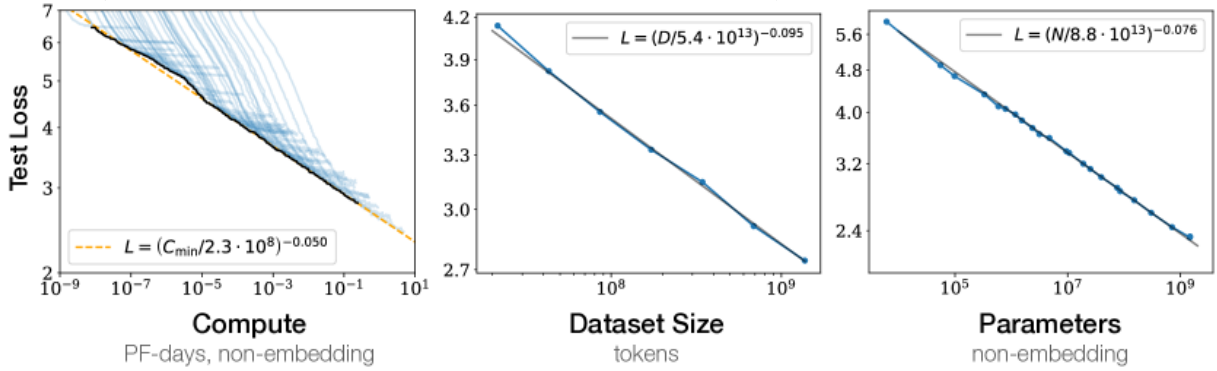
Can We Derive Scaling Law From First Principles?

Jiaxuan Zou

2025 年 12 月 30 日

摘要

LLM 预训练中损失函数随模型规模 N、数据规模 D 与计算量 C 呈现稳定的幂律（scaling laws），并且在给定算力预算下存在近似幂律的 compute-optimal 前沿，也就是说 optimal compute point 的连线也构成幂律。（一共四种常见 scaling law）已有理论常依赖对核谱或 Teacher 系数的特定假设。本文旨在不预设 NTK 线性化或特定谱分布的前提下，提出一种基于数据统计特性的第一性解释框架：证明幂律源于语言数据的无尺度长尾结构与梯度下降的隐式排序偏好。



我们将语言建模抽象为对服从 Zipf 分布（$p_k \propto k^{-\alpha}$）的模式（modes）的逐步掌握过程。基于梯度下降优先降低主导误差项的“贪婪覆盖”性质，总体可约损失被建模为未掌握模式权重的尾部积分。在此框架下，我们解析推导得到模型规模 N、数据规模 D 与计算量 C 的幂律缩放形式，其指数分别为 $-\gamma(\alpha-1)$、$-\frac{\alpha-1}{\alpha}$ 与 $-\frac{\alpha-1}{\alpha\beta}$。进一步地，我们在统一的瓶颈几何下重构了 compute-optimal 前沿，证明 Kaplan 与 Chinchilla 工作的一致性。

1 引言

LLM 预训练表现出测试 loss 随模型规模 N、数据规模 D 及计算预算 C 呈幂律衰减的 scaling laws。这背后一定有深刻的机理未被挖掘清楚，但从理论上推导该规律面临深度网络非凸优化与 feature learning 的复杂性挑战。现有理论常依赖于 NTK、KRR 线性化近似或预设特征谱的幂律分布，这种处理方式虽简化了数学结构，却在一定程度上引入了先验假设的循环论证风险。

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。

所以再三取舍之下，本文提出一种不依赖特定核谱假设的有效理论框架，从数据统计特性与优化动力学的第一性原理出发推导 Scaling Laws。我们将语言建模抽象为对可数模式集合（Modes）的序贯学习过程。由于语言数据普遍存在的长尾特性，模式的重要性（即出现概率 p_k）按其秩 k 服从无尺度的 Zipf 分布：

$$p_k \propto k^{-\alpha}, \qquad \alpha > 1. \tag{1}$$

在此基础上，我们引入梯度下降算法的**隐式偏好（Implicit Bias）**假设：在资源受限条件下，优化过程并非均匀地降低所有模式的误差，而是表现出频率依赖的排序机制，即优先拟合高频主导模式。

Remark 1.2. 我觉得这个假设并不是很强，因为在 NN+GD 的设定下，NN 经常表现出低复杂度的 implicit bias，或者说按照某种特定顺序来拟合子任务。

基于上述设定，总体可约损失 ΔL 可形式化为未掌握模式的加权和：

$$\Delta L \approx \sum_{k \geq 1} p_k \, q_k, \tag{2}$$

其中 q_k 为模式 k 的残差强度。在这一框架下，有限的资源（参数量、样本数或训练步数）实质上对模式空间施加了一个由优化动力学决定的低通滤波。因此，ΔL 的渐近行为主要由 Zipf 分布的尾部积分主导，从而解析地导出了幂律缩放形式。

Remark 1.3. 其实我仍旧是不满意这一点的，因为我们还是在假设里引入了一个 Zipf 幂律，那么最后导出一个幂律的 scaling law，似乎就变得很 trivial。

进一步地，我们将 Scaling Laws 的宏观形态统一为三类资源瓶颈的竞争几何。在非饱和缩放窗口内，总体损失由模型容量瓶颈 ε_N、数据统计瓶颈 ε_D 与优化收敛瓶颈 ε_τ 中的主导项决定：

$$\Delta L(N, D, \tau) \asymp \max\big(\varepsilon_N(N), \varepsilon_D(D), \varepsilon_\tau(\tau)\big). \tag{3}$$

该 max 结构不仅解释了单变量学习曲线的两段式特征，还为计算最优（Compute-Optimal）前沿提供了直接的几何推导：最优策略即为在算力约束下使各瓶颈项达到均衡。

本文的主要贡献如下：

1. **第一性原理推导：**在不引入 NTK 线性化或预设谱分布的前提下，仅基于数据的 Zipf 统计特性与梯度的排序偏好，严格推导了 N, D, C 的幂律缩放形式。
2. **得到指数的解析解：**将 Scaling 指数解析表达为数据长尾指数 α、架构效率 γ 与优化偏好 β 的函数，建立了宏观规律与微观机制的显式关联。
3. **统一了之前的一些工作：**证明了 Kaplan 与 Chinchilla 的经验规律实为同一瓶颈曲面在不同约束路径（如是否充分收敛）下的投影，并给出了计算最优前沿的解析解。

