学习收缩困难尾部：可解线性模型中的依赖训练的推理扩展

Noam LeviÉ

洛桑联邦理工学院（EPFL），瑞士洛桑
noam.levi@epfl.ch

arXiv:2601.03764v1 [cs.LG] 7 Jan 2026

摘要

我们在一个可解的最后一层微调模型中分析神经网络缩放定律，其中目标具有内在的实例异质性难度。在我们的潜在实例难度（LID）模型中，每个输入的目标方差由从重尾分布中抽取的潜在“精度”决定。虽然泛化损失遵循标准缩放定律，但我们的主要贡献是将其与推理联系起来。pass@k 失败率呈幂律衰减，$k^{-\beta_{\text{eff}}}$，但观测到的指数 β_{eff} 依赖于训练。它会随样本量 N 增加，直至达到由难度分布尾部决定的内在极限 β 后趋于饱和。这种耦合揭示，学习会收缩误差分布的“困难尾部”：随着模型的泛化误差改善，pass@k 曲线会变陡，直到不可约目标方差占据主导地位。LID 模型为这种行为提供了可检验的闭式预测，包括一条算力分配规则：在饱和前优先训练，之后则优先增加推理尝试。我们通过模拟以及两个真实数据代理验证了这些预测：CIFAR-10H（人工标注方差）和一项数学教师—学生蒸馏任务。

1 引言

大规模机器学习模型的非凡成功，与经验观察和理论理解的缩放定律密切相关；这些定律刻画了随着数据、模型规模或训练时间增加而带来的性能提升（Kaplan 等，2020；Hestness 等，2017；Rosenfeld 等，2019；Bahri 等，2024；Maloney 等，2022；Bordelon 等，2024）。这些定律对于预测学习曲线以及在数据集规模 N和参数数量 P较大的情形下优化资源分配至关重要。大多数经典分析聚焦于训练 方面，通过数据分布的特性（如谱衰减）（Bartlett 等，2020；Hastie 等，2020；Bordelon 等，2020）或易于处理的模型类别（Maloney 等，2022；Tay 等，2022），建立泛化损失 $\mathcal{L}_{\text{gen}}$ 与 N（有时也包括 P）之间的关系。

一种互补范式强调推理时计算：重复尝试、best-of-N以及搜索，即使无需进一步训练，也能在困难推理任务上取得显著提升，通常在验证器下通过 pass@k 进行评估（Snell 等，2024；Brown 等，2024）。尽管利用推理时计算的实用方法正在快速发展（见第 2 节），并且已有一些解释说明其成功原因（Levi，2024；Schaeffer 等，2025a），但一个基本的理论问题仍然存在：训练进展如何塑造推理时的性能扩展？

在许多现实场景中，输入—输出关系并非确定性的；有些实例在本质上比其他实例具有更大的方差（Arpit 等人，2017年；Northcutt 等，2021）。这种实例层面的异质性会影响两个阶段：在训练期间，我们针对每个输入只能观察到一次带噪声的实现；而在推理期间，我们可能会将模型预测与 pass@k下的多个新实现进行比较。这促使我们构建一个通过实例难度明确联系训练与推理的最简模型。

设定。 我们采用一种刻意简化但可分析的框架，以反映微调中的常见做法：在高维固定特征上进行最后一层（线性）回归，且目标本质上是随机的。每个实例$\mathbf{x}$ 都带有一个潜在精度 $\tau_{\mathbf{x}}$（即其“难度”），从重尾分布中抽取，它决定了目标围绕均值$\mathbf{x}^\top \boldsymbol{\theta}^*$的方差。训练

阶段对每个输入观察一次实现值 $y \sim Y^*_{\mathbf{x}}$，并拟合一个线性头（岭回归/普通最小二乘）。在推理时，我们使用完美验证器评估 pass@k：对每个测试输入抽取 k 个新的实现值，并判断其中是否至少有一个落在模型预测值的固定容差范围内。

概述与贡献。

- **用于线性微调的可解 LID 模型**。我们在高维线性设定下形式化了潜在实例难度（LID）模型。对于每个输入仅使用一次实现进行训练，等价于岭回归/普通最小二乘法，并恢复了关于样本量 N、维度 d 和谱指数 α 的既有泛化标度关系（包括经典情形下的 $1/N$ 尾部，此时平均目标方差有限，即 $\beta > 2$）。
- **通过双尾定律实现训练-推理耦合**。我们表明，在 pass@k 下，单次试验成功概率的分布会出现两个正则变差分量：一个内在尾部由潜在难度分布决定（指数为 β），另一个有限N 尾部则由模型相对于均值目标的误差支配（指数为 γ（N）$\propto 1/\mathcal{L}_{\text{gen}}$（$N$））。对试验取平均后，可得到混合 pass@$k$ 定律，由此产生有效推理指数 β_{eff}（N）$=\min\{\beta, \gamma(N)\}$。因此，观测到的 pass@$k$ 斜率随 N 增大并在内在难度指数 β 处趋于饱和。

- **预测与启示**。理论预测：(i) 在 (N, k) 空间中存在一条交叉曲面，将有限-N（偏差主导）区域与内在尾部区域分隔开；(ii) $\beta_{\text{eff}}(N)$ 曲线会趋于饱和；(iii) 即使斜率已进入平台期，N 仍能持续改善前因子。这些结论导出一条简单的算力分配规则：持续投入训练，直至 $\beta_{\text{eff}}(N)$ 接近 β，随后优先增加推理尝试。
- **模拟与真实数据代理的证据**。受控模拟证实了 $1/N$训练尾部（截距也正确）、pass@k 随 N 增大而变陡，以及 $\hat{}\beta_{\text{eff}}(N)$ 趋于饱和。两个真实数据代理表现出类似现象：CIFAR-10H（人工标注方差），以及使用 GSM8K 的数学推理教师–学生蒸馏场景。

综合来看，这些结果为一种广泛使用的微调范式提供了清晰、可检验的基线，统一了训练缩放与推理缩放。该模型明确说明了测试时计算何时能够发挥作用、其收益能达到多大程度，以及这些收益如何取决于训练。在 LID 线性微调场景中，标准的 $1/N$ 型泛化缩放与 N 并存，同时实例难度呈重尾分布。随着 N 增长，模型的平均平方偏差缩小，误差质量集中到一组不断缩小的内在难例上。在推理方面，这表现为 pass@k 失败曲线 $\mathcal{L}_{\text{inf}}(k; N) \sim \tilde{P}(N)\ k^{-\beta_{\text{eff}}(N)}$，其斜率 $\beta_{\text{eff}}(N)$ 随 N 增大，并最终趋于平台期，达到内在难度指数 β。换言之，训练逐步“缩小”误差分布的难例尾部，直至不可约实例随机性占据主导，测试时缩放指数趋于饱和。

2 相关工作

Here，我们重点回顾与本文贡献最为密切的既有研究。关于 n addi更全面细致的文献综述，请读者参阅附录 A 及其中的参考文献。

泛化缩放定律。大量研究已证实，深度网络的泛化损失通常会随着数据集规模 N 或模型参数等资源的变化，按照可预测的幂律缩放（Hestness 等，2017；Kaplan 等，2020；Hoffmann and et al., 2022）。理论框架试图依据数据的性质（例如其谱衰减）或模型架构来解释这些定律（Bahri 等人，2021；Maloney 等，2022）。我们的工作建立在泛化损失随 N变化的标准缩放关系之上，该关系构成了统一模型中的“训练”部分。

推理时缩放。另一条研究路线表明，即使不进一步训练，只要增加推理时的计算量，困难推理任务的表现就能得到显著提升（Snell 等，2024；Brown 等，2024）。主流方法是生成多个候选解并选出其中最佳者，通常使用 pass@k 指标进行评估。虽然已有研究提出了一些解释这一现象的理论模型（Levi，2024；Schaeffer 等，2025a），但这些模型通常将推理单独分析。

据我们所知，目前仍缺少一种简单且可解析求解的模型，能够将训练进展（即泛化误差的下降）与推理性能的缩放联系起来。我们的工作旨在提供这样一种统一的视角。

本文其余部分的组织如下：在第3节中，我们介绍 LID 模型。我们在第 4 节给出训练与推理缩放定律的主要结果。在第5节中，我们展示在一个本质上具有随机标注的数据集上进行线性微调的两个示例，表明 LID 是某些现实世界任务的合理代理指标。最后，我们在第6节总结全文。

3 潜在实例难度设定

现实世界的数据集呈现出显著的实例级异质性：一些图像标签含义模糊，因此标注者分歧更大（Peterson 等，2019；Northcutt 等，2021）；一些推理问题在本质上比其他问题更难，导致输出更加多变。标准的同质噪声假设忽视了这种复杂性。为应对这一问题，并将其与训练和推理过程中观察到的不同缩放行为联系起来（尤其是在可以根据正确解验证多次推理尝试时；参见第1节），我们引入了这一设定。直观而言，增加训练难度的因素（更难学习均值）也会影响推理的可靠性（更难匹配新的实现结果）。

最后一层微调视角。 我们在常见的微调范式中实例化潜在实例难度(LID)：冻结表征生成特征$\mathbf{x} \in \mathbb{R}^d$，我们据此学习一个线性头$\mathbf{x}^\top\boldsymbol{\theta}$。每个实例都具有一个潜在的精度 （即其“容易程度”） $\tau_{\mathbf{x}}$，它控制目标围绕均值$\mathbf{x}^\top\boldsymbol{\theta}^*$的方差。训练时，每个$\mathbf{x}$只观测到一个 实现值 y；推理时，pass@k 会将模型的预测与来自同一实例特定目标分布的 k 个新实现值进行比较。

定义 3.1 （潜在实例难度（LID）模型）。 观测值（$\mathbf{x}$ y）的数据生成过程如下：

1. **特征**。 输入特征向量$\mathbf{x} \in \mathbb{R}^d$ 从分布 p（$\mathbf{x}$）中抽取，其均值为零$\mathbb{E}[\mathbf{x}] = 0$，协方差为 $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$。我们假设特征值 σ_j^2 按幂律衰减$\sigma_j^2 \propto j^{-(1+\alpha)}$，其中 $j = 1, \ldots\ldots, d$，且 $\alpha > 0$。

2. **潜在难度**。与$\mathbf{x}$ 相关联的是一个潜在的难度精度 $\tau_{\mathbf{x}} \in (0, \infty)$，它独立于$\mathbf{x}$抽取，服从

$$\tau_{\mathbf{x}} \sim \mathrm{Gamma}(\text{shape} = \beta/2,\ \text{rate} = 1), \tag{1}$$

其中，$\beta > 0$ 控制近零尾部。β 越小，极低精度（高内在方差）的概率质量越大，对应“困难”实例。

3. **随机目标**。给定（$\mathbf{x}$ $\tau_{\mathbf{x}}$），实例目标服从高斯分布，其均值为均值关系 $f^*(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\theta}^*$，方差与 $\tau_{\mathbf{x}}$ 成反比：

$$Y_{\mathbf{x}}^* \sim \mathcal{N}\big(\text{mean} = \mathbf{x}^T\boldsymbol{\theta}^*,\ \text{variance} = \sigma_\eta^2/\tau_{\mathbf{x}}\big), \quad \text{where } \sigma_\eta^2 > 0 \text{ is a global scale.} \tag{2}$$

4. **训练标签**。观测到的标签是一次单独的实现

$$y \sim Y_{\mathbf{x}}^* \quad \text{equivalently} \quad y = \mathbf{x}^T\boldsymbol{\theta}^* + \eta,\ \eta \sim \mathcal{N}(0,\ \sigma_\eta^2/\tau_{\mathbf{x}}). \tag{3}$$

注释。(i) 条目 1 中的幂律谱概括了实践中观察到、并用于线性缩放分析的良性特征区域（Maloney 等，2022；Levi 和 Oz，2023；2024）。(ii) 式 (1) 中选用伽马分布族是出于解析便利性；所有决定推理缩放的结果都只取决于近零尾部 $\Pr(\tau_{\mathbf{x}} \le t) \asymp t^{\beta/2}$，因此，具有相同尾指数的其他分布会得到相同的指数（Levi，2024）。(iii) 假设 $\tau_{\mathbf{x}}$ 与$\mathbf{x}$ 相互独立可简化表述；允许二者相关是自然的扩展，主要影响常数和交叉位置，而非指数。(iv) 对于一个新的测试特征，我们将模型相对于均值目标的偏差记为 $\mathcal{E}_{\text{gen}}(\mathbf{x}) := \mathbf{x}^\top\hat{\boldsymbol{\theta}} - \mathbf{x}^\top\boldsymbol{\theta}^*$；其分布由训练过程和样本量 N 决定（见第 4 节）。

