \;\ge\; \frac{1}{2e}\, S(D),$$

其中最后一步用 $D \ge 4$ 时 $e^{-D^{1/2}} \le e^{-2} \le 1$。故可取 $c_1 = 1/(2e)$，完成证明。 □

5.2 数据 scaling 的解析推导（Zipf 情形）

定理 5.1 (数据规模 Scaling). 设 $p_k = \frac{1}{Z} k^{-\alpha}$，其中 $\alpha > 1$、$Z = \zeta(\alpha)$。令 $\Delta L(D)$ 由 (11) 定义。则当 $D \to \infty$ 时

$$\boxed{\Delta L(D) \;\asymp\; D^{-(\alpha-1)/\alpha}}. \tag{16}$$

更精确地，有渐近

$$\Delta L(D) \;\sim\; c_\alpha\, D^{-(\alpha-1)/\alpha}, \qquad c_\alpha = \frac{1}{\alpha}\Big(\frac{1}{Z}\Big)^{1/\alpha}\Gamma\Big(1-\frac{1}{\alpha}\Big).$$

证明. 由 命题 5.1，只需计算 $S(D)=\sum_{k\ge 1} p_k e^{-Dp_k}$ 的渐近，再将结论转回 $\Delta L(D)$。

代入 $p_k=\frac{1}{Z}k^{-\alpha}$，得

$$S(D)=\frac{1}{Z}\sum_{k=1}^{\infty} k^{-\alpha}\exp\!\Big(-\frac{D}{Z}k^{-\alpha}\Big).$$

令尺度参数

$$a:=\Big(\frac{D}{Z}\Big)^{1/\alpha}, \qquad g(y):=y^{-\alpha}e^{-y^{-\alpha}},\quad y>0.$$

则注意到

$$k^{-\alpha}=\frac{1}{a^\alpha}\Big(\frac{k}{a}\Big)^{-\alpha}, \qquad \exp\!\Big(-\frac{D}{Z}k^{-\alpha}\Big)=\exp\!\big(-a^\alpha k^{-\alpha}\big)=\exp\!\Big(-\Big(\frac{k}{a}\Big)^{-\alpha}\Big),$$

从而

$$S(D)=\frac{1}{Z}\,a^{-(\alpha-1)}\cdot\underbrace{\frac{1}{a}\sum_{k=1}^{\infty} g\Big(\frac{k}{a}\Big)}_{=:R(a)}. \tag{17}$$

下面证明 $R(a)\to\int_0^\infty g(y)\,dy$。首先 $g\ge 0$，且当 $y\to\infty$ 时 $g(y)\sim y^{-\alpha}$ 可积 $(\alpha>1)$；当 $y\downarrow 0$ 时 $g(y)=y^{-\alpha}e^{-y^{-\alpha}}$ 超指数衰减亦可积，故 $g\in L^1(0,\infty)$。对任意 $M>0$，

$$\frac{1}{a}\sum_{k=1}^{\lfloor aM\rfloor} g\Big(\frac{k}{a}\Big)\;\xrightarrow[a\to\infty]{}\;\int_0^M g(y)\,dy$$

为标准的黎曼和收敛。再对尾部，用 $g\in L^1(0,\infty)$ 得

$$\sup_{a\ge 1}\frac{1}{a}\sum_{k>\lfloor aM\rfloor} g\Big(\frac{k}{a}\Big)\;\le\;\int_M^\infty g(y)\,dy\;\xrightarrow[M\to\infty]{}\;0,$$

从而可交换极限并得到

$$R(a)=\frac{1}{a}\sum_{k=1}^{\infty} g\Big(\frac{k}{a}\Big)\;\xrightarrow[a\to\infty]{}\;\int_0^\infty g(y)\,dy.$$

计算该积分：作换元 $u=y^{-\alpha}$，则 $y=u^{-1/\alpha}$，$dy=-(1/\alpha)u^{-1/\alpha-1}du$，从而

$$\int_0^\infty g(y)\,dy=\int_0^\infty y^{-\alpha}e^{-y^{-\alpha}}\,dy=\frac{1}{\alpha}\int_0^\infty e^{-u}u^{-1/\alpha}\,du=\frac{1}{\alpha}\Gamma\Big(1-\frac{1}{\alpha}\Big).$$

代回 (17)，并用 $a^{-(\alpha-1)}=(D/Z)^{-(\alpha-1)/\alpha}$，得

$$S(D)\;\sim\;\frac{1}{Z}\Big(\frac{D}{Z}\Big)^{-(\alpha-1)/\alpha}\cdot\frac{1}{\alpha}\Gamma\Big(1-\frac{1}{\alpha}\Big)=\frac{1}{\alpha}\Big(\frac{1}{Z}\Big)^{1/\alpha}\Gamma\Big(1-\frac{1}{\alpha}\Big)\,D^{-(\alpha-1)/\alpha}.$$

最后由 命题 5.1 的主阶等价 $\Delta L(D)\asymp S(D)$，即可得到 (16)；并且由于上式给出了 $S(D)$ 的渐近常数，在该 Zipf 情形下亦可推出同一常数对应 $\Delta L(D)\sim c_\alpha D^{-(\alpha-1)/\alpha}$。 □

Remark 5.1 (统计识别阈值的推广). 若学习模式需要至少出现 $m_0\ge 1$ 次，可取

$$q_k(D):=\Pr[X_k<m_0], \qquad X_k\sim\mathrm{Binomial}(D,p_k),$$

并相应定义 $\Delta L_{m_0}(D):=\sum_{k\ge 1}p_k\,q_k(D)$。在长尾主导区间 $p_k\ll 1$ 下，可用泊松近似（或直接

用二项分布的标准界）得到

$$\Pr[X_k < m_0] \approx \sum_{j=0}^{m_0-1} e^{-Dp_k} \frac{(Dp_k)^j}{j!}.$$

该表达是 e^{-Dp_k} 乘以 Dp_k 的多项式，因此主贡献仍来自满足 $Dp_k \asymp 1$ 的秩区间。在 Zipf 情形下由 $Dp_{k_\star} \asymp 1$ 得 $k_\star \asymp (D/Z)^{1/\alpha}$，并进而得到 $\Delta L_{m_0}(D) \asymp D^{-(\alpha-1)/\alpha}$：指数不变，仅常数依赖于 m_0。