全文结构安排如下：第 2 章定义模式空间与 Zipf 假设；第 3 章形式化优化动力学的排序机制；第 4–6 章分别推导三类资源的 Scaling Law；第 7–8 章构建统一瓶颈几何并推导 Compute-Optimal 前沿的 scaling law，综上得到四种 scaling law。

2 预备：模式空间与 Zipf 假设

本节给出预备知识和前提设定，将语言建模形式化为对可数集内的模式（Modes）进行序贯学习的过程。该框架旨在抽象微观网络机制（如注意力路由、特征组合），从而在统计物理层面推导 Scaling Laws。

2.1 模式空间与重要性排序

模式定义 我们将“模式”定义为语言分布中可被统计识别并用于降低预测不确定性的结构单元，记为索引 $k \in \mathbb{N}$。模式的具体形式不限于 Token 或 n-gram，涵盖句法依赖、事实映射及长程关联等任意可重复的统计结构。该定义的物理实质在于：存在一个可数的结构集合，其对总体损失的贡献具有可加性且服从长尾分布。

重要性排序 定义 p_k 为模式 k 在数据分布中的统计权重（如出现频率或对 loss 的贡献）。不失一般性，假设模式按重要性降序排列：

$$p_1 \geq p_2 \geq \cdots \geq 0, \quad \sum_{k=1}^{\infty} p_k = 1. \tag{4}$$

我们在推导中假设模式的未掌握状态对总体损失（Cross-Entropy）的边际贡献与 p_k 成正比。

理论动机 采用模式化视角的合理性基于以下三点性质：

1. **可加性：** 总体损失在期望意义下近似为各独立模式预测误差的加权和。
2. **有序性：** 梯度下降优先拟合高频（这里的高频不是傅里叶变换语境下的高频）、高贡献的模式，使得 k 对应于学习的时间顺序或难易程度。
3. **可解耦性：** 神经网络的表示能力与泛化机制将被封装于后续定义的容量映射 $M(N)$ 与动力学方程中，从而允许我们专注于数据统计特性对 Scaling 行为的主导作用。

2.2 Zipf 统计与损失函数的谱分解

语言数据的核心统计特性在于其无尺度（Scale-free）的重尾结构。我们首先通过 Zipf 律将模式的重要性分布形式化，进而基于交叉熵的期望可加性，给出可约损失（可约损失就是损失的意思）的线性分解形式。

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：

$$p_k = \frac{1}{Z} k^{-\alpha}, \quad \alpha > 1, \tag{5}$$

其中 $Z = \zeta(\alpha)$ 为黎曼 Zeta 函数归一化常数。（用黎曼 Zeta 函数挺花哨，但其实就是个求和）

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 (5)。

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。

基于上述统计结构，我们将总体测试损失 L 分解为不可约误差 E（如 Bayes Risk）与可约损失 ΔL。在宏观尺度上，利用对数似然的期望线性性质，我们将 ΔL 建模为模式空间上的加权残差和：

$$\Delta L := L - E \approx \sum_{k=1}^{\infty} p_k q_k. \tag{6}$$

其中，$q_k \in [0,1]$ 定义为模式 k 的残差强度，表示该模式在当前模型状态下的未拟合程度。式 (6) 提供了一个从微观模式收敛到宏观损失下降的解析桥梁：训练过程被抽象为残差序列 $\{q_k\}$ 在 Zipf 权重 $\{p_k\}$ 下的逐步衰减。这一分解在理论上捕捉了“高频模式主导早期下降，长尾模式主导后期收敛”的动力学特征。

3 优化动力学的隐式谱偏好

基于前述损失分解 $\Delta L \approx \sum p_k q_k$，本节引入优化算法引入的关键动力学约束。我们假设随机梯度下降（SGD）在具有长尾分布的模式空间中并非均匀收敛，而是表现出显著的隐式谱偏好：模型倾向于优先拟合高频（并非傅里叶变换语境下的频率）、高贡献的模式。这种偏好在宏观上导致了模式学习的有序性。

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。

我们首先给出一个关于残差强度 $\{q_k\}$ 分布形态的广义假设，该假设独立于具体的优化时间或模型容量，仅描述资源受限时的稳态特征。

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），将模式空间划分为“已拟合”与“未拟合”两个主导区域。即对于 $k \ll k_\star$，有 $q_k \to 0$；对于 $k \gg k_\star$，有 $q_k \to 1$。

假设 3.1 将复杂的优化轨迹抽象为模式秩空间中的单调波前推进过程，$k_\star$ 描述在给定资源下的有效覆盖范围。为了导出关于训练时间（Compute）的具体 scaling 指数，我们仍需给出残差强度 $q_k(\tau)$ 随内禀训练时间 τ 的演化律。直观上，模式 k 的学习速度应由两类因素共同决定：其一是该模式在训练过程中被“看见”的频率，其二是每次被看见时参数更新对该模式误差的有效纠错强度。下面我们用尽可能最小的、且与具体网络细节解耦的假设，将这一乘法结构形式化，并推出后文所需的频率依赖指数衰减形式。

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 $\eta > 0$，$\lambda_k > 0$ 为该模式在当前参数化下的有效纠错系数（可理解为局部线性化下的有效曲率/可学习性）。

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，我们采用最简幂律参数化：

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

其中 $\beta = 1$ 对应“各模式在被观测到时的单次纠错强度近似相同”的基线情形；$\beta > 1$ 则刻画了优化/表征对头部模式的偏好：高频模式不仅被更频繁观测到，而且在每次观测到时也得到更强的有效纠错。

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，且取 $q_k(0) = 1$。当 $\eta\lambda_k$ 足够小、τ 足够大时，有

$$q_k(\tau) \;\approx\; \exp\big(-c\,\tau\, p_k^{\beta}\big), \qquad c := \eta\lambda_0. \tag{7}$$