推论 3.2. 目标围绕其均值的平均方差为

$$\mathbb{E}\big[(Y_{\mathbf{x}}^* - \mathbf{x}^T\boldsymbol{\theta}^*)^2\big] = \mathbb{E}_{\mathbf{x}}[\mathrm{Var}(Y_{\mathbf{x}}^* \mid \mathbf{x})] = \sigma_\eta^2\,\mathbb{E}\left[\frac{1}{\tau_{\mathbf{x}}}\right]. \tag{4}$$

对于 $\tau_{\mathbf{x}} \sim$ 伽马分布($\beta/2$,1), $\mathbb{E}[1/\tau_{\mathbf{x}}] = \frac{\Gamma}{(\beta/2-1)\,\Gamma(\beta/2)} = \frac{2}{\beta-2}$ 当 $\beta > 2$ 时。

假设 3.3（有限平均目标方差）。 我们假设 $\mathbb{E}[(Y^*_{\mathbf{x}} - \mathbf{x}^T\boldsymbol{\theta}^*)^2] < \infty$，即 $\beta > 2$。

假设 3.3 支持对 $\mathcal{L}_{\text{gen}}$ （训练/泛化缩放）进行标准的高维岭回归/普通最小二乘分析。重要的是，我们关于推理时结果（pass@k 缩放）仅依赖于小-τ 尾部；即使在 $\beta \le 2$ 时，这些结果在指数意义上仍然成立（尽管常数和渐近分析的起始点可能会变化），前提是学得的预测器是一致的。

3.1 训练设置

我们考虑从数据集 $\mathcal{D}_N = \{(\mathbf{x}_i,\ y_i)\}_{i=1}^N$中学习均值关系 $f^*(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\theta}^*$，其中每个 y_i 都是根据定义 3.1 抽样得到的单个 实现。在微调的视角下，$\mathbf{x}_i$ 是来自预训练骨干网络的冻结特征，我们只训练一个线性头。学习者观察 ˆ（$\mathbf{x}_i,\ y_i$），并通过针对实际观测到的 标签最小化岭回归目标函数来估计 $\boldsymbol{\theta}$，而评估始终针对均值目标进行：

$$\mathcal{L}_{\text{train}}(\hat{\boldsymbol{\theta}}) \;=\; \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\theta}})^2 \;+\; \lambda\|\hat{\boldsymbol{\theta}}\|_2^2, \tag{5}$$

其中，$\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ （最后一层参数）， $\lambda \ge 0$ 是一个较小的正则化项。

该最小化器具有标准的闭式解

$$\hat{\boldsymbol{\theta}}_\lambda \;=\; \arg\min_{\hat{\boldsymbol{\theta}}} \mathcal{L}_{\text{train}}(\hat{\boldsymbol{\theta}}) \;=\; (N^{-1}X^\top X + \lambda I_d)^{-1}\, N^{-1}X^\top\mathbf{y}, \tag{6}$$

其中， $X \in \mathbb{R}^{N\times d}$将特征堆叠起来，$\mathbf{y} \in \mathbb{R}^N$表示方程 3 中的实现标签。对于适定情形，我们将取无岭回归极限 $\lambda \to 0$；在过参数化情形 $N < d$ 下，我们使用标准的˜缩放 $\lambda = \lambda/N$，以恢复最小范数插值器（见第 4 节和附录 C）。

相对于均值目标的偏差。在全文中，我们都针对均值信号$\mathbf{x}^\top\boldsymbol{\theta}^*$评估泛化，并将模型的逐实例偏差记为

$$\mathcal{E}_{\text{gen}}(\mathbf{x}) \;:=\; \mathbf{x}^\top\hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top\boldsymbol{\theta}^*. \tag{7}$$

测试（泛化）损失为 $\mathcal{L}_{\text{gen}}(N,\lambda) = \mathbb{E}_{\mathbf{x},\mathcal{D}_N}\left[\mathcal{E}_{\text{gen}}(\mathbf{x})^2\right]$。在假设 3.3 下，有效训练噪声具有有限的平均方差 $\sigma^2_{\text{noise}} = \sigma_\eta^2\,\mathbb{E}[1/\tau_{\mathbf{x}}] = \frac{2\sigma_\eta^2}{\beta-2}$，因此可以应用标准的高维岭回归/普通最小二乘工具。在第 4 节中，我们分析 $\mathcal{L}_{\text{gen}}$ 如何随 N、 d 和特征谱指数 α进行缩放，以及由此产生的 $\mathcal{E}_{\text{gen}}(\mathbf{x})$ 分布如何控制有限-N 推理行为（pass@k）和有效推理指数 $\beta_{\text{eff}}(N)$。

4 缩放定律的理论分析

我们分析两条相互耦合的定律：(i) 泛化损失 $\mathcal{L}_{\text{gen}}$ 对样本量 N 的依赖关系，以及 (ii) pass@k 推理失败 $\mathcal{L}_{\text{inf}}(k;\ N)$ 对试验次数 k 的依赖关系。在我们的微调视角下（对冻结特征进行最后一层回归）， $\mathcal{L}_{\text{gen}}$ 控制逐实例偏差的分布

$$\mathcal{E}_{\text{gen}}(\mathbf{x}) := \mathbf{x}^\top\hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top\boldsymbol{\theta}^*,$$

这反过来决定了 $\mathcal{L}_{\text{inf}}(k;\ N)$ 在有限 N 下的行为。随着 N 增大， $\mathcal{L}_{\text{gen}}(N)$ 减小，且 $\mathcal{L}_{\text{inf}}(k;\ N)$ 中由 $\mathbf{x}$ 引起的惩罚逐渐消退；观测到的推断斜率从有限 N 区域过渡到潜在难度渐近线 $-\beta$。我们通过实证方式，利用 $\beta_{\text{eff}}(N)$ 量化这一过渡（图 1，右）。

4.1 训练缩放定律 $\mathcal{L}_{\text{GEN}}$ 与 N

我们以均值目标为基准评估泛化，因此测试损失为

$$\mathcal{L}_{\text{gen}}(N,\lambda) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\,\mathcal{D}_N}\Big[\big(\mathbf{x}^\top\hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top\boldsymbol{\theta}^*\big)^2\Big]. \tag{8}$$

在高维情形下， $\mathcal{L}_{\text{gen}}$ 在很大程度上取决于 N–d 比率以及 $\boldsymbol{\Sigma}$的谱；我们遵循标准做法，将其分解为不同的情形（Belkin 等人，2019年；Hastie 等，2020年）。在假设 3.3 下，

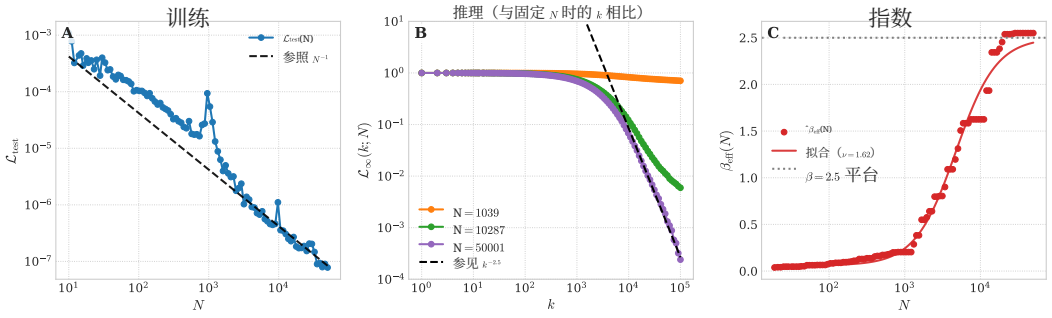

图 1：**LID 设定下的训练与推理时标度律**。 左： 泛化误差$\mathcal{L}_{\text{gen}}$ 与 N的关系，展示了双下降和两种经典区域。中： Pass@k 推理失败率 $\mathcal{L}_{\text{inf}}(k;\, N)$ 与 k的关系，针对若干 $N > d$，在均值已充分学习后，其渐近斜率为 $-\beta$（黑色虚线）。右： 有效推理指数 $\beta_{\text{eff}}(N)$，由固定 k窗口内的局部双对数斜率得出。点线标示渐近值 β；实线为经验拟合$\beta_{\text{eff}}(N) = \beta - \Delta/(1 + c_\beta N^\nu)$。我们使用 $\lambda = 10^{-9}$和 $\sigma_\eta = 10^{-3}$。

the 有效标签噪声的平均方差 $\sigma^2_{\text{noise}} = \sigma^2_\eta\, \mathbb{E}[1/\tau_{\mathbf{x}}]$ 有限，而经典的岭回归/OL S resu这些定律适用（附录 C 给出了采用本文记号的简明推导）。

过参数化区域（$N < d$）。 在无岭（或轻度正则化）插值下，误差由最小范数偏差和数据谱指数 α 决定（Bartlett 等，2020；Mei 和 Montanari，2020；Hastie 等，2020；Maloney 等，2022）。对于幂律谱 $\sigma_j^2 \propto j^{-(1+\alpha)}$，可得

$$\mathcal{L}_{\text{gen}}(N) \;\propto\; P_N\, N^{-\alpha}, \qquad N < d, \tag{9}$$

其中， P_N 是一个良性前因子，概括了谱和教师对齐的影响。此处符号 $\propto$ 表示，当 $N \to \infty$沿此区域变化时， $\mathcal{L}_{\text{gen}}(N)/\big(N^{-\alpha}\big)$ 始终有界，且远离零和无穷大（参见 B 节）。

欠参数化区域（$N > d$）。当样本数超过参数数时，方差项占主导地位，并按参数化速率衰减，

$$\mathcal{L}_{\text{gen}}(N) \;\propto\; \sigma_\eta^2\, \mathbb{E}\left[\frac{1}{\tau_{\mathbf{x}}}\right]\, \frac{d}{N}, \qquad N \gg d, \tag{10}$$

其斜率为通常的 $1/N$，常数项由噪声控制，反映出均值目标是通过单次标签学习得到的。当$\mathbf{x}$ 服从协方差为 $\mathbf{\Sigma}$ 的高斯分布时，精确的常数可以用逆 Marchenko–Pastur 定律表示（附录 C）。

过渡区域（$N \approx d$）。在插值附近，估计量会变得病态，且 $\mathcal{L}_{\text{gen}}$ 呈现峰值（“双重下降”）；其高度和宽度取决于 λ 以及谱（Belkin 等人，2019年）。

在过参数化区域，模型的表现受限于其解析数据低方差谱模态的速度，因此形成了 $N^{-\alpha}$ 尾部。一旦模型的样本数超过参数数，单次标签带来的有效噪声便占据主导地位， $\mathcal{L}_{\text{gen}}(N)$也转而受方差限制，并按 d/N衰减。这种经典行为构成了背景，供我们据此解读训练依赖的推理缩放变化。

图 1 左图确认了 LID 设定下的这些缩放关系，并提供了用于解读下文所讨论的有限-N 推理行为的 $\mathcal{L}_{\text{gen}}$（N）基线。

4.2 推理缩放规律 $\mathcal{L}_{\text{INF}}$ 与 k的对比

我们通过 pass@k 指标（Snell 等，2024；Brown 等，2024）分析推理，即在固定容差内， k 次独立试验中至少有一次与目标匹配的概率。 在 LID 设定下，对于具有潜在精度 $\tau_{\mathbf{x}}$ 的测试样本$\mathbf{x}$,

$$Y^*_{\mathbf{x}} \sim \mathcal{N}\big(\mathbf{x}^\top \boldsymbol{\theta}^*,\, \sigma_\eta^2/\tau_{\mathbf{x}}\big),$$

以及模型输出 $\hat{}y = \mathbf{x}^\top \hat{}\boldsymbol{\theta}_\lambda$。在试验 j 中，我们抽取 $y_j \sim {Y^*}_{\mathbf{x}}$， 误差为

$$e_j := \hat{y} - y_j = \big(\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*\big) - \eta_j = \mathcal{E}_{\text{gen}}(\mathbf{x}) - \eta_j, \qquad \eta_j \sim \mathcal{N}\big(0,\, \sigma_\eta^2/\tau_{\mathbf{x}}\big).$$

对于给定的容差 $\delta > 0$，若满足 $|e_j| \le \delta$，则试验成功。

假设 4.1（完美验证）。 我们假设存在一个完美验证器：对于给定的$\mathbf{x}$，只要 k 次试验中至少有一次满足 $|e_j| \le \delta$，就判定总体成功。