由 定理 5.1，数据 scaling 指数为 $\alpha_D = (\alpha-1)/\alpha$。在该 missing-mass 代理下，$\alpha \to 1^+$ 时 $\alpha_D \to 0$，对应尾部极重时仅增加数据带来的下降会变得缓慢。

Remark 5.2 (模型规模 vs. 数据规模：两类 scaling law 的共同机制与差异). Section 4 与 Section 5 中的两类 scaling law 表面上形式不同：模型规模情形直接出现了对 Zipf 权重的截断尾和 $\sum_{k>k_\star(N)} p_k$；而数据规模情形仍是对全体 $k \geq 1$ 的求和 $\sum_{k\geq 1} p_k(1-p_k)^D$（或其指数近似 $\sum_{k\geq 1} p_k e^{-Dp_k}$）。但二者的本质机制是同构的：*有限资源在秩空间诱导一个有效覆盖前沿 $k_\star$，可约误差由未覆盖尾部的概率质量主导。*

模型规模 scaling：容量诱导的（近似）硬截断 在模型规模受限时，假设 3.1 and 4.1 抽象地刻画了" 能被有效拟合的模式集合"：存在临界秩 $k_\star(N) \asymp M(N)$，使得 $k \ll k_\star(N)$ 的模式残差 $q_k(N) \approx 0$，而 $k \gg k_\star(N)$ 的模式残差 $q_k(N) \approx 1$。因此

$$\Delta L(N) \approx \sum_{k\geq 1} p_k q_k(N) \asymp \sum_{k>k_\star(N)} p_k,$$

即由 Zipf 尾部质量直接给出幂律指数。

数据规模 scaling：覆盖诱导的软截断 在数据规模受限时，即使求和指标仍是 $k \geq 1$，但此时每一项都被一个随 D 变化的因子

$$q_k(D) = \Pr[X_k = 0] = (1-p_k)^D \approx e^{-Dp_k}$$

所调制：当 $Dp_k \gg 1$ 时 $q_k(D) \approx 0$（模式几乎必然出现过）；当 $Dp_k \ll 1$ 时 $q_k(D) \approx 1$（模式大概率未出现）。决定 $q_k(D)$ 从 1 过渡到 0 的临界尺度就是

$$Dp_k \asymp 1 \quad \Longleftrightarrow \quad p_k \asymp \frac{1}{D}.$$

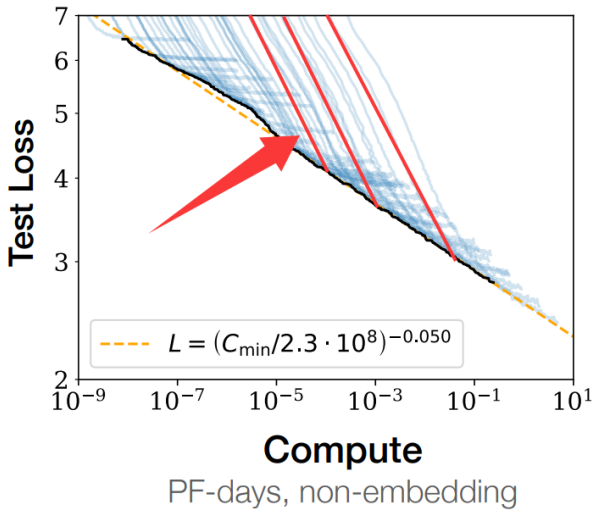
对 Zipf 分布 $p_k \asymp k^{-\alpha}$，这等价于 $k_\star(D) \asymp D^{1/\alpha}$。由于 e^{-Dp_k} 在 $k_\star(D)$ 附近表现为平滑阶跃函数，主阶上可把其视为对尾部的软截断，从而得到

$$\Delta L(D) = \sum_{k\geq 1} p_k e^{-Dp_k} \approx \sum_{k\gtrsim k_\star(D)} p_k \asymp \sum_{k>k_\star(D)} p_k \asymp k_\star(D)^{-(\alpha-1)} \asymp D^{-(\alpha-1)/\alpha}.$$

因此，数据 scaling 的幂律同样来自 Zipf 尾部质量，只不过前沿并非由容量假设硬性给出，而是由" 在 D 个样本里是否见到该模式" 的统计覆盖结构自发产生的软前沿。

6 第三种 Scaling Law：训练时间/算力 Scaling

本节在模型容量与数据覆盖均不构成主导瓶颈的条件下，推导 ΔL 随训练时间 τ（或累计 FLOPs）的下降规律。我们使用 引理 3.1 给出的残差动力学，并在 Zipf 分布下对求和做渐近估计。



由 (2) 与 (4)，

$$\Delta L(\tau) \approx \sum_{k\geq 1} p_k \, \exp\big(-c\,\tau\, p_k^{\beta}\big). \tag{18}$$

该求和的主导秩区间由 $c\tau p_k^{\beta} \approx 1$ 决定，即 $p_k \approx (c\tau)^{-1/\beta}$。Zipf 分布把 p_k 与 k 连接起来，从而产生幂律指数。

定理 6.1 (训练时间/算力 scaling law). 设 $p_k = \frac{1}{Z}k^{-\alpha}$，其中 $\alpha > 1$、$Z = \zeta(\alpha)$，并且 (4) 成立。则当 $\tau \to \infty$，

$$\boxed{\Delta L(\tau) \asymp \tau^{-(\alpha-1)/(\alpha\beta)}}. \tag{19}$$

更精确地，连续近似给出

$$\Delta L(\tau) \sim \tilde{c}_{\alpha,\beta}\, \tau^{-(\alpha-1)/(\alpha\beta)}, \qquad \tilde{c}_{\alpha,\beta} = \frac{1}{\alpha\beta}\left(\frac{1}{Z}\right)^{1/\alpha} c^{-(\alpha-1)/(\alpha\beta)}\Gamma\Big(\frac{\alpha-1}{\alpha\beta}\Big).$$

证明. 将 $p_k = \frac{1}{Z}k^{-\alpha}$ 代入 (18)，进行连续近似：

$$\Delta L(\tau) \approx \int_1^{\infty} \frac{1}{Z}x^{-\alpha} \exp\left(-c\,\tau\left(\frac{1}{Z}\right)^{\beta} x^{-\alpha\beta}\right) dx.$$