等价地，在连续时间极限下满足一阶衰减动力学

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

证明. 由 假设 3.3，对固定 k 定义累计观测次数 $n_k(\tau) := \sum_{t=0}^{\tau-1} I_{t,k}$，则

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

由 假设 3.2 的大数定律，$n_k(\tau) = \tau p_k + o(\tau)$（以高概率或在期望意义下）。当 $\eta\lambda_k$ 小时，

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

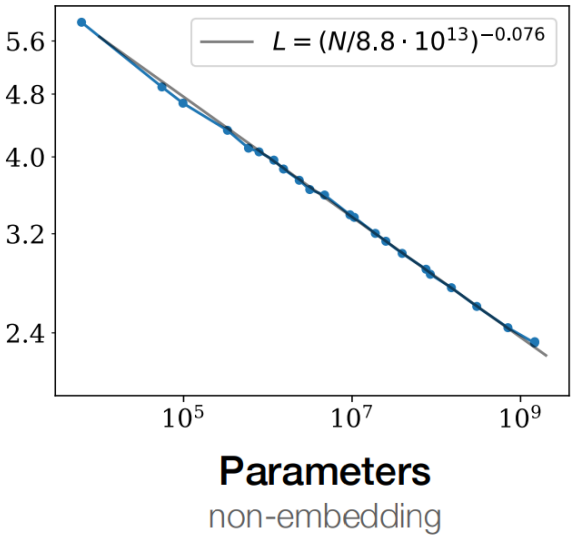
再代入 假设 3.4 的 $\lambda_k = \lambda_0 p_k^{\beta-1}$，得到 $q_k(\tau) \approx \exp(-\eta\lambda_0\,\tau\, p_k^{\beta})$，即 (7)。 □

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。

结合 假设 3.1 与 引理 3.1，我们可以将“有效覆盖前沿”的推进与 $q_k(\tau)$ 的频率依赖收敛联系起来：在给定 τ 时，满足 $c\,\tau\,p_k^{\beta} \gg 1$ 的模式残差已被显著压低，而满足 $c\,\tau\,p_k^{\beta} \ll 1$ 的尾部模式仍近似保持未拟合。由此，总体可约损失的主导项将由 Zipf 权重在该临界秩附近及其尾部的贡献控制，从而能够解析导出后续的 time/compute scaling 指数。

4 第一种 Scaling Law：模型规模 Scaling

本节希望在“数据与优化已到位”的条件下，导出模型规模 scaling law，即：



假定训练数据规模与训练时间足以使得误差主要由模型容量限制决定，从而 ΔL 的主阶只随 N 变化。核心问题是将参数量 N 诱导的有效覆盖前沿 $k_\star$ 与尾部未覆盖质量联系起来。

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

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

在很多架构与参数化下，可进一步用幂律参数化容量映射

$$M(N) \propto N^{\gamma}, \qquad \gamma > 0. \tag{8}$$

下文的结论首先以 $M(N)$ 表示；仅在额外假设 (8) 时才将其写为 N 的幂律。

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

$$\sum_{k>(1+\varepsilon)k_\star(N)} p_k \lesssim \Delta L(N) \lesssim \sum_{k>(1-\varepsilon)k_\star(N)} p_k, \tag{9}$$

特别地，

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

证明. 由 假设 3.1 的单调性与有效前沿性质，$q_k(N)$ 在 $k_\star(N)$ 两侧分别趋近于 0 与 1（在 $(1\pm\varepsilon)k_\star$ 的分离区间内），代入 $\Delta L(N) \approx \sum_k p_k q_k(N)$ 并利用 $p_k \ge 0$ 即得 (9)。令 ε 为常数并取 $N \to \infty$，得到数量级等价 $\Delta L(N) \asymp \sum_{k>k_\star(N)} p_k$，再由 $k_\star(N) \asymp M(N)$ 得结论。 □

前面的 假设 4.1 其实只是为了得到临界秩 $k_\star$ 和我们的模型规模同增同减，从而把问题转为一个尾部的估计，到这里应该可以很自然地导出一个幂律。

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

$$\sum_{k>M} p_k = \frac{1}{Z}\sum_{k>M} k^{-\alpha} \sim \frac{1}{Z(\alpha-1)} M^{-(\alpha-1)}, \qquad M \to \infty. \tag{10}$$

结合 命题 4.1 得

$$\boxed{\Delta L(N) \;\propto\; M(N)^{-(\alpha-1)}}. \tag{11}$$

若进一步满足 (8)，则

$$\boxed{\Delta L(N) \;\propto\; N^{-\gamma(\alpha-1)}}. \tag{12}$$

证明. 对 $\alpha > 1$，由积分判别法

$$\sum_{k>M} k^{-\alpha} \sim \int_M^\infty x^{-\alpha}\,dx = \frac{1}{\alpha-1} M^{-(\alpha-1)}.$$

乘以 $1/Z$ 得 (10)。再由 命题 4.1 将 $\Delta L(N)$ 归约为尾和，得到 (11)；若 $M(N) \propto N^\gamma$，代入即得 (12)。 □

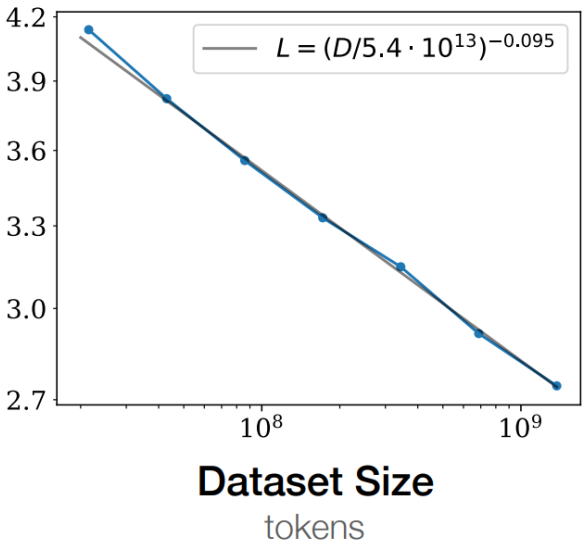
Remark 4.1 (软前沿与正则变差). 若 $q_k(N)$ 在 $k_\star(N)$ 附近为平滑过渡，只要对任意固定 $\varepsilon \in (0,1)$ 仍满足 命题 4.1 的分离极限，则 (9) 的夹逼仍成立，幂律指数不变。若将 Zipf 推广为正则变差 $p_k = k^{-\alpha}L(k)$，则 (11) 变为 $\Delta L(N) \asymp M(N)^{-(\alpha-1)}L(M(N))$，主指数仍为 $\alpha - 1$。