令 $p(\mathbf{x}\ \tau_{\mathbf{x}})$ 表示单次试验的失败概率

$$p(\mathbf{x}, \tau_{\mathbf{x}}) \coloneqq \mathbb{P}_{\eta \sim \mathcal{N}(0, \sigma_\eta^2/\tau_{\mathbf{x}})}\big(|\mathcal{E}_{\text{gen}}(\mathbf{x}) - \eta| > \delta\big).$$

假设各次试验之间条件独立，则 pass@k 失败概率为

$$\mathcal{L}_{\text{inf}}(k; N) \coloneqq \mathbb{E}_{\mathbf{x},\, \tau_{\mathbf{x}},\, \mathcal{D}_N}\big[p(\mathbf{x}, \tau_{\mathbf{x}})^{\,k}\big] = 1 - \text{pass@}k. \tag{11}$$

备注（模型抽样与目标抽样的 pass@k）。 在大型语言模型（LLM）的实际 pass@k 中，人们会抽取 k 个模型 输出，并根据固定目标对其进行验证；而在这里，我们抽取 k 个目标实现，并将其与固定预测器 $\hat{}y = \mathbf{x}^\top \hat{}\boldsymbol{\theta}_\lambda$ 进行比较。在容差窗口较小且局部密度平滑的情况下，单次尝试成功概率与在预测间隙 $B_N(\mathbf{x})$ 处评估的局部密度成正比：$= \mathcal{E}_{\text{gen}}(\mathbf{x})$。因此，随机性来自模型还是目标，只会影响常数项，不会影响 pass@k 曲线的渐近指数；我们在附录 D.2 中对这种等价性进行了形式化说明。

无偏渐近尾部。 我们首先考察均值学习得足够充分时的渐近标度：对于那些主导失败的小-$\tau_{\mathbf{x}}$ 实例，其逐实例偏差 $B_N(\mathbf{x})$ 相对于 $\sigma_\eta/\sqrt{\tau_{\mathbf{x}}}$ 可以忽略。在此情形下，对 δ 进行小窗口展开可得

$$1 - p(\mathbf{x}, \tau_{\mathbf{x}}) = c_\delta \sqrt{\tau_{\mathbf{x}}}\big(1 + o(1)\big) \quad \text{as } \delta \downarrow 0,$$

其中常数为 $c_\delta = 2\delta/(\sqrt{2\pi}\,\sigma_\eta)$；详情见附录 D。随后，在伽马先验 $\tau_{\mathbf{x}} \sim$ 伽马分布$(\beta/2,$ 1) 下对拉普拉斯–斯蒂尔切斯变换进行陶伯分析，可得以下形式化表述。

命题 4.2 (无偏 pass@k 标度)。 假设假设 *3.3* 以及附录 *E.2* 中的正则变差条件和光滑性条件成立，并考虑极限无偏预测器（或一列预测器，使得 $B_N(\mathbf{x}) \to 0$ 依分布收敛）。那么，当 $k \to \infty$ 时，

$$\mathcal{L}_{inf}(k) = \tilde{P}_{\beta,\delta,\sigma_\eta}\, k^{-\beta}\big(1 + o(1)\big), \qquad \tilde{P}_{\beta,\delta,\sigma_\eta} = \frac{2\,\Gamma(\beta)}{\Gamma(\beta/2)\, c_\delta^{\beta}}, \tag{12}$$

其中 $c_\delta = 2\delta/(\sqrt{2\pi}\,\sigma_\eta)$。

因此，渐近斜率 $-\beta$ 完全由 LID 先验的小-$\tau_{\mathbf{x}}$ 尾部决定，且与 d/N无关。

有限-N 修正与依赖训练的指数。

对于有限的 N，逐实例偏差 $B_N(\mathbf{x}) = \mathcal{E}_{\text{gen}}(\mathbf{x})$ 不可忽略，并会在中等 $\tau_{\mathbf{x}}$ 下压低每次试验的成功概率。对于较小的 δ，高斯累积分布函数展开给出

$$1 - p(\mathbf{x}, \tau_{\mathbf{x}}) = \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\delta}{\sigma_\eta} \sqrt{\tau_{\mathbf{x}}}\, \exp\Big(-\frac{B_N(\mathbf{x})^2\, \tau_{\mathbf{x}}}{2\sigma_\eta^2}\Big)\big(1 + o(1)\big), \tag{13}$$

当 $\delta \downarrow 0$ 时，该结果在 $(0, \infty)$ 的紧子集上一致成立；见附录 E.2。在标准线性泛化下， $B_N(\mathbf{x})$ 近似服从高斯分布，其 $\mathbb{E}[B_N(\mathbf{x})] = 0$，并且

$$\text{Var}[B_N(\mathbf{x})] = \Theta\big(\mathcal{L}_{\text{gen}}(N)\big).$$

利用方程 13 以及针对拉普拉斯–斯蒂尔切斯变换的陶伯论证（详见附录 E），我们得到 $\mathcal{L}_{\text{inf}}(k;\ N)$ 的双尾混合律。

定理 4.3（pass@ 的双尾混合律k）。 设附录 *E.2* 中的假设成立，并假设B_N（$\mathbf{x}$）是中心次高斯变量，且其方差$[B_N$（$\mathbf{x}$）$] \asymp \mathcal{L}_{gen}$（$N$），并且独立于 $\tau_{\mathbf{x}}$。那么˜ ˜存在正常数 P和 P_N（N），以及函数 γ（N）> 0，使得当 $k \to \infty$ 时，

$$\mathcal{L}_{inf}(k; N) = \tilde{P}\, k^{-\beta} + \tilde{P}_N(N)\, k^{-\gamma(N)}\big(1 + o(1)\big), \tag{14}$$

with

$$\gamma(N) = \Theta\Big(\frac{1}{\text{Var}[B_N(\mathbf{x})]}\Big) = \Theta\Big(\frac{1}{\mathcal{L}_{gen}(N)}\Big). \tag{15}$$

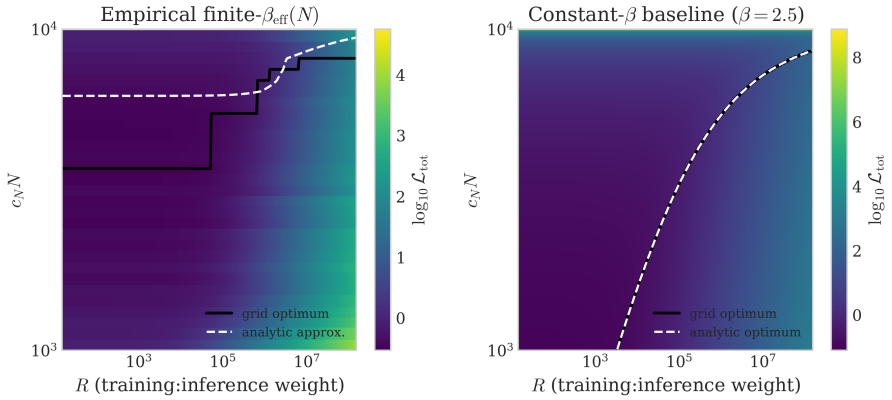

图 2：**依赖训练的推理指数下的计算分配**。 左图： 经验有限-N情形。我们绘制 $\log_{10} \mathcal{L}_{\text{tot}}(N,\ k;\ R)$，其中 R 为训练与推理权重比， $C = Nc_N + kc_k$. 黑色：网格最优解 $N_\star$；白色虚线：根据方程 21 得到的解析近似，使用 $\beta_{\text{eff}}(N)$（$\mathcal{L}_{\text{inf}}$的局部双对数斜率）及其离散导数。右图： 基于相同数据校准的常数-β 基线；白色虚线表示闭式解析最优解。有限-N面板中的等高线明显向较大的 N偏移，这与 $-\beta'_{\text{eff}}(N)\log(\cdot)$修正一致。

方程 14 中的第一项是由 β 决定的内在 LID 尾部；第二项是在中等 $\tau_{\mathbf{x}}$ 下由大偏差 $|B_N(\mathbf{x})|$ 导致的有限-N修正。

推论 4.4 （依赖训练的有效指数）。 固定一个 k-窗口 $[k_1, k_2]$，使得对于每个N，方程 *14* 中的 $k^{-\beta}$ 项或 $k^{-\gamma(N)}$ 项在该窗口内占主导。于是，当 $k_1 \to \infty$时，

$$\beta_{\text{eff}}(N; [k_1, k_2]) = \min\{\beta, \gamma(N)\} + o(1), \tag{16}$$

其中 $\beta_{\text{eff}}(N; [k_1, k_2])$ 由方程 *23* 定义。特别是，随着 N 增大且 $\mathcal{L}_{gen}(N)$ 减小， $\gamma(N)$ 增大，而有效推理指数 $\beta_{\text{eff}}(N)$ 则单调趋近于内在 *LID* 指数 β。

两个指数 β 和 $\gamma(N)$ 起着不同的作用。内在指数 β 完全源自实例难度分布的重尾；一旦均值目标已学会，额外训练也无法改善它。相比之下， $\gamma(N)$ 则体现了有限-N 惩罚：当模型的平均误差仍然很大时，许多测试点即使其潜在的$\tau_{\mathbf{x}}$ 适中，表现也会如同“实际上很难”一样。随着训练降低 $\mathcal{L}_{\text{gen}}(N)$，这种人为造成的难尾部会消失，观测到的 pass@k 曲线随之变陡，直到 $\beta_{\text{eff}}(N)$ 仅受 β 限制。

在实践中，我们用以下经验拟合来概括这种单调饱和现象：

$$\beta_{\text{eff}}(N) = \beta - \frac{\Delta}{1 + c_\beta N^\nu}, \tag{17}$$

其中，(Δ, c_β, ν) 对 k 窗口和估计量的依赖较弱。图 1 的中间面板和右侧面板展示了：(i) 固定 $N > d$ 时， $\mathcal{L}_{\text{inf}}(k;\ N)$ 的斜率随 N 增大而变陡；(ii) $\beta_{\text{eff}}(N)$ 随 N 增大，并在 β处达到平台。

4.3 计算资源分配权衡：训练计算-依赖型 $\beta_{\text{eff}}(N)$

在刻画了有限-N 推理缩放之后，我们可以结合训练和推理规律，研究两者之间的最优预算分配。

我们考虑一个固定的计算资源预算 C，必须在训练样本（每个成本为 c_N）和推理试验（每个成本为 c_k） $C = Nc_N + kc_k$之间分配，满足 $k = \frac{C-N}{c_k}$， $\tilde{N} := Nc_N$. 我们最小化一个加权目标函数

$$\mathcal{L}_{\text{tot}}(\tilde{N}) = R\,\mathcal{L}_{\text{gen}}(N) + \mathcal{L}_{\text{inf}}\Big(k = \tfrac{C-\tilde{N}}{c_k};\ N\Big), \qquad \text{training/inference weight} = R > 0. \tag{18}$$

在经典（欠参数化）情形 $N \gg d$下，我们有 $\mathcal{L}_{\text{gen}}(N) \propto P_N N^{-\gamma}$，其中 $\gamma = 1$（第 4 节.1）；更一般地，可以取取决于 d/N 的 $\gamma \in \{1, \alpha\}$。对于推理，定理 4.3

建议在实际的 k-窗口内对 $\mathcal{L}_{\text{inf}}(k;\, N)$ 建模为

$$\mathcal{L}_{\text{inf}}(k; N) \approx \tilde{P}(N)\, k^{-\beta_{\text{eff}}(N)}, \qquad \beta_{\text{eff}}(N) \uparrow \beta \text{ as } N \to \infty, \tag{19}$$

其中， $P(N)$是缓变前因子，用于捕捉残余偏差效应。此处的“≈”应理解为在固定的 k 窗口内进行的经验拟合，而非渐近性陈述；当 $k \to \infty$ 时的渐近行为由定理 4.3 给出。

在 $N = Nc_N$ 和 $k = (C - N)/c_k$ 的条件下，受预算约束的目标变为

$$\mathcal{L}_{\text{tot}}(\tilde{N}) = R\, P_N\, c_N^{\gamma}\, \tilde{N}^{-\gamma} + \tilde{P}(N)\, c_k^{\beta_{\text{eff}}(N)}\, (C - \tilde{N})^{-\beta_{\text{eff}}(N)}. \tag{20}$$