记

$$a := c\left(\frac{1}{Z}\right)^{\beta} > 0,$$

则指数项为 $\exp(-a\tau x^{-\alpha\beta})$。作换元

$$u = a\tau x^{-\alpha\beta} \quad \Longleftrightarrow \quad x = \left(\frac{a\tau}{u}\right)^{1/(\alpha\beta)}.$$

于是

$$dx = -\frac{1}{\alpha\beta}(a\tau)^{1/(\alpha\beta)} u^{-1/(\alpha\beta)-1}\, du.$$

并且

$$x^{-\alpha} = \Big(\frac{a\tau}{u}\Big)^{-1/\beta} = (a\tau)^{-1/\beta} u^{1/\beta}.$$

代入积分得

$$\begin{aligned}\Delta L(\tau) &\approx \int_{u=a\tau}^{0} \frac{1}{Z}\,(a\tau)^{-1/\beta} u^{1/\beta}\, e^{-u} \left[-\frac{1}{\alpha\beta}(a\tau)^{1/(\alpha\beta)} u^{-1/(\alpha\beta)-1}\, du\right] \\ &= \frac{1}{Z}\cdot\frac{1}{\alpha\beta}\,(a\tau)^{-(\alpha-1)/(\alpha\beta)} \int_0^{a\tau} e^{-u}\, u^{(\alpha-1)/(\alpha\beta)-1}\, du.\end{aligned}$$

当 $\tau \to \infty$ 时，上限 $a\tau \to \infty$。由于 $(\alpha-1)/(\alpha\beta) > 0$，积分收敛到

$$\int_0^{\infty} e^{-u}\, u^{(\alpha-1)/(\alpha\beta)-1}\, du = \Gamma\Big(\frac{\alpha-1}{\alpha\beta}\Big).$$

因此

$$\Delta L(\tau) \sim \frac{1}{Z}\cdot\frac{1}{\alpha\beta}\, a^{-(\alpha-1)/(\alpha\beta)} \Gamma\Big(\frac{\alpha-1}{\alpha\beta}\Big)\, \tau^{-(\alpha-1)/(\alpha\beta)}.$$

将 $a = c(1/Z)^{\beta}$ 展开并整理得到定理中的常数表达式与指数 (19)。 □

Remark 6.1 (与物理算力的对应). 若训练过程中模型规模 N 与 batch size B 固定，则单步计算量满足 $\mathrm{FLOPs}_{\mathrm{step}} \asymp NB$（稠密 Transformer 仅差一个实现常数），总算力

$$C_{\mathrm{run}} = \mathrm{FLOPs}_{\mathrm{step}} \cdot \tau \;\asymp\; \tau.$$

因此 (19) 等价于

$$\Delta L(C_{\mathrm{run}}) \;\asymp\; C_{\mathrm{run}}^{-(\alpha-1)/(\alpha\beta)}.$$

若 N 或 B 随训练改变，则应以实际累计 FLOPs 替代 τ 作归一化；在该归一化下，指数仍由 (α, β) 决定。

7 Rethink Compute Scaling 的最弱条件

前文在 Section 6 中以指数核 $q_k(\tau) \approx \exp(-c\tau p_k^{\beta})$ 为代表，推导了 $\Delta L(\tau)$ 的幂律下降。本节尝试做一个再反思，反推使该结论成立所需要的最弱结构条件。核心结论是：幂律指数来自 Zipf 长尾与单参数自相似的换元结构；核的具体形状只进入前因子，且恰好由 Mellin 变换给出。

问题重述 我们希望解释的现象是：在某个非饱和的 scaling 区间内，可约损失随训练时间（或等价的累计算力）呈稳定幂律

$$\Delta L(\tau) \;\sim\; K\,\tau^{-s}, \qquad \tau \to \infty, \tag{20}$$

其中 $s > 0$ 与数据分布与优化动力学有关，我们首先要问：什么样的微观结构能在求和意义下产生稳定的幂律?

模式分解必须是非负加权求和 为了让 (20) 表示可约误差的下降，一个几乎不可再弱的起点是模式分解：存在可数模式集合 $\{1, 2, \dots\}$，权重序列 $p_k \geq 0$ 且 $\sum_k p_k = 1$，以及残差强度 $q_k(\tau) \in [0, 1]$，使得

$$\Delta L(\tau) \;\approx\; \sum_{k\geq 1} p_k\, q_k(\tau). \tag{21}$$

这一步并不涉及网络谱、线性化或可学习性先验，只要求宏观误差可以写成各模式残差的非负加权和。从倒推角度看，(21) 之所以重要，是因为幂律的来源必然是某类" 尾部求和" 的渐近；没有这种可加结构，就谈不上用长尾统计去解释幂律。

长尾必须足够重　接下来，(20) 的幂律形式本身就暗示：主导误差不会来自有限个头部模式，而必须来自随 τ 增大不断向尾部移动的一段秩区间。换言之，p_k 必须是长尾的。最自然且足够一般的形式是正则变差（Zipf 是其特例）：

$$p_k \;=\; \frac{1}{Z}\,k^{-\alpha}L_p(k), \qquad \alpha>1, \tag{22}$$

其中 $L_p(\cdot)$ 为缓变函数，Z 为归一化常数（可吸收入 L_p 的定义）。在后续计算里，为了突出指数结构，我们仍以标准 Zipf 形式 $L_p\equiv 1$ 叙述。

满足单参数自相似　若残差 $q_k(\tau)$ 随 τ 单调减小，则在固定 τ 下通常存在一个过渡区间：对足够频繁的模式（小 k，大 p_k）有 $q_k(\tau)\approx 0$，对足够稀有的模式（大 k，小 p_k）有 $q_k(\tau)\approx 1$，而主导贡献（决定 $\Delta L(\tau)$ 主阶数量级的）来自过渡带。为了让 (20) 的指数 s 稳定，我们并不需要知道 $q_k(\tau)$ 的全局形式；真正必要的是：在决定主导项的过渡带附近，$q_k(\tau)$ 只能依赖于一个单一的缩放变量。最弱且清晰的表达是存在 $\beta>0$ 与尺度常数 $a>0$，以及某个核函数 $g:[0,\infty)\to[0,1]$，使得在主导秩窗口内有

$$q_k(\tau) \;\approx\; g\big(a\tau p_k^{\beta}\big), \tag{23}$$

并且这种近似并非点态（对每个固定 k）成立，而是必须覆盖主导窗口 $k\sim k_\star(\tau)$。我们进一步考察所谓 $k\sim k_\star(\tau)$ 的主导窗口，像前文说的，