由 (12) 可见模型 scaling 指数为 $\alpha_N = \gamma(\alpha-1)$。因此，一方面，数据统计特性通过 α 起主导作用。数据的长尾越重（$\alpha \to 1^+$），尾部概率质量衰减越慢，导致模型 Scaling 越平缓。这意味着 Tokenizer 粒度、数据清洗策略等改变语料分布 α 的操作，将直接导致 Scaling 斜率的系统性漂移。

另一方面，模型架构效率通过 γ 施加线性调制。在数据分布不变（固定 α）的情况下，Scaling 指数 α_N 与架构效率 γ 成正比。这为架构优劣提供了一个与参数量解耦的评价指标：在相同参数规模增长下，能够解析更多有效模式（更大 γ）的架构将表现出更陡峭的 Loss 下降曲线。

5 第二种 Scaling Law：数据规模 Scaling

本节希望导出数据规模 scaling law，即：



在模式分解

$$\Delta L(D) \approx \sum_{k\geq 1} p_k \, q_k(D)$$

下，我们用“训练集中是否观测到模式 k”作为数据瓶颈的最简代理。设训练集由 D 个 token 组成，并视为从真实分布独立抽样；令 $X_k \sim \mathrm{Binomial}(D, p_k)$ 为模式 k 在训练集中的出现次数。定义

$$q_k(D) := \Pr[X_k = 0] = (1-p_k)^D. \tag{13}$$

当 $p_k \ll 1$ 且 Dp_k 中等时，有标准近似

$$(1-p_k)^D = \exp\big(D\log(1-p_k)\big) \approx e^{-Dp_k}. \tag{14}$$

因此

$$\Delta L(D) \approx \sum_{k\geq 1} p_k (1-p_k)^D \approx \sum_{k\geq 1} p_k \, e^{-Dp_k}. \tag{15}$$

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 (15) 给出。则当 $D \to \infty$ 时

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

更精确地，连续近似下有

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

证明. 由 $p_k = \frac{1}{Z}k^{-\alpha}$，用 (15) 的连续近似可写

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

作换元 $u = \frac{D}{Z}x^{-\alpha}$，则 $x = (\frac{D}{Zu})^{1/\alpha}$ 且

$$dx = -\frac{1}{\alpha}\left(\frac{D}{Z}\right)^{1/\alpha}u^{-1/\alpha-1}\,du.$$

代入得

$$\Delta L(D) \approx \frac{1}{Z}\cdot\frac{1}{\alpha}\left(\frac{D}{Z}\right)^{-(\alpha-1)/\alpha}\int_0^{D/Z} e^{-u}\,u^{-1/\alpha}\,du.$$

当 $D\to\infty$ 时，上限 $D/Z\to\infty$，且因 $\alpha>1$ 有 $1-\frac{1}{\alpha}>0$，故积分收敛到 $\Gamma(1-\frac{1}{\alpha})$，从而

$$\Delta L(D) \asymp D^{-(\alpha-1)/\alpha},$$

并得到常数 c_α 的表达式。 □

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

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

在 $p_k \ll 1$ 的区间，可用 Poisson 近似写为

$$\Pr[X_k < m_0] \approx \sum_{j=0}^{m_0-1} e^{-Dp_k}\frac{(Dp_k)^j}{j!}. \tag{17}$$

主导贡献来自满足 $Dp_k \lesssim m_0$ 的尾部模式；由 $Dp_{k_\star} \asymp m_0$ 得

$$k_\star \asymp \left(\tfrac{D}{Zm_0}\right)^{1/\alpha}, \qquad \sum_{k>k_\star} p_k \asymp k_\star^{-(\alpha-1)} \asymp D^{-(\alpha-1)/\alpha},$$

因此幂律指数仍为 $(\alpha-1)/\alpha$，仅常数发生变化。

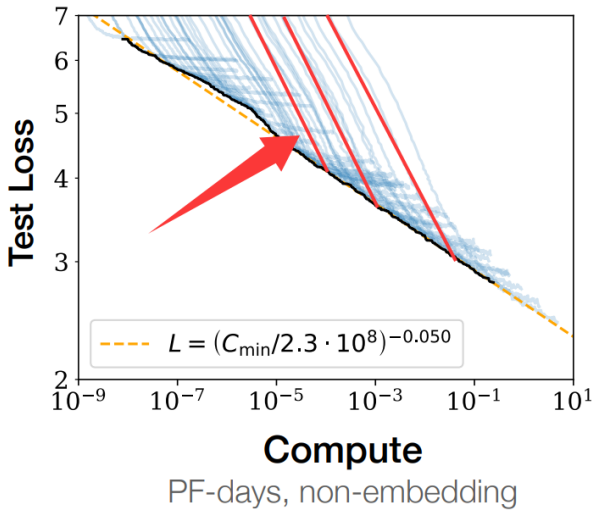
Remark 5.2 (两点直接推论). 由 定理 5.1，数据 scaling 指数为

$$\alpha_D = \frac{\alpha-1}{\alpha}.$$

其一，$\alpha\to 1^+$ 时 $\alpha_D\to 0$，即在该 surrogate 下 $\Delta L(D)$ 随 D 的下降可以任意缓慢。其二，若数据处理（如 tokenizer、去重、过滤、采样/合成策略）改变了有效的尾指数 α，则观测到的 α_D 将随之系统性变化；在同一近似窗口内，α 增大对应 α_D 增大。

6 第三种 Scaling Law：训练时间/算力 Scaling

本节希望在模型容量与数据覆盖均不构成主导瓶颈的条件下，导出训练时间/算力 scaling law，即图中平行的红线：



沿用模式分解

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

其中 τ 为内禀训练时间（例如优化步数或等价的 token-steps），$q_k(\tau) \in [0,1]$ 为模式 k 在时间 τ 的残差强度。

沿用 引理 3.1，我们有

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

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 (7) 成立。则当 $\tau \to \infty$，

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

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. 由 (18) 与 $p_k = \frac{1}{Z} k^{-\alpha}$，作连续近似