命题 *4.5*（训练依赖的 $\beta_{\text{eff}}(N)$ 下的最优分配）。假设 $\mathcal{L}_{gen}(N) \propto P_N N^{-\gamma}$，且在实际的 k 窗口内，*pass@*k 失败率满足 $\mathcal{L}_{inf}(k;\, N) \approx \tilde{P}(N)k^{-\beta_{\text{eff}}(N)}$，其中 $\beta_{\text{eff}}(N)$ 关于 N 可微。令 $N = N/c_N$。$\mathcal{L}_{\text{tot}}(N)$ 的任一内部最优解 $N_* \in (0, C)$ 都满足

$$R\, P_N\, c_N^{\gamma}\, \gamma\, \tilde{N}^{-(\gamma+1)} = \tilde{P}(N)\, c_k^{\beta_{\text{eff}}(N)}\, (C - \tilde{N})^{-\beta_{\text{eff}}(N)} \left[\frac{\beta_{\text{eff}}(N)}{C - \tilde{N}} - \frac{\beta'_{\text{eff}}(N)}{c_N}\, \ln\!\Big(\frac{C - \tilde{N}}{c_k} \Big) \right]. \tag{21}$$

如果 $P(N)$ 是缓变的，且在相关范围内 $|\beta'_{\text{eff}}(N)|$ 较小，则公式 *21* 可简化为准静态平衡

$$R\, P_N\, c_N^{\gamma}\, \gamma\, \tilde{N}^{-(\gamma+1)} \approx \tilde{P}\, c_k^{\beta}\, \beta\, (C - \tilde{N})^{-(\beta+1)}, \tag{22}$$

其中， P是一个与 N无关的常数；这里的近似意味着两边相差至多缓变的乘法因子。

与常数-β 条件相比，方程 21 包含一个新的对数项，与 $-\beta'_{\text{eff}}(N) \ln((C - \tilde{N})/c_k)$成正比。当预算使得 $k = (C - \tilde{N})/c_k$低于LID 主导区间（因此 $\beta'_{\text{eff}}(N) > 0$且 $\ln((C - N)/c_k) > 0$）时，该项会提高训练的边际收益，使最优值向更大的 N偏移。一旦 $\beta_{\text{eff}}(N)$达到饱和（因此$\beta'_{\text{eff}}(N) \approx 0$），方程 21 就简化为方程 22 中的常指数平衡关系（恢复经典权衡，其中 β被 $\beta_{\text{eff}}(N)$取代）。
具体而言：

- 欠参数化，接近饱和。当 $\gamma = 1, \beta'_{\text{eff}}(N) \approx 0$且 $P(N)$缓变时，方程 22 给出了准确的分配规则：持续增加训练，直到 N增益$\propto \tilde{N}^{-2}$的边际收益与 k增益 $\propto (C - \tilde{N})^{-(\beta_{\text{eff}}+2)}$的边际收益相等。

- 有限N，次渐近 k。当 $\beta_{\text{eff}}(N)$ 仍在增加时，$-\beta'_{\text{eff}}(N) \log(\cdot)$ 项意味着，额外训练带来的价值严格高于准静态近似的预测；在有效斜率稳定之前，最优策略会将更多资源投入 N。

在有限-N 区间内，转向更大的 N 这一趋势在图 2 中显而易见。

5 真实应用的实证结果

为将 LID 模型与现实设定联系起来，我们进行了两项实验：在从 CIFAR-10H 提取的特征上训练线性分类器，以及使用自回归大语言模型在推理任务上进行教师–学生蒸馏实验。我们在以下章节讨论这些实验的结果，更多细节见附录 F 和 H。

5.1 测试案例 I：使用预训练特征的 CIFAR-10H 中的 LID

在此，我们使用 CIFAR-10（Krizhevsky，2012）及其配套的 CIFAR-10H 数据集中的人工标注分布（Peterson 等，2019）验证 LID 的预测。我们冻结一个预训练 ResNet-18（He 等，2015），并仅在骨干网络特征之上微调一个线性头，将人工标注分歧视为依赖实例的噪声（高方差 $\Rightarrow$ 低精度），以匹配 LID 设定。完整的实验方案和实现细节见附录 F。

图 3（上）展示了 $\mathcal{L}_{\text{gen}}$与 N，以及 $\mathcal{L}_{\text{inf}}(\cdot;\, N)$ 与 k[1] 的关系。泛化曲线在 $N \approx d$ 附近呈现出熟悉的过渡区域，并在 $N \gg d$ 时表现出 $1/N$ 衰减，这与我们的线性岭回归分析一致。对于

[1]这里， $s = 32$ 和 $\delta = 0.05$遵循附录 F 中的记号。

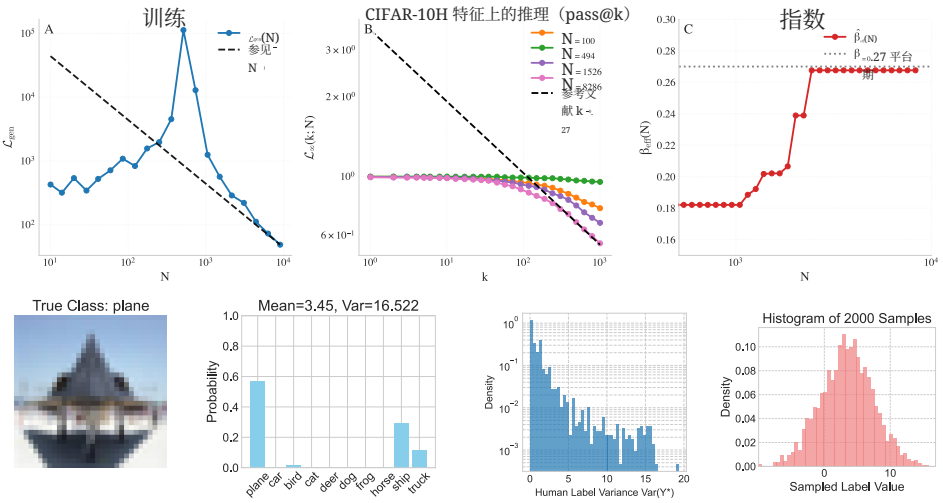

图 3：**CIFAR-10H 上的训练与推理缩放（冻结骨干网络、线性头）。顶行：**左：经典情形下带有 N^{-1}尾部的泛化损失 $\mathcal{L}_{\text{gen}}(N)$。中：CIFAR-10H 上多个 N的 Pass@k失败 $\mathcal{L}_{\text{inf}}(k;\ N)$；虚线表示以最大-$N$曲线为锚点的 $k^{-\beta}$ 参考线。右：根据局部双对数斜率估计的有效推理斜率 $\beta_{\text{eff}}(N)$；叠加了趋于内在尾指数 β 的饱和拟合 $\beta_{\text{eff}}(N) = \beta - \Delta/(1 + c_\beta N^\nu)$。**底行：**一张 CIFAR-10 图像、其人工标签分布，以及用于采样训练/推理标签的高斯概率密度函数，展示通过标签方差体现的实例难度。

推理而言，每条固定-N曲线都近似遵循幂律 $k^{-\beta_{\text{eff}}(N)}$；我们将 $\beta_{\text{eff}}(N)$ 估计为中心 k 窗口内的局部双对数斜率。正如第 4.2 节所预测的，$\beta_{\text{eff}}(N)$ 随 N 增加而趋于饱和，这反映了从偏差受限窗口（有限-N 均值误差）到 LID 主导的尾部这一过渡区域。

底行展示了人工标注分歧如何引发实例难度：较宽的标签直方图（较大的实例噪声方差）会在方程 48 中产生更宽的高斯分布，使得在较小 k 下 Pass@k 成功更加罕见。这与合成 LID 模型（第 4.2 节）中 $k^{-\beta}$ 定律背后的机制相一致，但需要注意，经验难度分布并不完全服从伽马分布；尽管如此，斜率行为仍然稳健，并由高方差实例的普遍程度所决定。

要点。 (i) 线性头微调设置重现了岭回归分析所预测的 $\mathcal{L}_{\text{gen}}$ 缩放规律。(ii) pass@k失败曲线呈现对数–对数线性关系，其斜率 $\beta_{\text{eff}}(N)$ 随训练数据增加而增大并最终进入平台期，这与有限-N 理论一致，并支持第 4.3 节中的计算资源分配结果。(iii) 将人类不确定性视为依赖实例的噪声，为 LID 提供了一个现实的测试平台：尽管偏离理想化的伽马先验或 $\tau_{\mathbf{x}}$ 与 $\mathbf{x}$ 之间的独立性假设，定性缩放规律和依赖训练的推理指数仍然存在。

5.2 测试案例 II：GSM8K 教师–学生蒸馏

我们通过一个小型 GSM8K（Cobbe 等，2021）蒸馏任务补充 CIFAR-10H 实验，该任务旨在探究 LID 训练–推理耦合。一个 Flan-T5-XL（Chung 等，2022）教师在 GSM8K 训练集上生成解题轨迹；一个 Flan-T5-small 学生使用 LoRA（Hu 等，2021）进行微调（r=8），训练子集大小为 $N \in [10,\ 6309]$（每道题一个教师样本）。我们在 GSM8K 测试集上使用两项指标进行评估：(i) 一个严格贪心 代理指标，用于衡量训练损失：该指标解码单个学生答案，仅解析明确的“最终答案”数值，并在容差 $\delta = 10^{-3}$ 下报告仅有效样本均方误差（经过 1% 缩尾处理）；以及 (ii) pass@k失败 $\mathcal{L}_{\text{inf}}(k;\ N)$，以及从中心 k-窗口估计得到的ˆ有效斜率 $\beta_{\text{eff}}(N)$。

图 4 显示，严格贪心损失随 N^2 小幅下降，而 $\mathcal{L}_{\text{inf}}(k;\ N)$ 变陡，且 ˆ$\beta_{\text{eff}}(N)$ 先上升后趋于饱和；这正是 LID 模型（§4.2，附录 E）所预测的双尾行为。完整的设置和评估细节见附录 H。

[2]尽管测试损失因较大的数值不匹配而不稳定（有些最终答案的数值就是比其他答案大），但整体趋势仍在下降，推理指标也保持稳定。

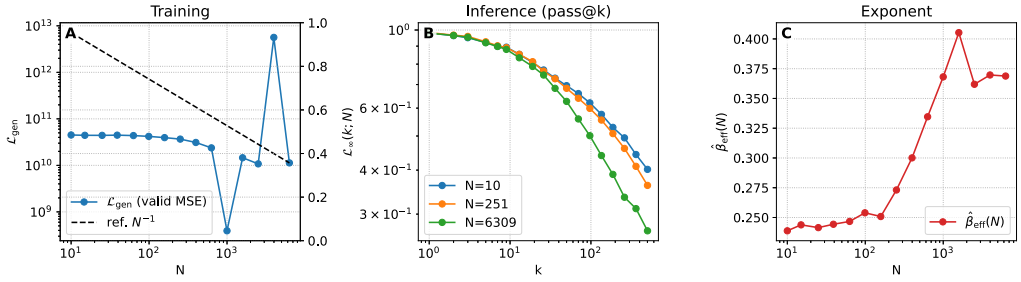

图 4：**GSM8K 蒸馏（教师→学生）**。左：严格贪心解码下的仅有效样本均方误差，对比 N（1% 温莎化），并显示 N^{-1}参考值。中：$\mathcal{L}_{\text{inf}}$（k；N）与 k的关系，展示三种 N取值。右：有效推断斜率 $\hat{}\ \beta_{\text{eff}}$（$N$），该值随 N增加而上升并趋于平台期，这与 LID 双尾定律一致。

6 讨论与结论

本文引入了潜在实例难度（LID）模型，这是一种简单且可求解的最后一层微调框架，统一了训练和推理的缩放定律。我们对具有内在的实例异质性难度的任务进行建模，并证明，尽管泛化损失 $\mathcal{L}_{\text{gen}}$ 遵循已确立的、关于样本量 N 和数据谱 α 的缩放规律，但其改进会直接且非平凡地影响推理性能。

我们的核心贡献是推导出一种**依赖训练的推理指数**。pass@k失败率 $\mathcal{L}_{\text{inf}}$（k）按幂律衰减，但其指数 β_{eff}（N）并非固定不变。对于训练不足的模型，该指数起初较小；随着训练样本数 N增加，指数逐渐增大，最终在一个内在极限 β处饱和，而该极限由任务真实难度分布的尾部决定。这一机制揭示了，相对于均值目标降低模型误差，如何让推理时计算更有效；但这种效果存在上限，数据中的不可约随机性决定了收益递减的临界点。

这一统一视角带来了可付诸实践的洞见。它预测了最优资源分配策略会出现明确的转折点：当模型训练不足且 $\beta_{\text{eff}}(N) < \beta$ 时，获取更多训练数据的边际收益很高。一旦模型训练充分且推理指数达到饱和，进一步的收益最好通过投入更多推理时计算来获取。