- **头部（小 k，大 p_k）**：由于 $a\tau p_k^{\beta}\gg 1$，所以 $g(\cdot)\approx 0$，这些项几乎不贡献；
- **极尾（大 k，小 p_k）**：由于 $a\tau p_k^{\beta}\ll 1$，所以 $g(\cdot)\approx 1$，单个项贡献约 $\approx p_k$，但 p_k 本身很小，需要很多项叠加；
- **中间过渡区**：当

  $$a\tau p_k^{\beta}\approx 1$$

  时，g 从接近 1 过渡到接近 0，这些 k 的贡献既不被 g 压到 0，也不完全是 1，而是 "恰好在切换"。这段区间通常给出主导的幂律标度。

于是定义**临界秩** $k_\star(\tau)$ 为满足

$$a\tau p_{k_\star}^{\beta} \asymp 1 \tag{24}$$

的解。在 Zipf 情形 $\big(p_k=\frac{1}{Z}k^{-\alpha}\big)$ 下，

$$k_\star(\tau)\asymp\Big(\frac{a\tau}{Z^{\beta}}\Big)^{1/(\alpha\beta)}. \tag{25}$$

之所以这段叫 "主导"，是因为幂律指数来自把求和的主要贡献集中到这段过渡区，然后做换元得到的 τ^{-s}。如果对 $q_k(\tau)\approx g(a\tau p_k^{\beta})$ 的近似只在很头或很尾成立，但在这段过渡区不成立，那么推导出来的指数就没有保障，因为真正贡献最大的那批项没有被控制住。因此，从倒推角度看，指数核 $q_k(\tau)\approx\exp(-c\tau p_k^{\beta})$ 并不是必要的；必要的是 $q_k(\tau)$ 在主导窗口上具备单参数缩放结构 (23)。为了保证该结构确实对应" 学得越久残差越小"，我们只需要 $g(0)=1$、$g(u)\to 0$，以及 g 有界（最好再假设非增，以便后续把求和替换为积分）。

在该结构下，(21) 变为

$$\Delta L(\tau) \;\approx\; \sum_{k\geq 1} p_k \, g\big(a\tau p_k^{\beta}\big). \tag{26}$$

此时 compute scaling 的问题被彻底转化为一个长尾序列与缩放核的渐近求和问题。

换元显示指数由 (α,β) 决定，核只影响常数 在 Zipf 情形 $p_k = \frac{1}{Z}k^{-\alpha}$ 下，用连续近似（其严格化只需 p_k 单调且 g 有界/非增）把 (26) 视为积分：

$$\Delta L(\tau) \;\approx\; \int_1^{\infty} \frac{1}{Z} x^{-\alpha} \, g\Big(a\tau (1/Z)^{\beta} x^{-\alpha\beta}\Big)\, dx.$$

令 $A := a(1/Z)^{\beta}$，作换元 $u = A\tau x^{-\alpha\beta}$，则直接得到

$$\Delta L(\tau) \;\approx\; \frac{1}{\alpha\beta}\Big(\frac{1}{Z}\Big)^{1/\alpha} a^{-s}\, \tau^{-s} \int_0^{A\tau} u^{s-1} g(u)\, du, \qquad s := \frac{\alpha-1}{\alpha\beta}. \tag{27}$$

这一步已经揭示了 (20) 的指数来自何处：s 完全由长尾指数 α 与频率敏感度 β 决定，而与 g 的具体形状无关。剩下的问题是：何时可以把 $\int_0^{A\tau}$ 推到 $\int_0^{\infty}$，从而得到稳定常数 K？

倒推到这里，最弱且可检验的充分条件自然出现：只要

$$\int_0^{\infty} u^{s-1} g(u)\, du \;\in\; (0,\infty), \tag{28}$$

则 (27) 中的截断积分收敛到一个有限非零常数，于是得到纯幂律

$$\Delta L(\tau) \;\sim\; \frac{1}{\alpha\beta}\Big(\frac{1}{Z}\Big)^{1/\alpha} a^{-s} \Big[\int_0^{\infty} u^{s-1} g(u)\, du\Big]\, \tau^{-s}. \tag{29}$$

括号内正是核 g 在点 s 处的 Mellin 变换 $M_g(s)$。当 $g(u) = e^{-u}$ 时，它退化为 $\Gamma(s)$，即前文指数核推导得到的常数。

总结 综上，从 (20) 的纯幂律结论反推，compute scaling law 的最弱结构可概括为三件事：其一，误差在模式空间上具有非负加权求和的分解 (21)；其二，权重序列在秩空间上具有长尾（至少正则变差，Zipf 为代表）(22)；其三，动力学在主导窗口上满足单参数缩放 (23)，并且缩放核 g 的 Mellin 常数在 $s = \frac{\alpha-1}{\alpha\beta}$ 处有限非零 (28)。在这些条件下，指数 s 由 (α,β) 唯一确定，而核的具体形状只通过 Mellin 变换进入前因子。这解释了为何指数核并非本质：它只是一个便于计算 $M_g(s)$ 的特例，而不是幂律指数的来源。

8 联合资源下的组合律与两段式结构

前文分别在单一资源受限的情形下得到了幂律衰减。实际预训练中 N, D, τ 同时有限，因此需要一个组合律刻画 $\Delta L(N, D, \tau)$ 的主导阶，并解释经验拟合中常用的幂律加和形式。本节给出一个结构性结论：在非饱和的 scaling 区间内，ΔL 与三项瓶颈的最大者同阶；加和形式可视为对 max 的平滑替代，不改变主导指数。

定义三个单变量瓶颈项

$$\varepsilon_N(N) = A\, N^{-\alpha_N}, \qquad \varepsilon_D(D) = B\, D^{-\alpha_D}, \qquad \varepsilon_\tau(\tau) = G\, \tau^{-\alpha_\tau}, \tag{30}$$