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

记 $a := c\left(\frac{1}{Z}\right)^{\beta}$，换元 $u = a\tau x^{-\alpha\beta}$，则 $x = \left(\frac{a\tau}{u}\right)^{1/(\alpha\beta)}$ 且

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

代入得

$$\Delta L(\tau) \approx \frac{1}{Z} \cdot \frac{1}{\alpha\beta}\, (a\tau)^{-(\alpha-1)/(\alpha\beta)} \int_0^{a\tau} e^{-u}\, u^{(\alpha-1)/(\alpha\beta)-1}\, du.$$

当 $\tau \to \infty$ 时，上限 $a\tau \to \infty$，积分收敛到 $\Gamma\big(\frac{\alpha-1}{\alpha\beta}\big)$，从而得到 (19) 及常数表达式。 □

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 导出总体 loss 和两段式结构

前文分别得到在单一资源受限时的幂律衰减。实际预训练中 N, D, τ 同时有限，因此需要一个组合律来刻画 $\Delta L(N, D, \tau)$，并解释经验拟合中常见的“幂律加和”形式。本节给出一个结构性结论：在非饱和的 scaling window 内，总体可约损失与三项瓶颈的 max 同阶（说白了就是被最差的那一项控制），而加和只是其平滑替代。

我们以三个单变量瓶颈项表征三类资源在该窗口内的主导阶

$$\varepsilon_N(N) = A\,N^{-\alpha_N}, \qquad \varepsilon_D(D) = B\,D^{-\alpha_D}, \qquad \varepsilon_\tau(\tau) = G\,\tau^{-\alpha_\tau}, \tag{20}$$

其中 $A, B, G > 0$ 为与实现细节相关的常数，$\alpha_N, \alpha_D, \alpha_\tau > 0$ 为 scaling 指数。在本文的 Zipf–模式框架中，$\alpha_N = \gamma(\alpha-1)$、$\alpha_D = \frac{\alpha-1}{\alpha}$、$\alpha_\tau = \frac{\alpha-1}{\alpha\beta}$，但以下结构推导不依赖其具体形式。

命题 7.1 (loss 的 Max 结构). 在 scaling window 内，若存在与常数无关的比较记号使得

$$\Delta L(N, D, \tau) \;\gtrsim\; \varepsilon_N(N), \qquad \Delta L(N, D, \tau) \;\gtrsim\; \varepsilon_D(D), \qquad \Delta L(N, D, \tau) \;\gtrsim\; \varepsilon_\tau(\tau), \tag{21}$$

并且存在一类训练策略满足联合上界

$$\Delta L(N, D, \tau) \;\lesssim\; \varepsilon_N(N) + \varepsilon_D(D) + \varepsilon_\tau(\tau), \tag{22}$$

则有数量级等价

$$\boxed{\Delta L(N, D, \tau) \;\asymp\; \max\big(\varepsilon_N(N),\, \varepsilon_D(D),\, \varepsilon_\tau(\tau)\big).} \tag{23}$$

证明. 由 (21) 得 $\Delta L \gtrsim \max(\varepsilon_N, \varepsilon_D, \varepsilon_\tau)$。另一方面，对任意非负 x, y, z 有范数等价

$$\max(x, y, z) \;\le\; x + y + z \;\le\; 3\max(x, y, z), \tag{24}$$

将 $(x, y, z) = (\varepsilon_N, \varepsilon_D, \varepsilon_\tau)$ 代入并结合 (22) 得 $\Delta L \lesssim \max(\varepsilon_N, \varepsilon_D, \varepsilon_\tau)$。两侧合并即得 (23)。 □

(23) 给出“主导瓶颈决定误差阶”的结论。工程拟合中常采用加和形式

$$\widetilde{\Delta L}(N, D, \tau) := \varepsilon_N(N) + \varepsilon_D(D) + \varepsilon_\tau(\tau), \tag{25}$$

其与 Max 在数量级上等价由 (24) 直接保证；加和的作用是提供瓶颈切换处的连续表达，便于回归拟合。更一般地，可用

$$M_p(x, y, z) := \big(x^p + y^p + z^p\big)^{1/p}, \qquad p \in [1, \infty)$$

定义一族平滑聚合算子；对固定 p，M_p 与 max 仍然范数等价，因此不会改变主导幂律指数。例如在论文 Scaling Laws for Neural Language Models [1] 中给出的：

$$L(N,D)=\left[\left(\frac{N_c}{N}\right)^{\frac{\alpha_N}{\alpha_D}}+\frac{D_c}{D}\right]^{\alpha_D}$$

下面我们尝试导出论文 Scaling Laws for Neural Language Models [1] 中的两段式 loss 形式。固定一次训练 run 的 (N,D)，将静态瓶颈记为

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

由 (23) 得到随训练时间变化的统一表达

$$\Delta L(N,D,\tau)\ \asymp\ \max\big(\varepsilon_{\text{stat}}(N,D),\,\varepsilon_\tau(\tau)\big). \tag{26}$$

定义切换点 $\tau_\star$ 为两项同阶时的时间尺度，即

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

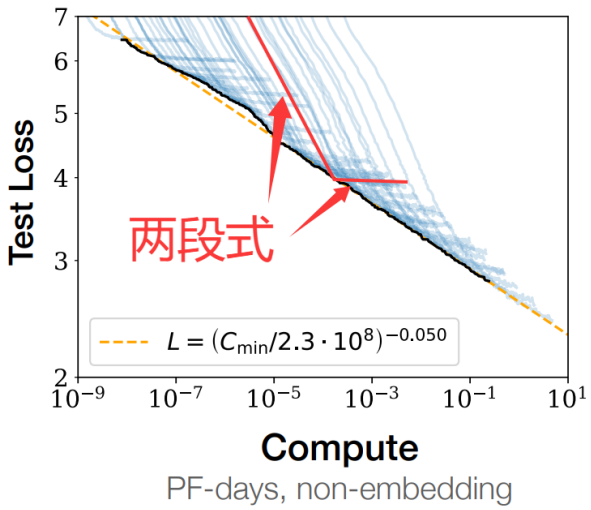
于是当 $\tau\ll\tau_\star$ 时，$\varepsilon_\tau(\tau)\gg\varepsilon_{\text{stat}}$，由 (26) 得

$$\Delta L(N,D,\tau)\asymp\varepsilon_\tau(\tau)\propto\tau^{-\alpha_\tau},$$

而当 $\tau\gg\tau_\star$ 时，$\varepsilon_\tau(\tau)\ll\varepsilon_{\text{stat}}$，继续训练不改变数量级，