LID 模型虽然简单，却提供了有价值的理论基线。它清晰地区分了数据结构（α）与数据异质性（β）各自的作用，同时也解释了二者如何通过训练过程相互耦合。通过为测试时计算在何时以及应增加多少才能发挥作用提供一种闭式、可检验的理论，它为更有原则地理解现代机器学习中的资源分配迈出了第一步。

尽管我们在采用标量回归的最后一层微调中实例化了 LID，但其核心机制——pass@k 失败由一个内在 难度尾部（索引 β）与一个由 $\mathcal{L}_{\text{gen}}(N)$ 控制的依赖训练的 尾部混合而成——与模型无关。除了微调之外，还可以通过允许表征不断演化来分析从头训练，这实际上会改变特征谱（例如衰减指数 α），从而改变有限-N 尾部的收缩速率；此时饱和水平可能取决于架构。对于分类任务，实例难度可以等同于标签歧义/噪声，而 pass@k 表示 k 次抽样中至少有一次命中正确类别的事件，由此得到相同的双尾混合以及会达到饱和的 $\beta_{\text{eff}}(N)$。在强化学习中，难度对应于回报方差，因此 pass@k 表示 k 条轨迹采样中包含近似最优回报的概率；同样，LID 的预测也适用。对于大语言模型推理，增加预训练词元会降低平均误差，并在经验上使 pass@k 斜率变陡；我们推测 $\beta_{\text{eff}}(N)$ 会随规模增长，并在某个内在难度指数处达到饱和，同时呈现相同的可观测特征：随着训练改进，pass@k 曲线先变陡，随后趋于平台。

局限性。为清晰起见，我们分析了一个线性末层模型；将理论推广到多层非线性架构是重要的下一步，而随机特征模型或核模型则是自然的桥梁。我们还将内在难度分布视为任务固定不变；对于复杂的大语言模型推理，更强的模型可能既能降低偏差，也能有效地“简化”任务，从而改变表面上的尾指数 β。因此，我们的结果应理解为：对于一个固定的架构，$\beta_{\text{eff}}(N)$如何上升并趋近于依赖于架构的平台值；如何量化这一平台值在不同模型间的变化，仍有待研究。最后，我们研究的是单输出回归；如何将关于训练-推理耦合的 LID 视角推广到完全结构化的自回归输出（如大型语言模型中的输出），是未来研究面临的一项关键挑战。

7 致谢

感谢 Yohai Bar-Sinai、Alon Beck、Francesco Cagnetta、Joshua Kazdan、Nadav Outmezguine 和 Zohar Ringel 进行富有成效的讨论。NL 得到了洛桑联邦理工学院 AI4science/AI 中心计划的支持。

参考文献

Jared Kaplan、Sam McCandlish、Tom Henighan、Tom B Brown、Benjamin Chess、Rewon Child、Scott Gray、Alec Radford、Jeffrey Wu 和 Dario Amodei。《神经语言模型的缩放定律》。*arXiv* 预印本 *arXiv:2001.08361*，2020年。

Joel Hestness、SharanNarang、Newsha Ardalani、GregoryDiamos、HeewooJun、Hassan Kianinejad、Md Patwary、Mostofa Ali、Yang Yang 和 Yanqi Zhou。《深度学习的缩放具有可预测性：实证研究》。*arXiv* 预印本 *arXiv:1712.00409*，2017年。

Jonathan S. Rosenfeld、Amir Rosenfeld、Yonatan Belinkov 和 Nir Shavit。《跨尺度泛化误差的构造性预测》，2019年。网址：https://arxiv.org/abs/1909. 12673。

Yasaman Bahri、Ethan Dyer、Jared Kaplan、Jaehoon Lee 和 Utkarsh Sharma。《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。

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

Blake Bordelon、Alexander Atanasov 和 Cengiz Pehlevan。神经网络缩放定律的一个动力学模型，2024。URL https://arxiv.org/abs/2402.01092。

Peter L. Bartlett、Philip M. Long、Gábor Lugosi 和 Alexander Tsigler。线性回归中的良性过拟合。《美国国家科学院院刊》，117(48):30063–30070，2020年4月。ISSN 1091-6490。doi: 10.1073/pnas.1907378117。URL http://dx.doi.org/10.1073/ pnas.1907378117。

Trevor Hastie、Andrea Montanari、Saharon Rosset 和 Ryan J. Tibshirani。高维无岭最小二乘插值中的意外发现，2020年。网址：https://arxiv.org/abs/ 1903.08560。

Blake Bordelon、Abdulkadir Canatar 和 Cengiz Pehlevan。核回归和宽神经网络中的谱依赖学习曲线。载于国际机器学习大会，第 1024–1034 页。PMLR，2020年。

Yi Tay、Mostafa Dehghani、Samira Abnar、Hyung Won Chung、William Fedus、Jinfeng Rao、Sharan Narang、Vinh Q. Tran、Dani Yogatama 和 Donald Metzler。缩放定律与模型架构：归纳偏置如何影响缩放？，2022年。网址：https://arxiv.org/abs/2207. 10551。

Charlie Snell、Jaehoon Lee、Kelvin Xu 和 Aviral Kumar。以最优方式扩展大语言模型测试时计算，可能比扩展模型参数更有效，2024。网址：https://arxiv.org/abs/ 2408.03314。

Bradley Brown、Jordan Juravsky、Ryan Ehrlich、Ronald Clark、Quoc V. Le、Christopher Ré 和 Azalia Mirhoseini。《大型语言猴子：通过重复采样扩展推理计算》，2024 年。网址：https://arxiv.org/abs/2407.21787。

Noam Levi。《推理缩放定律的一个简单模型》，2024 年。网址：https://arxiv.org/abs/ 2410.16377。

Rylan Schaeffer、Joshua Kazdan、John Hughes、Jordan Juravsky、Sara Price、Aengus Lynch、Erik Jones、Robert Kirk、Azalia Mirhoseini 和 Sanmi Koyejo。《大型语言猴子如何获得其力量（定律）？》，2025a。网址：https://arxiv.org/abs/2502.17578。

Devansh Arpit、Stanislaw Jastrzebski、Nicolas Ballas、David Krueger、Emmanuel Bengio、MaxinderS. Kanwal、Tegan Maharaj、Asja Fischer、Aaron Courville、Yoshua Bengio 和 Simon Lacoste-Julien。《深入探究深度网络中的记忆》，2017年。网址：https://arxiv.org/abs/ 1706.05394。

Curtis G. Northcutt、Anish Athalye 和 Jonas Mueller。《测试集中的普遍标签错误会破坏机器学习基准》，2021年。网址：https://arxiv.org/abs/2103.14749。

Jordan Hoffmann 等人。《训练计算量最优的大语言模型》。《arXiv 预印本 arXiv:2203.15556》，2022年。

Yasaman Bahri、Ethan Dyer、Jared Kaplan、Jaehoon Lee 和 Utkarsh Sharma。《神经网络缩放定律的解释》。《arXiv 预印本 arXiv:2102.06701》，2021年。

Joshua C. Peterson、Ruairidh M. Battleday、Thomas L. Griffiths 和 Olga Russakovsky。《人类不确定性使分类更稳健》，2019年。网址：https://arxiv.org/abs/1908. 07086。

Noam Levi 和 Yaron Oz。《混沌与湍流的普适统计结构和缩放定律》，2023。网址：https://arxiv.org/abs/2311.01358。

Noam Levi 和 Yaron Oz。复杂数据集的底层缩放定律与普适统计结构，2024。网址：https://arxiv.org/abs/2306.14975。

Mikhail Belkin、Daniel Hsu、Siyuan Ma 和 Soumik Mandal。调和现代机器学习实践与经典偏差—方差权衡。《美国国家科学院院刊》，116(32):15849–15854，2019年7月。ISSN 1091-6490。doi：10.1073/pnas.1903070116。网址：http://dx.doi.org/10.1073/pnas.1903070116。

Song Mei 和 Andrea Montanari。随机特征回归的泛化误差：精确渐近分析与双下降曲线，2020年。网址：https://arxiv.org/abs/1908. 05355。

Alex Krizhevsky。从微型图像中学习多层特征。多伦多大学，2012年5月。网址：https://www.cs.toronto.edu/~kriz/ learning-features-2009-TR.pdf。

Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。《用于图像识别的深度残差学习》，2015年。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。《训练验证器解决数学应用题》，2021年。网址：https://arxiv.org/abs/2110.14168。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Yunxuan Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma、Albert Webson、Shixiang Shane Gu、Zhuyun Dai、Mirac Suzgun、Xinyun Chen、Aakanksha Chowdhery、Alex Castro-Ros、Marie Pellat、Kevin Robinson、Dasha Valter、Sharan Narang、Gaurav Mishra、Adams Yu、Vincent Zhao、Yanping Huang、Andrew Dai、Hongkun Yu、Slav Petrov、Ed H. Chi、Jeff Dean、Jacob Devlin、Adam Roberts、Denny Zhou、Quoc V. Le 和 Jason Wei。《指令微调语言模型的扩展》，2022年。网址：https://arxiv.org/abs/2210.11416。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。《 LoRA：大语言模型的低秩适配》，2021 年。网址：https: //arxiv.org/abs/2106.09685。

Tom Henighan、Jared Kaplan、Mor Katz、Mark Chen、Christopher Hesse、Jacob Jackson、Hee- woo Jun、TomB.Brown、PrafullaDhariwal、ScottGray、Chris Hallacy、Benjamin Mann、Alec Radford、Aditya Ramesh、Nick Ryder、Daniel M. Ziegler、John Schulman、Dario Amodei 和 Sam McCandlish。《自回归生成建模的缩放定律》，2020年。网址：https://arxiv.org/abs/2010.14701。

Utkarsh Sharma 和 Jared Kaplan。数据流形维度导出的缩放定律。机器学习研究杂志，23(9):1–34，2022年。网址：http://jmlr.org/papers/v23/ 20-1111.html。S Spigler、M Geiger、S d'Ascoli、L Sagun、G Biroli 和 M Wyart。由欠参数化向过参数化的拥塞转变会影响深度学习中的泛化。《物理学杂志 *A*：数学与理论》，52(47):474001，2019年10月。ISSN 1751-8121。doi: 10.1088/1751-8121/ab4c8b。网址：http://dx.doi.org/10.1088/1751-8121/ab4c8b。Ethan Caballero、Kshitij Gupta、Irina Rish 和 David Krueger。破碎神经网络缩放定律，2023。网址：https://arxiv.org/abs/2210.14891。Preetum Nakkiran、Gal Kaplun、Yamini Bansal、Tristan Yang、Boaz Barak 和 Ilya Sutskever。深度双重下降：更大的模型和更多数据何以有害。《统计力学杂志：理论与实验》，2021(12):124003，2021年。Ben Sorscher、Robert Geirhos、Shashank Shekhar、Surya Ganguli 和 Ari S Morcos。超越神经网络缩放定律：通过数据剪枝突破幂律缩放。*arXiv* 预印本 *arXiv:2206.14486*，2022年。Nolan Dey、Bin Claire Zhang、Lorenzo Noci、Mufan Li、Blake Bordelon、Shane Bergsma、Cengiz Pehlevan、Boris Hanin 和 Joel Hestness。别偷懒：Completep 实现计算高效的深度 Transformer，2025年。网址：https://arxiv.org/abs/2505.01618。Danny Hernandez、Jared Kaplan、Tom Henighan 和 Sam McCandlish。迁移的缩放定律，2021年。网址：https://arxiv.org/abs/2102.01293。Haowei Lin、Baizhou Huang、Haotian Ye、Qinyu Chen、Zihao Wang、Sujian Li、Jianzhu Ma、Xiaojun Wan、James Zou 和 Yitao Liang。通过修正缩放定律选择要微调的大语言模型，2024年。网址：https://arxiv.org/abs/2402.02314。

Robert Irvine、Douglas Boubert、Vyas Raina、Adian Liusie、Ziyi Zhu、Vineet Mudupalli、Aliaksei Korshuk、Zongyi Liu、Fritz Cremer、Valentin Assassi、Christie-Carol Beauchamp、Xiaoding Lu、Thomas Rialan 和 William Beauchamp。《奖励聊天机器人与数百万用户进行现实世界中的互动》，2023。网址：https://arxiv.org/abs/2303.06135。