其中 $A, B, G > 0$ 为常数，$\alpha_N, \alpha_D, \alpha_\tau > 0$ 为 scaling 指数。在本文 Zipf–模式框架中，$\alpha_N = \gamma(\alpha-1)$、$\alpha_D = \frac{\alpha-1}{\alpha}$、$\alpha_\tau = \frac{\alpha-1}{\alpha\beta}$，但以下结论不依赖其具体形式。

命题 8.1 (loss 的 Max 结构). 在 scaling 区间内，若存在下界

$$\Delta L(N,D,\tau) \gtrsim \varepsilon_N(N), \qquad \Delta L(N,D,\tau) \gtrsim \varepsilon_D(D), \qquad \Delta L(N,D,\tau) \gtrsim \varepsilon_\tau(\tau), \tag{31}$$

并且存在训练策略使得联合上界成立

$$\Delta L(N,D,\tau) \lesssim \varepsilon_N(N) + \varepsilon_D(D) + \varepsilon_\tau(\tau), \tag{32}$$

则有数量级等价

$$\boxed{\Delta L(N,D,\tau) \asymp \max\big(\varepsilon_N(N),\, \varepsilon_D(D),\, \varepsilon_\tau(\tau)\big).} \tag{33}$$

证明. 由 (31) 得到

$$\Delta L(N,D,\tau) \gtrsim \max(\varepsilon_N, \varepsilon_D, \varepsilon_\tau).$$

另一方面，对任意 $x, y, z \geq 0$，有基本不等式

$$\max(x,y,z) \leq x + y + z \leq 3\max(x,y,z). \tag{34}$$

将 $(x,y,z) = (\varepsilon_N, \varepsilon_D, \varepsilon_\tau)$ 代入并结合 (32)，得

$$\Delta L(N,D,\tau) \lesssim \varepsilon_N + \varepsilon_D + \varepsilon_\tau \leq 3\max(\varepsilon_N, \varepsilon_D, \varepsilon_\tau),$$

因此

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

与下界合并即得 (33)。 □

由 (33)，当固定 (N,D) 并随 τ 增加训练时间时，ΔL 的主导项从 $\varepsilon_\tau(\tau)$ 过渡到静态瓶颈 $\max(\varepsilon_N(N), \varepsilon_D(D))$。因此学习曲线具有两段结构：早期由收敛瓶颈控制，后期由静态瓶颈控制。

固定一次训练运行的 (N,D)，记静态瓶颈为

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

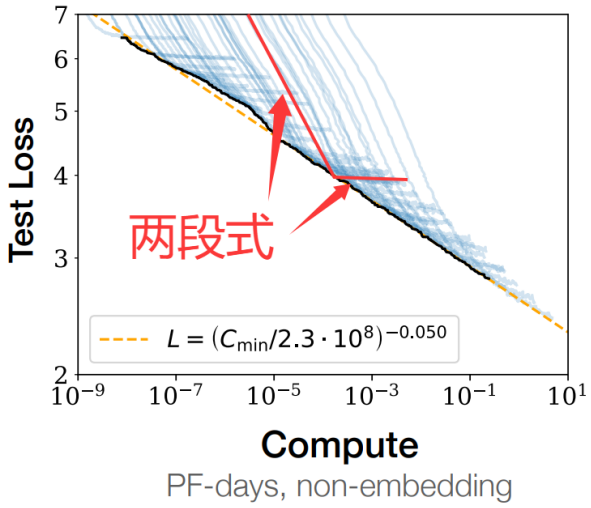
则

$$\Delta L(N,D,\tau) \asymp \max\big(\varepsilon_{\text{stat}}(N,D),\, \varepsilon_\tau(\tau)\big). \tag{35}$$

定义切换时间 $\tau_\star$ 为两者同阶时的尺度：

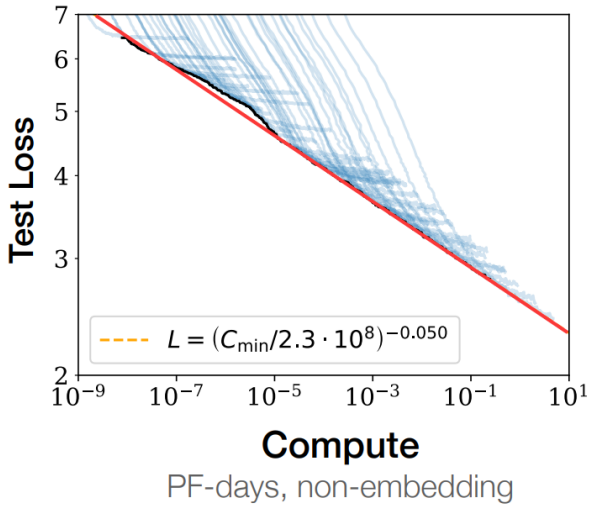
$$\varepsilon_\tau(\tau_\star) = \varepsilon_{\text{stat}}(N,D) \quad \Longleftrightarrow \quad \tau_\star = \Big(\frac{G}{\varepsilon_{\text{stat}}(N,D)}\Big)^{1/\alpha_\tau}.$$

于是 $\tau \ll \tau_\star$ 时 $\Delta L(N,D,\tau) \asymp \varepsilon_\tau(\tau) \propto \tau^{-\alpha_\tau}$；$\tau \gg \tau_\star$ 时 $\Delta L(N,D,\tau) \asymp \varepsilon_{\text{stat}}(N,D)$。



9 第四种 Scaling Law：Optimal compute point Scaling

本节推导两段式学习曲线转折点处的 loss–compute 幂律关系。该转折点由 ε_τ 与静态瓶颈同阶确定，并可进一步写成算力 C 的函数。



定义 9.1 (Optimal compute point / 转折点). 给定一次训练运行的 (N, D)，定义其 optimal compute point 为任意满足

$$\varepsilon_\tau(\tau_\star) \;=\; \varepsilon_{\text{stat}}(N, D) \tag{36}$$

的 $\tau_\star = \tau_\star(N, D)$。并定义该点对应的损失高度与算力分别为

$$\Delta L_\star(N, D) := \Delta L(N, D, \tau_\star), \qquad C_\star(N, D) := \kappa' N \tau_\star(N, D),$$