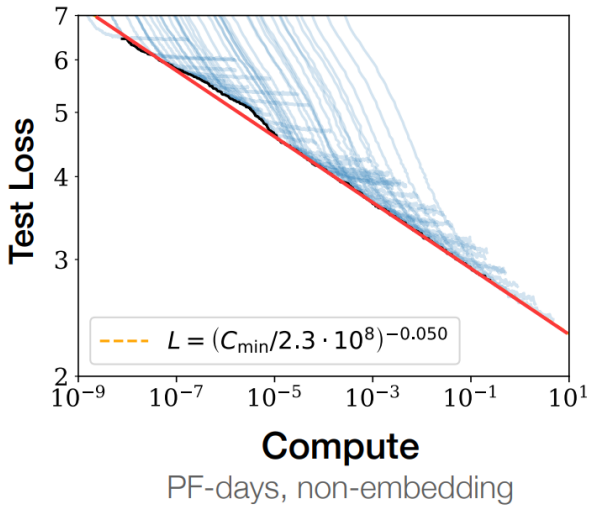
$$\Delta L(N,D,\tau)\asymp\varepsilon_{\text{stat}}(N,D)=\max\big(\varepsilon_N(N),\varepsilon_D(D)\big).$$

这给出固定 (N,D) 时学习曲线的两段式结构，并将平台高度与切换时间用 (20) 的幂律显式参数化。



8 第四种 Scaling Law：Optimal compute point Scaling

本节我们希望导出 optimal compute point 的 scaling law，即红线：



在 Section 7 的两段式结构

$$\Delta L(N, D, \tau) \asymp \max\left(\varepsilon_{\text{stat}}(N, D), \varepsilon_\tau(\tau)\right), \qquad \varepsilon_{\text{stat}}(N, D) := \max\left(\varepsilon_N(N), \varepsilon_D(D)\right)$$

下，我们将下降段与平台段的转折点定义为使两项同阶的训练时间尺度。

定义 8.1 (Optimal compute point / 转折点). 给定一次训练运行的 (N, D)，定义其 *optimal compute point*（亦即两段式转折点）为任意满足

$$\varepsilon_\tau(\tau_\star) = \varepsilon_{\text{stat}}(N, D) \tag{27}$$

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。

命题 8.1 (Optimal compute point 的显式表达). 在 (20) 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 8.1 中的 $\tau_\star, \Delta L_\star, C_\star$ 满足

$$\tau_\star(N, D) = \left(\frac{G}{\varepsilon_{\text{stat}}(N, D)}\right)^{1/\alpha_\tau} = \left(\frac{G}{\max(AN^{-\alpha_N}, BD^{-\alpha_D})}\right)^{1/\alpha_\tau}, \tag{28}$$

$$\Delta L_\star(N, D) \asymp \varepsilon_{\text{stat}}(N, D) = \max(AN^{-\alpha_N}, BD^{-\alpha_D}), \tag{29}$$

$$C_\star(N, D) \asymp N\left(\frac{G}{\max(AN^{-\alpha_N}, BD^{-\alpha_D})}\right)^{1/\alpha_\tau}. \tag{30}$$

特别地，在两种静态瓶颈主导区域内，

$$C_\star(N, D) \asymp \begin{cases} N^{1+\alpha_N/\alpha_\tau}, & AN^{-\alpha_N} \gtrsim BD^{-\alpha_D}, \\ N\, D^{\alpha_D/\alpha_\tau}, & BD^{-\alpha_D} \gtrsim AN^{-\alpha_N}. \end{cases} \tag{31}$$

证明. 由 (27) 与 $\varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau}$ 直接解得 (28)。将 $\tau = \tau_\star$ 代回两段式结构的定义即得 (29)。再由算力定义 $C_\star = \kappa' N \tau_\star$ 得 (30)，并由 $\max(\varepsilon_N, \varepsilon_D)$ 的两种主导情形得到 (31)。 □

下面给出一条将 optimal compute point 沿算力连线写成幂律的结论。它对应于在静态瓶颈均衡族上观察转折点轨迹。

假设 8.1 (静态瓶颈均衡族). 考虑一族训练运行 $\{(N, D)\}$ 满足

$$AN^{-\alpha_N} \asymp BD^{-\alpha_D}. \tag{32}$$

定理 8.1 (Optimal compute point 的 loss–compute 幂律). 在 假设 8.1 下，optimal compute point 的损失高度与算力满足

$$\boxed{\Delta L_\star(C) \propto C^{-\frac{\alpha_N \alpha_\tau}{\alpha_N + \alpha_\tau}}.} \tag{33}$$

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. 在 假设 8.1 下有 $\varepsilon_{\text{stat}}(N, D) \asymp AN^{-\alpha_N}$。由 命题 8.1 得

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

并且

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

消去 N 即得 $\Delta L_\star \asymp C_\star^{-\alpha_N/(1+\alpha_N/\alpha_\tau)} = C_\star^{-\alpha_N\alpha_\tau/(\alpha_N+\alpha_\tau)}$，从而得到 (33)。 □

Remark 8.1 (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

$$L(N, S) = \left(\frac{N_c}{N}\right)^{\alpha_N} + \left(\frac{S_c}{S}\right)^{\alpha_S}, \tag{34}$$

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

$$L_{\text{opt}}(C) \propto C^{-\frac{\alpha_N \alpha_S}{\alpha_N + \alpha_S}}. \tag{35}$$

将 定理 8.1 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 (33) 与 (35) 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。

9 Compute-Optimal：预算约束下的均衡条件

本节在瓶颈模型

$$\begin{aligned} \Delta L(N,D,\tau) &\asymp \max\Big(\varepsilon_N(N),\, \varepsilon_D(D),\, \varepsilon_\tau(\tau)\Big), \\ \varepsilon_N(N) = AN^{-\alpha_N}, &\quad \varepsilon_D(D) = BD^{-\alpha_D}, \quad \varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau} \end{aligned} \tag{36}$$