Niklas Muennighoff、Zitong Yang、Weijia Shi、Xiang Lisa Li、李飞飞、Hannaneh Hajishirzi、Luke Zettlemoyer、Percy Liang、Emmanuel Candès 和 Tatsunori Hashimoto。《s1：简单的测试时扩展》，2025年。网址：https://arxiv.org/abs/2501.19393。Joshua Kazdan、Rylan Schaeffer、Youssef Allouah、Colin Sullivan、Kyssen Yu、Noam Levi 和 Sanmi Koyejo。《大语言模型中 pass@k 扩展的高效预测》，2025年。网址：https://arxiv.org/abs/2510.05197。Rylan Schaeffer、Noam Levi、Brando Miranda 和 Sanmi Koyejo。《语言模型生成式评估的预训练扩展定律》，2025b。网址：https://arxiv.org/abs/2509. 24012。Feng Chen、Allan Raventos、Nan Cheng、Surya Ganguli 和 Shaul Druckmann。《重新思考测试时计算扩展中的微调：限制置信度可提升数学推理能力》，2025年。网址：https://arxiv.org/abs/2502.07154。Zhenyu Hou、Xin Lv、Rui Lu、Jiajie Zhang、Yujiang Li、Zijun Yao、Juanzi Li、Jie Tang 和 Yuxiao Dong。《通过强化学习与推理缩放推进语言模型推理》，2025年。网址：https://arxiv.org/abs/2501.11651。Kuang-Huei Lee、Ian Fischer、Yueh-Hua Wu、Dave Marwood、Shumeet Baluja、Dale Schuurmans 和 Xinyun Chen。《不断进化的大语言模型思维》，2025年。网址：https://arxiv.org/abs/ 2501.09891。Kanishk Gandhi、Denise Lee、Gabriel Grand、Muxin Liu、Winson Cheng、Archit Sharma 和 Noah D. Goodman。《搜索流（SoS）：学习在语言中进行搜索》，2024年。网址：https: //arxiv.org/abs/2404.03683。

Yangzhen Wu、Zhiqing Sun、Shanda Li、Sean Welleck 和 Yiming Yang。《推理缩放定律：使用语言模型进行问题求解时计算最优推理的实证分析》，2024年。网址：https://arxiv.org/abs/2408.00724。Jiacheng Liu、Andrew Cohen、Ramakanth Pasunuru、Yejin Choi、Hannaneh Hajishirzi 和 Asli Celikyilmaz。《别丢弃你的价值模型！利用价值引导的蒙特卡洛树搜索解码生成更多偏好文本》，2024年。网址：https://arxiv.org/abs/2309. 15028。Shun Zhang、Zhenfang Chen、Yikang Shen、Mingyu Ding、Joshua B. Tenenbaum 和 Chuang Gan。《使用大语言模型进行代码生成规划》，2023年。网址：https://arxiv.org/ abs/2303.05510。Andy Zhou、Kai Yan、Michal Shlapentokh-Rothman、Haohan Wang 和 Yu-Xiong Wang。《语言智能体树搜索统一了语言模型中的推理、行动与规划》，2024年。网址：https: // arxiv.org/abs/2310.04406。

Sehyun Choi、Tianqing Fang、Zhaowei Wang 和 Yangqiu Song。KCTS：结合令牌级幻觉检测的知识约束树搜索解码，2023。网址 https://arxiv.org/ abs/2310.09044。

Yuxi Xie、Kenji Kawaguchi、Yiran Zhao、Xu Zhao、Min-Yen Kan、Junxian He 和 Qizhe Xie。用于推理的自评估引导的束搜索，2023。网址 https://arxiv.org/abs/ 2305.00633。

Hunter Lightman、Vineet Kosaraju、Yura Burda、Harri Edwards、Bowen Baker、Teddy Lee、Jan Leike、John Schulman、Ilya Sutskever 和 Karl Cobbe。让我们逐步验证，2023。网址 https://arxiv.org/abs/2305.20050。

Peiyi Wang、Lei Li、Zhihong Shao、R. X. Xu、Damai Dai、Yifei Li、Deli Chen、Y. Wu 和 Zhifang Sui。Math-Shepherd：无需人工标注，逐步验证并强化大型语言模型，2024a。网址 https://arxiv.org/abs/2312.08935。

Zhilin Wang、Yi Dong、Olivier Delalleau、Jiaqi Zeng、Gerald Shen、Daniel Egert、Jimmy J. Zhang、Makesh Narsimhan Sreedhar 和 Oleksii Kuchaiev。Helpsteer2：用于训练顶尖奖励模型的开源数据集，2024b。网址：https://arxiv.org/abs/2406.08673。Huajian Xin、Daya Guo、Zhihong Shao、Zhizhou Ren、Qihao Zhu、Bo Liu、Chong Ruan、Wenda Li 和 Xiaodan Liang。Deepseek-prover：通过大规模合成数据推动大型语言模型中的定理证明，2024年。网址：https://arxiv.org/abs/2405.14333。Zachary Ankner、Mansheej Paul、Brandon Cui、Jonathan D. Chang 和 Prithviraj Ammanabrolu。大声批判奖励模型，2024。网址：https://arxiv.org/abs/2408.11791。Nikolaos Tsilivis、Eran Malach、Karen Ullrich 和 Julia Kempe。下一令牌预测之后的强化学习如何促进学习，2025年。网址：https://arxiv.org/abs/2510. 11495。Jack W Silverstein 和 Z. D. Bai。关于一类大维随机矩阵的特征值经验分布。《多元分析杂志》，54(2):175–199，1995。Romain Couillet 和 Zhenyu Liao。《机器学习的随机矩阵方法》。剑桥大学出版社，2022。doi：10.1017/9781009128490。

A 相关工作

泛化缩放定律 神经网络缩放定律（NSL）已被证明能够准确描述深度神经网络的泛化损失如何随规模（模型/数据集规模、计算量）增大而改善（Hestness 等，2017；Kaplan 等，2020；Hoffmann 等，2022）。例如，语言模型的交叉熵损失通常会在多个数量级上呈现幂律缩放（Kaplan 等，2020；Henighan 等，2020），这表明增加资源能够带来可量化的性能提升。理论框架旨在通过识别不同的缩放区间（例如方差受限、分辨率受限）来解释这些观察结果（Bahri 等人，2021；Sharma 和 Kaplan，2022；Spigler 等人，2019），有时还会将其与数据流形的内在维数联系起来（Sharma 和 Kaplan，2022；Bahri 等人，2021）。更精细的模型研究了训练时间带来的动态缩放演化（Bordelon 等，2024），以及刻画幂律区间之间过渡的“破碎”神经网络缩放定律（BNSL）（Caballero 等人，2023；Nakkiran 等人，2021）。这类转折凸显了外推的复杂性，并表明平台期可能源于策略欠佳，而非根本性限制（Sorscher 等人，2022；Dey 等人，2025）。除预训练之外，缩放定律也适用于微调表现与预训练模型规模、微调数据量之间的关系（Hernandez 等人，2021；Lin 等人，2024）。这些研究表明，预训练实际上扩充了微调数据，迁移收益遵循可预测但取决于任务的规律（Hernandez 等人，2021），使研究重点转向如何利用预训练服务多样化的下游应用。在这项工作中，我们主要关注标准的神经网络缩放定律设定，即缩放定律描述泛化损失如何随训练样本数量增加而改善。

推理时缩放

深度学习中扩展推理时计算的方法，通常涉及生成多个候选解，然后根据特定标准选出最佳解。这些标准包括在多数投票中选择出现频率最高的回答，或在 Best-of-N 中根据外部奖励选择最佳回答（Brown 等，2024；Irvine 等，2023；Levi，2024；Muennighoff 等，2025；Schaeffer 等，2025a；Kazdan 等，2025年；Schaeffer 等，2025b；Chen 等，2025年）。与重复采样不同，先前的序贯缩放方法让模型依次生成解答尝试，在先前尝试的基础上继续生成，并根据之前的结果不断改进每次尝试（Snell 等，2024；Hou 等，2025年；Lee 等，2025年）。基于树的搜索方法（Gandhi 等，2024；Wu 等，2024）提供了序贯缩放与并行缩放相结合的混合方案。例如，蒙特卡洛树搜索（MCTS）（Liu 等，2024；Zhang 等，2023；Zhou 等，2024；Choi 等，2023）和引导式束搜索（Xie 等，2023）。REBASE（Wu 等，2024）采用过程奖励模型，在树搜索过程中平衡利用与剪枝。实证结果表明，REBASE 的表现优于基于采样的方法和 MCTS（Wu 等，2024）。奖励模型在这些推理时缩放方法中发挥着关键作用（Lightman 等，2023；Wang 等，2024a;b；Wu 等，2024；Gandhi 等，2024；Liu 等，2024；Zhang 等，2023；Zhou 等，2024；Choi 等，2023；Xie 等，2023；Xin 等，2024；Ankner 等，2024）。奖励模型通常分为两种：结果奖励模型和过程奖励模型。结果奖励模型（Xin 等，2024；Ankner 等，2024）为完整解答评分，尤其适用于 Best-of-N 选择。相比之下，过程奖励模型（Lightman 等，2023；Wang 等，2024a；Wu 等，2024）评估单个推理步骤，适合用于引导基于树的搜索方法。Tsilivis 等人（2025年）近期分析了强化学习的影响。其他方法也探索了简单的测试时缩放技术（Muennighoff 等，2025年）。

B 记号与渐近约定

我们简要总结本文及正文中使用的主要记号和渐近约定。

度量。 相对于均值目标的泛化（测试）损失为

$$\mathcal{L}_{\text{gen}}(N,\lambda) \coloneqq \mathbb{E}_{\mathbf{x},\mathcal{D}_N}\left[\left(\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*\right)^2\right].$$

pass@k 失败概率为

$$\mathcal{L}_{\text{inf}}(k;N) \coloneqq \mathbb{P}\Big(\text{at least } k \text{ trials all fail under the pass@}k \text{ protocol}\Big),$$

使得 pass@$k = 1 - \mathcal{L}_{\text{inf}}(k;\ N)$。

渐近记号。对于依赖于变量 t （通常为 $t = k$或 $t = N$）的序列或函数 f 和 g ，我们记作：

- $f(t) \sim g(t)$，当 $t \to \infty$时，如果 $\lim_{t\to\infty} f(t)/g(t) = 1$。
- $f(t) = O(g(t))$，当 $t \to \infty$时，如果存在 $C > 0$和 t_0，使得对于所有$t \geq t_0$，$|f(t)| \leq C|g(t)|$。
- $f(t) = \Theta(g(t))$，当 $t \to \infty$时，如果 $f(t) = O(g(t))$，且 $g(t) = O(f(t))$。
- $f(t) \asymp g(t)$。

在所考虑的范围内，对于所有t，如果存在常数 $0 < c_1 \leq c_2 < \infty$使得 $c_1 \leq f(t)/g(t) \leq c_2$，则

对于在相关极限下趋于 0 的量，我们用 $o(1)$表示（该极限总会明确指出，例如 $k \to \infty$ 或 $\delta \to 0$）。当我们写出诸如此类的陈述时

$$f(k) = C\,k^{-\beta}\big(1 + o(1)\big) \quad (k \to \infty),$$

$o(1)$项对其他参数（例如 N、δ）的依赖关系，将根据上下文明确，或予以明确说明。

我们将符号 $\sim$ 保留用于上述意义上的渐近相等，并且不在正文的主要陈述中使用非正式符号“≈”。非正式近似总是与明确的极限（例如 $k \to \infty, \delta \to 0$）相关联，或指向附录中的精确渐近陈述。

局部双对数斜率。对于固定的 k 窗口 $[k_1, k_2]$ 和固定的 N，我们将经验有效推理指数定义为局部双对数斜率

$$\beta_{\text{eff}}(N; [k_1, k_2]) \coloneqq -\frac{\log \mathcal{L}_{\text{inf}}(k_2; N) - \log \mathcal{L}_{\text{inf}}(k_1; N)}{\log k_2 - \log k_1}. \tag{23}$$

在实验中，我们将 $[k_1, k_2]$ 设为一个中心窗口，在该窗口内，$\mathcal{L}_{\text{inf}}(\cdot; N)$ 的双对数曲线近似呈线性；具体窗口不会影响 $\beta_{\text{eff}}(N)$ 的定性行为。

C 带噪标签的高维岭回归

在本附录中，我们重新推导并分析正文所用的带加性标签噪声的高维岭线性回归的泛化误差。尽管这些结果在文献中已有充分研究（例如，Maloney 等人（2022）；Hastie 等人（2020）及其中的参考文献），我们仍给出简洁的推导，以突出其与潜在实例难度（LID）模型之间的联系，并为理解第 4.1 节讨论的训练缩放定律奠定基础。

C.1 模型设定

我们考虑一个线性模型，其中学习者旨在根据 N 个训练样本 $(\mathbf{x}_i, y_i)$估计真实的底层参数向量$\boldsymbol{\theta}^* \in \mathbb{R}^d$ 。输入特征 $\mathbf{x}_i \in \mathbb{R}^d$ 从一个零均值且协方差矩阵为 $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ 的分布中独立同分布地抽取。观测标签 y_i 按照以下方式生成