其中 $\kappa' > 0$ 为实现常数。

由定义可直接解出 $\tau_\star$ 并得到 $C_\star$ 的表达。为便于后续消元，我们先给出一般形式，再在特定族上写成单变量幂律。

命题 9.1 (Optimal compute point 的显式表达). 在 (30) 的幂律瓶颈假设下，即

$$\varepsilon_N(N) = AN^{-\alpha_N}, \quad \varepsilon_D(D) = BD^{-\alpha_D}, \quad \varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau}, \qquad A, B, G > 0,\ \alpha_N, \alpha_D, \alpha_\tau > 0,$$

则 定义 9.1 中的 $\tau_\star, \Delta L_\star, C_\star$ 满足

$$\tau_\star(N, D) \;=\; \Big(\frac{G}{\varepsilon_{\text{stat}}(N, D)}\Big)^{1/\alpha_\tau} \;=\; \Big(\frac{G}{\max(AN^{-\alpha_N},\, BD^{-\alpha_D})}\Big)^{1/\alpha_\tau}, \tag{37}$$

$$\Delta L_\star(N, D) \;\asymp\; \varepsilon_{\text{stat}}(N, D) \;=\; \max(AN^{-\alpha_N},\, BD^{-\alpha_D}), \tag{38}$$

$$C_\star(N, D) \;\asymp\; N\Big(\frac{G}{\max(AN^{-\alpha_N},\, BD^{-\alpha_D})}\Big)^{1/\alpha_\tau}. \tag{39}$$

特别地，在两种静态瓶颈主导区域内，

$$C_\star(N, D) \;\asymp\; \begin{cases} N^{1+\alpha_N/\alpha_\tau}, & AN^{-\alpha_N} \gtrsim BD^{-\alpha_D}, \\ N\, D^{\alpha_D/\alpha_\tau}, & BD^{-\alpha_D} \gtrsim AN^{-\alpha_N}. \end{cases} \tag{40}$$

证明. 由 (36) 与 $\varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau}$，解得

$$G\tau_\star^{-\alpha_\tau} = \varepsilon_{\text{stat}}(N, D) \quad \Longleftrightarrow \quad \tau_\star = \Big(\frac{G}{\varepsilon_{\text{stat}}(N, D)}\Big)^{1/\alpha_\tau},$$

即 (37)。代回 (35) 得

$$\Delta L_\star(N, D) \asymp \varepsilon_{\text{stat}}(N, D),$$

即 (38)。再由 $C_\star = \kappa' N\tau_\star$ 得 (39)。最后，根据 $\max(AN^{-\alpha_N}, BD^{-\alpha_D})$ 的两种主导情形，分别化简得到 (40)。 □

为了把 (N, D) 消去并得到 $\Delta L_\star$ 关于 $C_\star$ 的单变量幂律，需要指定一族运行。一个自然的选择是令静态两项同阶，从而避免在不同运行间引入不同主导项导致的指数切换。

假设 9.1 (静态瓶颈均衡族). 考虑一族训练运行 $\{(N, D)\}$ 满足

$$AN^{-\alpha_N} \;\asymp\; BD^{-\alpha_D}. \tag{41}$$

定理 9.1 (Optimal compute point 的 loss–compute 幂律). 在 假设 9.1 下，optimal compute point 的损失高度与算力满足

$$\boxed{\Delta L_\star(C) \;\propto\; C^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}}.} \tag{42}$$

等价地，存在常数 $c_1, c_2 > 0$ 使得对该族上所有运行有

$$c_1\, C_\star^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}} \;\le\; \Delta L_\star(N, D) \;\le\; c_2\, C_\star^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}}.$$

证明. 由 假设 9.1，静态瓶颈满足

$$\varepsilon_{\text{stat}}(N, D) \asymp AN^{-\alpha_N}.$$

由 命题 9.1，

$$\tau_\star(N,D) \asymp \Big(\frac{G}{A}\Big)^{1/\alpha_\tau} N^{\alpha_N/\alpha_\tau}, \qquad C_\star(N,D) \asymp N\tau_\star \asymp N^{1+\alpha_N/\alpha_\tau}.$$

同时

$$\Delta L_\star(N,D) \asymp \varepsilon_{\text{stat}}(N,D) \asymp AN^{-\alpha_N}.$$

消去 N：由 $C_\star \asymp N^{1+\alpha_N/\alpha_\tau}$ 得

$$N \asymp C_\star^{\frac{1}{1+\alpha_N/\alpha_\tau}} = C_\star^{\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}}.$$

代入 $\Delta L_\star \asymp N^{-\alpha_N}$ 得

$$\Delta L_\star \asymp C_\star^{-\alpha_N\cdot\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}} = C_\star^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}},$$

即 (42)。 □

Remark 9.1 (与 Kaplan [1] 的 compute-efficient 前沿一致性). Kaplan [1] 等在无限数据极限下拟合两变量学习曲线

$$L(N,S) \;=\; \Big(\frac{N_c}{N}\Big)^{\alpha_N} + \Big(\frac{S_c}{S}\Big)^{\alpha_S}, \tag{43}$$

并在预算约束 $C \propto NS$ 下最小化 $L(N,S)$，其最优前沿满足

$$L_{\text{opt}}(C) \;\propto\; C^{-\frac{\alpha_N\alpha_S}{\alpha_N+\alpha_S}}. \tag{44}$$

将 定理 9.1 中的训练时间指数识别为 Kaplan [1] 的步数指数 $\alpha_\tau = \alpha_S$，并将 $C_\star = \kappa' N\tau_\star$ 视作 $C \propto NS$ 的同一预算近似，则 (42) 与 (44) 逐项一致。此外，[1] 在考虑 critical batch 校正时给出 $L(C) \propto C^{-\alpha_C}$，其中 $\alpha_C^{-1} = \alpha_N^{-1} + \alpha_S^{-1} + \alpha_B^{-1}$；该情形对应于在算力定义中显式保留 batch 对预算的额外约束项。

10 Compute-Optimal：预算约束下的均衡条件

本节在瓶颈模型

$$\Delta L(N,D,\tau) \;\asymp\; \max\Big(\varepsilon_N(N),\, \varepsilon_D(D),\, \varepsilon_\tau(\tau)\Big), \qquad \varepsilon_N(N) = AN^{-\alpha_N},\ \varepsilon_D(D) = BD^{-\alpha_D},\ \varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau} \tag{45}$$