下推导给定算力预算 C 时的 compute-optimal 前沿。这里 $A, B, G > 0$ 与实现细节有关，$\alpha_N, \alpha_D, \alpha_\tau > 0$ 为三类单变量 scaling 指数。我们将 compute-optimal 前沿定义为约束优化问题的最小值轨迹：在给定 C 的可行集合内，选择 (N, D, τ) 以最小化 $\Delta L(N, D, \tau)$，并记对应最优值为 $\Delta L_{\text{opt}}(C)$，最优解（不必唯一）为 $(N_{\text{opt}}(C), D_{\text{opt}}(C), \tau_{\text{opt}}(C))$。

经验上，Kaplan 等 [1] 与 Hoffmann 等（Chinchilla）[2] 给出了不同的最优分配幂律（例如 $N \propto C^{0.73}$ 与 $N \propto C^{0.5}$）。下面给出一个解析解释：在同一 Max-bottleneck 结构 (36) 下，最优分配取决于哪个瓶颈在可行域内被消去为次主导项；Kaplan 与 Chinchilla 对应于同一高维损失曲面在不同主导瓶颈切片上的最优轨迹。

先考虑 τ 足够大以致收敛瓶颈次主导的情形。此时 ΔL 由 (N, D) 的静态瓶颈控制，而算力主要由参数规模与训练 token 数决定。对稠密 Transformer 的常用近似为

$$C = \kappa ND, \tag{37}$$

其中 $\kappa > 0$ 为实现常数。

命题 9.1 (Chinchilla 切片：ε_τ 次主导时的最优分配 [2]). 假设训练时间足够大使得

$$\varepsilon_\tau(\tau) \lesssim \max\big(\varepsilon_N(N),\, \varepsilon_D(D)\big), \tag{38}$$

从而

$$\Delta L(N, D) \asymp \max\big(AN^{-\alpha_N},\, BD^{-\alpha_D}\big). \tag{39}$$

在预算约束 $C = \kappa ND$ 下，约束最优化

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

的任一最优解都满足均衡条件

$$AN^{-\alpha_N} \asymp BD^{-\alpha_D}, \tag{40}$$

并且存在与 C 无关的常数使得

$$N_{\text{opt}}(C) \propto C^{\frac{\alpha_D}{\alpha_N+\alpha_D}}, \qquad D_{\text{opt}}(C) \propto C^{\frac{\alpha_N}{\alpha_N+\alpha_D}}, \qquad \Delta L_{\text{opt}}(C) \propto C^{-\frac{\alpha_N\alpha_D}{\alpha_N+\alpha_D}}. \tag{41}$$

证明. 在约束 $ND = C/\kappa$ 下写 $D = (C/\kappa)N^{-1}$，目标函数等价为

$$\phi(N) = \max\Big(AN^{-\alpha_N},\; B\kappa^{\alpha_D}C^{-\alpha_D}N^{\alpha_D}\Big).$$

第一项随 N 单调下降，第二项随 N 单调上升，因此极小值发生在两项同阶处，从而得到 (40)；代入解出 $N_{\text{opt}}(C), D_{\text{opt}}(C)$ 的幂律指数，并由最优点处两项同阶得到 $\Delta L_{\text{opt}}(C)$ 的指数。 □

再考虑数据瓶颈次主导的情形，即 D 足够大而训练步数（或等价内禀时间）τ 受限；这对应于 Kaplan 等 [1] 以步数/算力约束推导的最优分配。此时常用预算近似为

$$C = \kappa' N \tau, \tag{42}$$

其中 $\kappa' > 0$ 为实现常数。

命题 9.2 (Kaplan 切片：ε_D 次主导时的最优分配 [1]). 假设数据瓶颈次主导，即

$$\varepsilon_D(D) \lesssim \max\big(\varepsilon_N(N), \varepsilon_\tau(\tau)\big), \tag{43}$$

从而

$$\Delta L(N, \tau) \asymp \max\big(AN^{-\alpha_N}, G\tau^{-\alpha_\tau}\big). \tag{44}$$

在预算约束 $C = \kappa' N\tau$ 下，约束最优化

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

的任一最优解都满足均衡条件

$$AN^{-\alpha_N} \asymp G\tau^{-\alpha_\tau}, \tag{45}$$

并且存在与 C 无关的常数使得

$$N_{\text{opt}}(C) \propto C^{\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}}, \qquad \tau_{\text{opt}}(C) \propto C^{\frac{\alpha_N}{\alpha_N+\alpha_\tau}}, \qquad \Delta L_{\text{opt}}(C) \propto C^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}}. \tag{46}$$

证明. 同 命题 9.1。在约束 $N\tau = C/\kappa'$ 下写 $\tau = (C/\kappa')N^{-1}$，目标函数为两项一降一升的 max，极小点由两项同阶给出 (45)，进而解得 (46)。 □

由 命题 9.1 and 9.2 可见，Kaplan 与 Chinchilla 的最优分配并非相互矛盾：它们分别对应于同一 Max-bottleneck 损失曲面 (36) 在不同 “次主导假设” 下的约束最优轨迹。更具体地说，当可行域允许将 ε_τ 压到不超过静态瓶颈(38)时，最优点由 ε_N 与 ε_D 的均衡 (40) 决定；当可行域允许将 ε_D 压到不超过 $(\varepsilon_N, \varepsilon_\tau)$(43)时，最优点由 ε_N 与 ε_τ 的均衡 (45) 决定。两者共享同一第一性结构：在单调幂律瓶颈与乘性预算约束下，max 型目标的最优解发生在 “活跃瓶颈同阶” 的边界上。

Remark 9.1. “活跃瓶颈同阶” 是 ai 给的一个词，我想不到一个比较好的词。

最后，将本文 Zipf–模式框架得到的指数

$$\alpha_N = \gamma(\alpha - 1), \qquad \alpha_D = \frac{\alpha - 1}{\alpha}, \qquad \alpha_\tau = \frac{\alpha - 1}{\alpha\beta} \tag{47}$$