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta}^* + \eta_i, \tag{24}$$

其中，η_i 是样本 i 的标签噪声。在我们的 LID 模型（定义 3.1）中，η_i 是来自 $\mathcal{N}(0, \sigma_\eta^2/\tau_{\mathbf{x}_i})$ 的一次实现。对于当前推导，我们假设 η_i 是独立同分布的，且具有 $\mathbb{E}[\eta_i] = 0$和 $\mathbb{E}[\eta_i^2] = \sigma_{\text{noise}}^2 < \infty$。如果直接应用 LID，则 σ_{noise}^2 将替换为 $\mathbb{E}[\sigma_\eta^2/\tau_{\mathbf{x}}] = \sigma_\eta^2\mathbb{E}[1/\tau_{\mathbf{x}}]$。

学习者通过最小化以下损失，使用岭回归估计 $\boldsymbol{\theta}^*$

$$L(\hat{\boldsymbol{\theta}}) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\theta}})^2 + \frac{\lambda}{2}\|\hat{\boldsymbol{\theta}}\|_2^2, \tag{25}$$

其中，$\lambda \geq 0$ 是正则化参数。令 $X \in \mathbb{R}^{N\times d}$ 为训练特征矩阵（行$\mathbf{x}_i^\top$），$\mathbf{y} \in \mathbb{R}^N$ 为观测标签向量。其解为

$$\hat{\boldsymbol{\theta}}_\lambda = \Big(\frac{1}{N}X^\top X + \lambda \mathbf{I}_d\Big)^{-1}\frac{1}{N}X^\top \mathbf{y}. \tag{26}$$

C.2 泛化误差

泛化误差（或测试损失）衡量的是在具有真实均值目标$\mathbf{x}_{\text{test}}^{\top}\boldsymbol{\theta}^{*}$的未见测试数据$\mathbf{x}_{\text{test}}$上，预期的平方预测误差:

$$\mathcal{L}_{\text{gen}}(N,\lambda)=\mathbb{E}_{\mathcal{D}_N,\mathbf{x}_{\text{test}}}\left[\left(\mathbf{x}_{\text{test}}^{\top}\hat{\boldsymbol{\theta}}_{\lambda}-\mathbf{x}_{\text{test}}^{\top}\boldsymbol{\theta}^{*}\right)^{2}\right]. \tag{27}$$

令 $\Delta\boldsymbol{\theta}:=\hat{}\boldsymbol{\theta}_{\lambda}-\boldsymbol{\theta}^{*}$。则 $\mathcal{L}_{\text{gen}}=\mathbb{E}[\Delta\boldsymbol{\theta}^{\top}\boldsymbol{\Sigma}\Delta\boldsymbol{\theta}]$，其中期望是对训练数据 $\mathcal{D}_N$ 取的。将 $\mathbf{y}=X\boldsymbol{\theta}^{*}+\vec{\eta}$（其中 $\vec{\eta}$ 是噪声实现 η_i 的向量）代入方程 26，可得

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\lambda}&=\Big(\frac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\frac{1}{N}X^{\top}(X\boldsymbol{\theta}^{*}+\vec{\eta})\\&=\Big(\frac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\left(\frac{1}{N}X^{\top}X\boldsymbol{\theta}^{*}+\frac{1}{N}X^{\top}\vec{\eta}\right).\end{aligned}$$

因此，误差向量为

$$\begin{aligned}\Delta\boldsymbol{\theta}&=\hat{\boldsymbol{\theta}}_{\lambda}-\boldsymbol{\theta}^{*}\\&=\Big(\frac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\left(\frac{1}{N}X^{\top}X\boldsymbol{\theta}^{*}+\frac{1}{N}X^{\top}\vec{\eta}\right)-\boldsymbol{\theta}^{*}\\&=\Big(\frac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\left[\frac{1}{N}X^{\top}\vec{\eta}-\lambda\boldsymbol{\theta}^{*}\right].\end{aligned} \tag{28}$$

假设 $\boldsymbol{\theta}^{*}$ 是固定的（非随机），且 $\mathbb{E}[\vec{\eta}]=\mathbf{0},\mathbb{E}[\vec{\eta}\vec{\eta}^{\top}]=\sigma_{\text{noise}}^{2}\mathbf{I}_N$,

$$\mathbb{E}[\Delta\boldsymbol{\theta}]=-\Big(\tfrac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\lambda\boldsymbol{\theta}^{*}$$

and

$$\begin{aligned}\operatorname{Cov}(\Delta\boldsymbol{\theta})&=\Big(\tfrac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\mathbb{E}\Big[\Big(\tfrac{1}{N}X^{\top}\vec{\eta}\Big)\Big(\tfrac{1}{N}X^{\top}\vec{\eta}\Big)^{\top}\Big]\Big(\tfrac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\\&=\Big(\tfrac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}\frac{\sigma_{\text{noise}}^{2}}{N^{2}}X^{\top}X\Big(\tfrac{1}{N}X^{\top}X+\lambda\mathbf{I}_d\Big)^{-1}.\end{aligned} \tag{29}$$

因此

$$\begin{aligned}\mathcal{L}_{\text{gen}}&=\mathbb{E}[\Delta\boldsymbol{\theta}]^{\top}\boldsymbol{\Sigma}\mathbb{E}[\Delta\boldsymbol{\theta}]+\operatorname{Tr}\Big(\boldsymbol{\Sigma}\,\operatorname{Cov}(\Delta\boldsymbol{\theta})\Big)\\&=\lambda^{2}\,\boldsymbol{\theta}^{*\top}\Big(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I}_d\Big)^{-1}\boldsymbol{\Sigma}\Big(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I}_d\Big)^{-1}\boldsymbol{\theta}^{*}\\&\quad+\sigma_{\text{noise}}^{2}\,\operatorname{Tr}\Big(\boldsymbol{\Sigma}\big(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I}_d\big)^{-1}\hat{\boldsymbol{\Sigma}}\big(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I}_d\big)^{-1}\Big),\end{aligned} \tag{30}$$

其中 $\hat{}\boldsymbol{\Sigma}=\frac{1}{N}X^{\top}X$ 是经验协方差。

C.2.1 欠参数化区域（$N\gg d$）

在欠参数化极限 $N\gg d$ 且 $\lambda\to 0$ 时，偏差项消失，我们得到

$$\mathcal{L}_{\text{gen}}=\sigma_{\text{noise}}^{2}\operatorname{Tr}\Big(\boldsymbol{\Sigma}\,\hat{\boldsymbol{\Sigma}}^{-1}\Big). \tag{31}$$

在正文所讨论的设置中，总体协方差可写为 $\boldsymbol{\Sigma}=\boldsymbol{\Lambda}$，其中 $\boldsymbol{\Lambda}$ 为对角矩阵，其具有如方程 1 所示的衰减幂律谱。正如 Silverstein 和 Bai ˆ（1995）；Couillet 和 Liao（2022）所述，经验协方差矩阵可表示为 $\boldsymbol{\Sigma}=\boldsymbol{\Lambda}\mathbf{W}$，其中 $\mathbf{W}$ 是一个长宽比为 $\kappa=d/N$ 的 Wishart 矩阵。于是

$$\mathcal{L}_{\text{gen}}=\sigma_{\text{noise}}^{2}\operatorname{Tr}\big(\mathbf{W}^{-1}\big). \tag{32}$$

$\mathbf{W}^{-1}$ 的特征值服从逆 Marchenko–Pastur 定律（Couillet 和 Liao，2022年），这意味着

$$\mathbb{E}\big[\operatorname{Tr}(\mathbf{W}^{-1})\big]=\frac{d}{N-d-1}\qquad\text{for } N>d+1. \tag{33}$$

因此，在高样本极限 $N/d \to \infty$ 下，

$$\mathcal{L}_{\text{gen}}(N) \sim \sigma_{\text{noise}}^2 \, \frac{d}{N} \qquad (N \to \infty,\ N \gg d), \tag{34}$$

这在相差一个常数因子的意义下，恢复了公式 10 中使用的 d/N 尾部。

在插值阈值 $N \approx d$附近， $\mathcal{L}_{\text{gen}}$ 会出现熟悉的峰值，构成“双下降”的一侧（Belkin 等人，2019年）。

C.2.2 过参数化区域（$N \ll d$）

在过参数化区域中，通过令岭回归˜参数 $\lambda \to 0$ 在具有缩放关系 $\lambda = \lambda/N$ 的情况下趋于零，可以得到插值解（Bahri 等人，2021年）。在这种缩放下，测试损失分解为

$$\begin{aligned}\mathcal{L}_{\text{gen}} = \frac{\tilde{\lambda}^2}{N^2} \boldsymbol{\theta}^{*\top} \Big(\hat{\boldsymbol{\Sigma}} + \tfrac{\tilde{\lambda}}{N}\mathbf{I}_d\Big)^{-1} \boldsymbol{\Sigma} \Big(\hat{\boldsymbol{\Sigma}} + \tfrac{\tilde{\lambda}}{N}\mathbf{I}_d\Big)^{-1} \boldsymbol{\theta}^* \\ + \sigma_{\text{noise}}^2 \, \operatorname{Tr}\Big(\boldsymbol{\Sigma}\big(\hat{\boldsymbol{\Sigma}} + \tfrac{\tilde{\lambda}}{N}\mathbf{I}_d\big)^{-1} \hat{\boldsymbol{\Sigma}} \big(\hat{\boldsymbol{\Sigma}} + \tfrac{\tilde{\lambda}}{N}\mathbf{I}_d\big)^{-1}\Big).\end{aligned} \tag{35}$$

第一项（偏差）可以通过复制法或随机矩阵方法进行分析，从而得到如下形式的自洽方程:

$$\text{Bias}^2 = \frac{\tilde{\lambda}^2}{N^2} \sum_{i=1}^{d} \frac{\sigma_i^2}{\big(\sigma_i^2 + \frac{\tilde{\lambda}}{N}\big)^2}, \qquad \tilde{\lambda} = \sum_{i=1}^{d} \frac{\frac{\tilde{\lambda}}{N}\sigma_i^2}{\sigma_i^2 + \frac{\tilde{\lambda}}{N}}, \tag{36}$$

其中， $\{\sigma_i^2\}$ 是 $\boldsymbol{\Sigma}$的特征值；详细推导参见 Bordelon 等人（2020年）。对于幂律谱 $\sigma_i^2 \propto i^{-(1+\alpha)}$，求解公式 36 可得如下缩放关系:

$$\mathcal{L}_{\text{gen}}(N) \propto N^{-\alpha} \qquad (N \ll d),$$

除以依赖于谱和教师对齐的乘法常数外（Bartlett 等，2020；Hastie 等，2020）。过参数化区域中的噪声项表现与欠参数化情形类似，只是 d 和 N 的角色互换；它主要在插值阈值附近产生贡献，从而完成双下降图景（Maloney 等，2022）。

在 LID 模型中， σ_{noise}^2 被 $\sigma_\eta^2 \mathbb{E}[1/\tau_{\mathbf{x}}]$取代。因此，假设 3.3（$\beta > 2$）对于这些缩放关系以有限前因子成立至关重要；如果 $\beta \le 2$， $\mathbb{E}[1/\tau_{\mathbf{x}}]$ 就会发散，必须修改标准的岭回归分析。

固定 N下的推理缩放

我们给出在大 k 情况下，对式 (11) 中 pass@k 失败概率积分进行渐近求值的详细步骤

$$\mathcal{L}_{\text{inf}}(k) = \mathbb{E}_{\tau_{\mathbf{x}} \sim \text{Gamma}(\beta/2,1)} \left[[p(\mathbf{x}, \tau_{\mathbf{x}})]^k \right] \approx \mathbb{E}_{\tau_{\mathbf{x}} \sim \text{Gamma}(\beta/2,1)} [e^{-k c_\delta \sqrt{\tau_{\mathbf{x}}}}]. \tag{37}$$

这里， $c_\delta = 2\delta/(\sqrt{2}\pi\sigma_\eta)$ 和 $p(\mathbf{x}\ \tau_{\mathbf{x}})$ 在 $\tau_{\mathbf{x}}$较小时为 $\approx 1 - c_\delta\sqrt{\tau_{\mathbf{x}}}$。伽马分布 Gamma($\beta/2$, 1) 的概率密度函数为 $f(\tau) = \Gamma^{\ 1}{}_{(\beta/2)} \tau^{\beta/2-1} e^{-\tau}$。期望积分为

$$\mathcal{L}_{\text{inf}}(k) = \int_0^\infty e^{-k c_\delta \sqrt{\tau}} \frac{1}{\Gamma(\beta/2)} \tau^{\beta/2-1} e^{-\tau} d\tau. \tag{38}$$