下，推导给定算力预算 C 时的 compute-optimal 前沿。推导只使用两个结构性事实：(i) 在乘性预算约束下，max 型目标的最优解出现在活跃项同阶处；(ii) 一项随决策变量单调下降、另一项单调上升时，max 的极小值由交点给出。

经验上，Kaplan 等 [1] 与 Hoffmann 等（Chinchilla）[2] 给出了不同的最优分配幂律。下文给出统一解释：二者对应于同一 (45) 在不同可行域假设下的约束最优问题。

10.1 情形 A：收敛瓶颈次主导（$C \asymp ND$）

先考虑训练时间足够大，使得收敛瓶颈不超过静态瓶颈的情形：

$$\varepsilon_\tau(\tau) \;\lesssim\; \max\big(\varepsilon_N(N),\, \varepsilon_D(D)\big). \tag{46}$$

此时

$$\Delta L(N,D) \;\asymp\; \max\big(AN^{-\alpha_N},\, BD^{-\alpha_D}\big). \tag{47}$$

对稠密 Transformer 的常用预算近似为

$$C = \kappa N D, \tag{48}$$

其中 $\kappa > 0$ 为实现常数。

命题 10.1 (收敛瓶颈次主导时的最优分配 [2]). 在 (47) 与预算约束 $C = \kappa N D$ 下，约束最优化

$$\min_{N,D} \ \max\bigl(AN^{-\alpha_N},\, BD^{-\alpha_D}\bigr) \quad \text{s.t.} \quad ND = \frac{C}{\kappa}$$

的任一最优解都满足均衡条件

$$AN^{-\alpha_N} \asymp BD^{-\alpha_D}, \tag{49}$$

并且存在与 C 无关的常数使得

$$N_{\mathrm{opt}}(C) \propto C^{\frac{\alpha_D}{\alpha_N+\alpha_D}}, \qquad D_{\mathrm{opt}}(C) \propto C^{\frac{\alpha_N}{\alpha_N+\alpha_D}}, \qquad \Delta L_{\mathrm{opt}}(C) \propto C^{-\frac{\alpha_N\alpha_D}{\alpha_N+\alpha_D}}. \tag{50}$$

证明. 在约束 $ND = C/\kappa$ 下写

$$D = \frac{C}{\kappa} N^{-1}.$$

代入 (47) 的两项得到

$$\phi(N) = \max\Bigl(AN^{-\alpha_N},\ B\Bigl(\frac{C}{\kappa}N^{-1}\Bigr)^{-\alpha_D}\Bigr) = \max\Bigl(AN^{-\alpha_N},\ B\kappa^{\alpha_D}C^{-\alpha_D}N^{\alpha_D}\Bigr).$$

第一项随 N 单调下降，第二项随 N 单调上升，因此 $\phi(N)$ 的极小值在两项同阶处取得，即 $AN^{-\alpha_N} \asymp B\kappa^{\alpha_D}C^{-\alpha_D}N^{\alpha_D}$，等价于 (49)。

由同阶条件解出 N 的 C-指数：整理得

$$N^{\alpha_N+\alpha_D} \asymp C^{\alpha_D}, \quad \Longrightarrow \quad N_{\mathrm{opt}}(C) \propto C^{\alpha_D/(\alpha_N+\alpha_D)}.$$

再由 $D = C/(\kappa N)$ 得 $D_{\mathrm{opt}}(C) \propto C^{\alpha_N/(\alpha_N+\alpha_D)}$。最优损失在均衡点处两项同阶，因此

$$\Delta L_{\mathrm{opt}}(C) \asymp AN_{\mathrm{opt}}(C)^{-\alpha_N} \propto C^{-\alpha_N\alpha_D/(\alpha_N+\alpha_D)}.$$

□

10.2 情形 B：数据瓶颈次主导（$C \asymp N\tau$）

再考虑数据规模足够大，使得数据瓶颈不超过 $(\varepsilon_N, \varepsilon_\tau)$ 的情形：

$$\varepsilon_D(D) \lesssim \max\bigl(\varepsilon_N(N), \varepsilon_\tau(\tau)\bigr). \tag{51}$$

此时

$$\Delta L(N,\tau) \asymp \max\bigl(AN^{-\alpha_N}, G\tau^{-\alpha_\tau}\bigr). \tag{52}$$

常用预算近似为

$$C = \kappa' N \tau, \tag{53}$$

其中 $\kappa' > 0$ 为实现常数。

命题 10.2 (数据瓶颈次主导时的最优分配 [1]). 在 (52) 与预算约束 $C = \kappa' N\tau$ 下，约束最优化

$$\min_{N,\tau} \ \max\bigl(AN^{-\alpha_N},\, G\tau^{-\alpha_\tau}\bigr) \quad \text{s.t.} \quad N\tau = \frac{C}{\kappa'}$$

的任一最优解都满足均衡条件

$$AN^{-\alpha_N} \ \asymp \ G\tau^{-\alpha_\tau}, \tag{54}$$

并且存在与 C 无关的常数使得

$$N_{\mathrm{opt}}(C) \ \propto \ C^{\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}}, \qquad \tau_{\mathrm{opt}}(C) \ \propto \ C^{\frac{\alpha_N}{\alpha_N+\alpha_\tau}}, \qquad \Delta L_{\mathrm{opt}}(C) \ \propto \ C^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}}. \tag{55}$$

证明. 在约束 $N\tau = C/\kappa'$ 下写

$$\tau = \frac{C}{\kappa'}N^{-1}.$$

代入 (52) 得

$$\psi(N) = \max\Bigl(AN^{-\alpha_N},\, G\Bigl(\frac{C}{\kappa'}N^{-1}\Bigr)^{-\alpha_\tau}\Bigr) = \max\Bigl(AN^{-\alpha_N},\, G\kappa'^{\alpha_\tau}C^{-\alpha_\tau}N^{\alpha_\tau}\Bigr).$$

同样地，第一项随 N 单调下降，第二项随 N 单调上升，因此最优点在两项同阶处取得：

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

等价于 (54)。整理得

$$N^{\alpha_N+\alpha_\tau} \asymp C^{\alpha_\tau} \quad \Longrightarrow \quad N_{\mathrm{opt}}(C) \propto C^{\alpha_\tau/(\alpha_N+\alpha_\tau)}.$$