代入 (41) 与 (46)，即可得到 compute-optimal 前沿在机制参数 (α, β, γ) 下的显式幂律。特别地，在 Chinchilla 切片（(38)）下，

$$N_{\text{opt}}(C) \propto C^{\frac{1}{1+\alpha\gamma}}, \qquad D_{\text{opt}}(C) \propto C^{\frac{\alpha\gamma}{1+\alpha\gamma}}, \qquad \Delta L_{\text{opt}}(C) \propto C^{-\frac{\gamma(\alpha-1)}{1+\alpha\gamma}}, \tag{48}$$

而在 Kaplan 切片（(43)）下，

$$N_{\text{opt}}(C) \propto C^{\frac{1}{1+\alpha\beta\gamma}}, \qquad \tau_{\text{opt}}(C) \propto C^{\frac{\alpha\beta\gamma}{1+\alpha\beta\gamma}}, \qquad \Delta L_{\text{opt}}(C) \propto C^{-\frac{\gamma(\alpha-1)}{1+\alpha\beta\gamma}}. \tag{49}$$

因此，compute-optimal 前沿不仅给出 $\Delta L_{\text{opt}}(C)$ 的全局幂律指数，也给出最优资源配置 (N, D, τ) 的幂律指数；这些指数由 (47) 将宏观规律与可测的统计/动力学参数直接关联，从而为后续的实证检验提供了明确的可比较量。

10 结论

本文在模式分解

$$\Delta L \approx \sum_{k\geq 1} p_k q_k \tag{50}$$

与 Zipf 统计假设（假设 2.1：$p_k = \frac{1}{Z} k^{-\alpha}$，$\alpha > 1$）的基础上，引入“有序学习/有效前沿”（假设 3.1）与频率依赖的收敛动力学（引理 3.1），在不使用 NTK 线性化或预设核谱幂律的前提下，给出了四类常见 scaling law 的解析推导。核心机制是：有限资源在模式秩空间上诱导一个随资源推进的覆盖前沿 $k_\star$，而 ΔL 的主阶由 Zipf 尾部质量控制。

在本文框架下，三类单变量瓶颈可写为

$$\varepsilon_N(N) = AN^{-\alpha_N}, \qquad \varepsilon_D(D) = BD^{-\alpha_D}, \qquad \varepsilon_\tau(\tau) = G\tau^{-\alpha_\tau}, \tag{51}$$

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

$$\alpha_N = \gamma(\alpha - 1), \qquad \alpha_D = \frac{\alpha - 1}{\alpha}, \qquad \alpha_\tau = \frac{\alpha - 1}{\alpha\beta}. \tag{52}$$

对应的四类幂律结论可概括为（忽略与实现相关的常数）：

$$\text{(i) Model scaling:} \quad \Delta L(N) \asymp M(N)^{-(\alpha-1)} \sim N^{-\gamma(\alpha-1)}, \tag{53}$$

$$\text{(ii) Data scaling:} \quad \Delta L(D) \asymp D^{-(\alpha-1)/\alpha}, \tag{54}$$

$$\text{(iii) Time/compute scaling:} \quad \Delta L(\tau) \asymp \tau^{-(\alpha-1)/(\alpha\beta)}, \tag{55}$$

$$\text{(iv) Optimal point / frontier:} \quad \Delta L_\star(C) \propto C^{-\frac{\alpha_N \alpha_\tau}{\alpha_N + \alpha_\tau}} \quad \text{(Kaplan 型预算).} \tag{56}$$

在联合资源有限的 scaling window 内，总体可约损失由瓶颈的主导项决定（命题 7.1）：

$$\Delta L(N, D, \tau) \asymp \max\big(\varepsilon_N(N), \varepsilon_D(D), \varepsilon_\tau(\tau)\big), \tag{57}$$

从而 compute-optimal 的分配与前沿是一个确定的约束极小化问题：在乘性预算（如 $C \asymp ND$ 或 $C \asymp N\tau$）下，最优点满足“活跃瓶颈同阶”的均衡条件，并给出

$$\begin{aligned}
&\text{(Chinchilla 切片：}\varepsilon_\tau \text{ 次主导，}C \asymp ND\text{)：} \quad N_{\text{opt}}(C) \propto C^{\frac{\alpha_D}{\alpha_N+\alpha_D}},\ D_{\text{opt}}(C) \propto C^{\frac{\alpha_N}{\alpha_N+\alpha_D}},\ \Delta L_{\text{opt}}(C) \propto C^{-\frac{\alpha_N\alpha_D}{\alpha_N+\alpha_D}},\\
&\text{(Kaplan 切片：}\varepsilon_D \text{ 次主导，}C \asymp N\tau\text{)：} \quad N_{\text{opt}}(C) \propto C^{\frac{\alpha_\tau}{\alpha_N+\alpha_\tau}},\ \tau_{\text{opt}}(C) \propto C^{\frac{\alpha_N}{\alpha_N+\alpha_\tau}},\ \Delta L_{\text{opt}}(C) \propto C^{-\frac{\alpha_N\alpha_\tau}{\alpha_N+\alpha_\tau}}.
\end{aligned} \tag{58}$$

将 (52) 代入即可得到以 (α, β, γ) 参数化的最优分配与前沿指数（见 Section 9）。

上述结论给出了可检验的指数约束关系。设经验上拟合得到的单变量指数为 $(\hat{\alpha}_N, \hat{\alpha}_D, \hat{\alpha}_\tau)$，则该框架要求存在 (α, β, γ) 使得

$$\hat{\alpha}_D = \frac{\alpha - 1}{\alpha}, \qquad \hat{\alpha}_\tau = \frac{1}{\beta}\hat{\alpha}_D, \qquad \hat{\alpha}_N = \gamma(\alpha - 1), \tag{59}$$

并进一步要求 compute-optimal 前沿指数满足

$$\hat{\alpha}_{\text{opt}} = \frac{\hat{\alpha}_N \hat{\alpha}_D}{\hat{\alpha}_N + \hat{\alpha}_D} \ \text{(Chinchilla 型)} \qquad \text{或} \qquad \hat{\alpha}_{\text{opt}} = \frac{\hat{\alpha}_N \hat{\alpha}_\tau}{\hat{\alpha}_N + \hat{\alpha}_\tau} \ \text{(Kaplan 型).} \tag{60}$$

因此，本文框架不仅给出幂律形式，还给出指数之间的可证伪一致性条件。

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.