再由 $\tau = C/(\kappa' N)$ 得 $\tau_{\mathrm{opt}}(C) \propto C^{\alpha_N/(\alpha_N+\alpha_\tau)}$。最优损失由均衡点处任一活跃项给出，从而

$$\Delta L_{\mathrm{opt}}(C) \asymp AN_{\mathrm{opt}}(C)^{-\alpha_N} \propto C^{-\alpha_N\alpha_\tau/(\alpha_N+\alpha_\tau)}.$$

□

10.3 代入 Zipf–模式框架的指数映射

将本文框架得到的指数

$$\alpha_N = \gamma(\alpha-1), \qquad \alpha_D = \frac{\alpha-1}{\alpha}, \qquad \alpha_\tau = \frac{\alpha-1}{\alpha\beta} \tag{56}$$

代入 (50) 与 (55)，即可得到最优分配的显式幂律。特别地，在情形 A（预算 $C \asymp ND$）下，

$$N_{\mathrm{opt}}(C) \ \propto \ C^{\frac{1}{1+\alpha\gamma}}, \qquad D_{\mathrm{opt}}(C) \ \propto \ C^{\frac{\alpha\gamma}{1+\alpha\gamma}}, \qquad \Delta L_{\mathrm{opt}}(C) \ \propto \ C^{-\frac{\gamma(\alpha-1)}{1+\alpha\gamma}}, \tag{57}$$

而在情形 B（预算 $C \asymp N\tau$）下，

$$N_{\mathrm{opt}}(C) \ \propto \ C^{\frac{1}{1+\alpha\beta\gamma}}, \qquad \tau_{\mathrm{opt}}(C) \ \propto \ C^{\frac{\alpha\beta\gamma}{1+\alpha\beta\gamma}}, \qquad \Delta L_{\mathrm{opt}}(C) \ \propto \ C^{-\frac{\gamma(\alpha-1)}{1+\alpha\beta\gamma}}. \tag{58}$$

11 结论

本文以模式分解

$$\Delta L \approx \sum_{k\geq 1} p_k q_k \tag{59}$$

为起点，在 Zipf 统计假设（假设 2.1）下，引入有效前沿抽象（假设 3.1）与频率依赖的指数收敛动力学（引理 3.1），在不依赖 NTK 线性化或核谱幂律先验的前提下，解析推出四类常见 scaling law。推导的共同结构是：有限资源在秩空间诱导临界秩 $k_\star$，而 ΔL 的主阶由 Zipf 尾部质量决定。

三类单变量瓶颈可写为

$$\varepsilon_N(N) = AN^{-\alpha_N}, \qquad \varepsilon_D(D) = BD^{-\alpha_D}, \qquad \varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau}, \tag{60}$$

其中指数由数据统计参数 α、容量映射指数 γ 与动力学指数 β 给出

$$\alpha_N = \gamma(\alpha-1), \qquad \alpha_D = \frac{\alpha-1}{\alpha}, \qquad \alpha_\tau = \frac{\alpha-1}{\alpha\beta}. \tag{61}$$

对应的四类幂律结论（忽略实现相关常数）为

$$\text{(i) Model scaling:} \quad \Delta L(N) \asymp M(N)^{-(\alpha-1)} \sim N^{-\gamma(\alpha-1)}, \tag{62}$$

$$\text{(ii) Data scaling:} \quad \Delta L(D) \asymp D^{-(\alpha-1)/\alpha}, \tag{63}$$

$$\text{(iii) Time/compute scaling:} \quad \Delta L(\tau) \asymp \tau^{-(\alpha-1)/(\alpha\beta)}, \tag{64}$$

$$\text{(iv) Optimal point:} \quad \Delta L_\star(C) \propto C^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}} \quad \text{(在静态均衡族上).} \tag{65}$$

在联合资源有限的 scaling 区间内，总体可约损失由三项瓶颈中的主导者决定（命题 8.1）：

$$\Delta L(N,D,\tau) \asymp \max\big(\varepsilon_N(N), \varepsilon_D(D), \varepsilon_\tau(\tau)\big). \tag{66}$$

在乘性预算约束下，compute-optimal 解满足活跃瓶颈同阶，从而在两类常见预算形式下分别给出

$$\begin{aligned}
&(\text{预算} C \asymp ND,\ \varepsilon_\tau \text{ 次主导}): \quad N_{\text{opt}}(C) \propto C^{\frac{\alpha_D}{\alpha_N+\alpha_D}},\ D_{\text{opt}}(C) \propto C^{\frac{\alpha_N}{\alpha_N+\alpha_D}},\ \Delta L_{\text{opt}}(C) \propto C^{-\frac{\alpha_N\alpha_D}{\alpha_N+\alpha_D}},\\
&(\text{预算} C \asymp N\tau,\ \varepsilon_D \text{ 次主导}): \quad N_{\text{opt}}(C) \propto C^{\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}},\ \tau_{\text{opt}}(C) \propto C^{\frac{\alpha_N}{\alpha_N+\alpha_\tau}},\ \Delta L_{\text{opt}}(C) \propto C^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}}.
\end{aligned} \tag{67}$$

最后，上述推导给出了可证伪的指数一致性关系。若经验拟合得到单变量指数 $(\hat\alpha_N, \hat\alpha_D, \hat\alpha_\tau)$，则该框架要求存在 (α, β, γ) 使得

$$\hat\alpha_D = \frac{\alpha-1}{\alpha}, \qquad \hat\alpha_\tau = \frac{1}{\beta}\hat\alpha_D, \qquad \hat\alpha_N = \gamma(\alpha-1), \tag{68}$$

并进一步要求 compute-optimal 前沿指数满足

$$\hat\alpha_{\text{opt}} = \frac{\hat\alpha_N\hat\alpha_D}{\hat\alpha_N+\hat\alpha_D} \ \ (\text{预算} C \asymp ND) \qquad \text{或} \qquad \hat\alpha_{\text{opt}} = \frac{\hat\alpha_N\hat\alpha_\tau}{\hat\alpha_N+\hat\alpha_\tau} \ \ (\text{预算} C \asymp N\tau). \tag{69}$$

因此，该框架不仅给出幂律形式，也给出不同实验拟合指数之间的检验约束。

参考文献

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Jordan Hoffmann et al. Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.15556.