对于较大的 k，因子 $e^{-k c_\delta \sqrt{\tau}}$ 会对任意 $\tau > 0$ 极快衰减，使得积分的主要贡献来自 $\tau = 0$ 附近的区域。在该区域内，Gamma 概率密度函数中的项 $e^{-\tau}$ 近似为 1。因此，我们将积分近似为

$$\mathcal{L}_{\text{inf}}(k) \approx \frac{1}{\Gamma(\beta/2)} \int_0^\infty e^{-k c_\delta \sqrt{\tau}} \tau^{\beta/2-1} d\tau. \tag{39}$$

我们进行变量代换：令 $u = kc_\delta\sqrt{\tau}$。则 $\sqrt{\tau} = u/(kc_\delta)$，这意味着 $\tau = (u/(kc_\delta))^2 = u^2/(kc_\delta)^2$。微分为 $d\tau = \frac{2u}{(kc_\delta)^2}du$。将这些代入积分方程 39

$$\begin{aligned}\int_0^\infty e^{-u}\left(\frac{u^2}{(kc_\delta)^2}\right)^{\beta/2-1}\frac{2u}{(kc_\delta)^2}du &= \int_0^\infty e^{-u}\frac{u^{\beta-2}}{(kc_\delta)^{\beta-2}}\frac{2u}{(kc_\delta)^2}du \\ &= \frac{2}{(kc_\delta)^{\beta-2}(kc_\delta)^2}\int_0^\infty e^{-u}u^{\beta-1}du \\ &= \frac{2}{(kc_\delta)^\beta}\int_0^\infty u^{\beta-1}e^{-u}du.\end{aligned}$$

The 剩余积分就是伽马函数的定义，$\Gamma(\beta)$，其在 $\beta > 0$ 时收敛。

$$\int_0^\infty u^{\beta-1}e^{-u}du = \Gamma(\beta). \tag{40}$$

将其代回 $\mathcal{L}_{\text{inf}}(k)$ 的表达式中

$$\mathcal{L}_{\text{inf}}(k) \approx \frac{1}{\Gamma(\beta/2)}\frac{2\Gamma(\beta)}{(kc_\delta)^\beta} = \left(\frac{2\Gamma(\beta)}{\Gamma(\beta/2)(c_\delta)^\beta}\right)k^{-\beta}. \tag{41}$$

这证实了当 k 较大时，渐近标度 $\mathcal{L}_{\text{inf}}(k) \sim k^{-\beta}$。

D.1 陶伯步骤的假设

我们在以下温和条件下使用标准的拉普拉斯–斯蒂尔切斯陶伯论证：

1. **零点附近的正则变差难度**。潜在精度满足 $\Pr(\tau_x \le t) = t^{\beta/2}L(t)$，当 $t \downarrow 0$ 时成立，其中 L 缓变且在 $(0, t_0]$ 上有界。2. **一致小窗口展开**。对于某个 $c_\delta = \sqrt{2/\pi}\,\delta/\sigma_\eta$ 以及任意紧集 $[\tau_{\text{lo}}, \tau_{\text{hi}}] \subset (0, \infty)$，当 $\delta \downarrow 0$ 时，对 $\tau \in [\tau_{\text{lo}}, \overline{\tau_{\text{hi}}}]$ 一致成立，其中 $c_\delta = \sqrt{2/\pi}\,\delta/\sigma_\eta$ 且 $\eta \sim \mathcal{N}(0, \sigma_\eta^2/\tau)$。3. **模型误差**。$B_N(x) = x^\top(\hat{\theta}_\lambda - \theta^*)$ 是中心次高斯变量，方差为 $\mathrm{Var}[B_N] \asymp \mathcal{L}_{\text{sec}}(N)$，且独立于 τ_x（或为弱相关，使得在 τ_x 条件下仍保持次高斯尾部）。4. **独立试验与完美验证**。给定 (x, τ_x) 和训练集后，k 次比较是独立同分布的；若任一试验落在容差范围内，即判定成功。

$$s(B,\tau) := \Pr\left(|B - \eta| \le \delta \mid \tau\right) = c_\delta\sqrt{\tau}\,\exp\left(-\frac{\tau^2}{2\sigma_\eta^2}\right)(1 + o(1))$$

在条件 (1)–(4) 下，卡拉马塔型陶伯定理给出了方程 14 中的混合定律和方程 12 中的 $k^{-\beta}$尾部，常数如文中所述。

D.2 等价于模型抽样 PASS@k。

考虑另一种方案：抽取 k 个独立同分布的模型提议 $\tilde{y}_j = \hat{y} + \xi_j$，其提议噪声 ξ 具有有界且光滑的密度 f_ξ，并且独立于 τ_x；然后针对固定目标 $y^\star$ 进行验证。记 $B_N(x) = \hat{y} - m(x)$，则每次试验的成功概率为

$$\int_{|u|\le\delta} f_\xi\big(B_N(x) - u\big)\,du \;=\; 2\delta\, f_\xi\big(B_N(x)\big)\,(1 + o(1)) \quad (\delta \downarrow 0). \tag{42}$$

在我们的目标抽样协议中，相应的量为 $2\delta\, f_\eta\big(B_N(x);\ \tau_x\big)$ 其中 $f_\eta(\cdot;\tau_x) = \mathcal{N}(0, \sigma_\eta^2/\tau_x)$。因此，两种协议都可归结为 $(1-p)^k$，其中包含 $p \propto 2\delta$ 倍于位于 $B_N(x)$处的局部密度；由于 LID 中唯一的重尾来自 $f_\eta(\cdot;\tau_x) \propto \sqrt{\tau_x}$ 当 $\tau_x \downarrow 0$ 时，成功概率较小时的行为，以及由此产生的 $k^{-\beta}$ 尾部，在交换模型抽样与目标抽样后都不会改变（除前因子外）。只要 f_ξ 有界，且其自身不会引入依赖于 τ_x 的尾部，并且各次试验在条件上相互独立，这种等价性就成立。

E LID 线性模型中的训练依赖型推理缩放

我们将更详细地推导双尾混合律（定理 4.3）以及依赖于训练的有效指数（推论 4.4）。

E.1 设定

回顾 LID 设定：

- 特征$\mathbf{x} \in \mathbb{R}^d$取自 $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$，其特征值为 $\sigma_j^2 \propto j^{-(1+\alpha)}$。
- 教师是线性的：$m(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}^*$。
- 每个样本都有潜在精度 $\tau_\mathbf{x} \sim \Gamma(\beta/2, 1)$，且与 $\mathbf{x}$ 独立。
- 随机目标为 $Y^*_\mathbf{x} \sim \mathcal{N}\big(m(\mathbf{x}), \sigma_\eta^2/\tau_\mathbf{x}\big)$.^
- 训练阶段观测独立同分布的样本对 $(\mathbf{x}_i, y_i)$，其中 $y_i \sim Y^*_{\mathbf{x}_i}$，并通过 N 个样本拟合岭回归/普通最小二乘法，得到 $\boldsymbol{\theta}_\lambda(N)$。

在测试时，我们评估

- the training loss $\mathcal{L}_{\text{gen}}(N) = \mathbb{E}_\mathbf{x}[(\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*)^2]$,
- the inference loss $\mathcal{L}_{\text{inf}}(k; N)$ under a perfect verifier with tolerance $\delta > 0$,

对于每个测试$\mathbf{x}$，我们抽取 k 个独立同分布的 $y_j \sim Y^*_\mathbf{x}$，并在 $\min_{j \le k} |\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - y_j| \le \delta$ 时判定成功。

E.2 陶伯分析的假设

我们在以下条件下使用标准的拉普拉斯–斯蒂尔切斯陶伯论证：**假设 E.1**（难度尾部和小窗口成功率展开）。

1. **难度的近零尾部**。潜在精度的累积分布函数呈正则变化

$$\Pr(\tau_\mathbf{x} \le t) = t^{\beta/2} L(t) \quad \text{as } t \downarrow 0,$$

其中 L 缓变，且 $\beta > 0$。（当 $\Gamma(\beta/2, 1)$ 时，有 $L(t) \to 1/\Gamma(\beta/2 + 1)$。）

2. **小窗口成功形式**。对于容差 $\delta > 0$ 以及任意紧集 $[\tau_{\text{lo}}, \tau_{\text{hi}}] \subset (0, \infty)$，存在

$$c_\delta = \sqrt{\frac{2}{\pi}} \frac{\delta}{\sigma_\eta}$$

使得

$$s(B, \tau_\mathbf{x}) := \mathbb{P}(|B - \eta| \le \delta \mid \tau_\mathbf{x}) = c_\delta \sqrt{\tau_\mathbf{x}} \, \exp\Big(-\frac{B^2 \tau_\mathbf{x}}{2\sigma_\eta^2}\Big)\big(1 + o(1)\big) \tag{43}$$

当 $\delta \downarrow 0$ 时，关于 $\tau_\mathbf{x} \in [\tau_{\text{lo}}, \tau_{\text{hi}}]$ 一致成立，其中 $\eta \sim \mathcal{N}(0, \sigma_\eta^2/\tau_\mathbf{x})$。

Ass假设 E.2（模型误差分布）。预测偏差 $B_N(\mathbf{x}) = \mathbf{x}^\top(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)$ 满足：

- $\mathbb{E}[B_N(\mathbf{x})] = 0$,
- $B_N(\mathbf{x})$ 是次高斯的，且 $\mathrm{Var}[B_N(\mathbf{x})] \asymp \mathcal{L}_{\text{gen}}(N)$,
- $B_N(\mathbf{x})$ 与 $\tau_\mathbf{x}$ 相互独立（或弱相关，使得对 $\tau_\mathbf{x}$ 进行条件化后仍保有次高斯尾部）。

假设 E.1 是一个温和的正则变差与光滑性条件。对于具有高斯特征和幂律协方差的线性岭回归设置（附录 C），假设 E.2 成立；在此设置下，$B_N(\mathbf{x})$ 近似服从高斯分布，其方差按 $\mathcal{L}_{\text{gen}}(N)$ 缩放。

E.3 单次尝试成功概率的双尾定律

定义单次尝试成功概率

$$S_N(\mathbf{x}, \tau_{\mathbf{x}}) \coloneqq 1 - p(\mathbf{x}, \tau_{\mathbf{x}}) = \mathbb{P}(|B_N(\mathbf{x}) - \eta| \le \delta \mid \tau_{\mathbf{x}}),$$

并令 S_N 表示从测试分布中抽取 $(\mathbf{x}\ \tau_{\mathbf{x}})$ 得到的随机变量。我们首先刻画其分布的小-S_N 尾部。

命题 E.3（的双尾定律 S_N）。在假设 *E.1–E.2* 下，存在常数 $A > 0$、B（N）> 0 以及函数 γ（N）> 0，使得当 $s \downarrow 0$ 时，

$$\mathbb{P}\big(S_N \le s\big) = A\, s^{\beta} + B(N)\, s^{\gamma(N)}\big(1 + o(1)\big), \tag{44}$$

with

$$\gamma(N) = \Theta\Big(\frac{1}{\mathrm{Var}[B_N(\mathbf{x})]}\Big) = \Theta\Big(\frac{1}{\mathcal{L}_{gen}(N)}\Big), \tag{45}$$

and

$$A = \frac{1}{\Gamma(\beta/2 + 1)}\, c_\delta^{-\beta}, \qquad c_\delta = \sqrt{\frac{2}{\pi}}\,\frac{\delta}{\sigma_\eta}.$$

证明概要。当 $\tau_{\mathbf{x}}$ 较小时， $B_N(\mathbf{x})$ 在公式 43 中可忽略，且 $S_N \sim c_\delta \sqrt{\tau_{\mathbf{x}}}$。由于当 $t \downarrow 0$ 时，$\Pr(\tau_{\mathbf{x}} \le t)\ \sim t^{\beta/2}/\Gamma(\beta/2 + 1)$，我们有

$$\mathbb{P}(S_N \le s) \supset \mathbb{P}\Big(\tau_{\mathbf{x}} \le (s/c_\delta)^2\Big) \sim \frac{1}{\Gamma(\beta/2+1)}\Big(\frac{s}{c_\delta}\Big)^{\beta},$$

从而得到 As^β 项。

对于适中的 $\tau_{\mathbf{x}} = \Theta(1)$，较小的 S_N 源于较大的 $|B_N(\mathbf{x})|$。对固定的$\tau_{\mathbf{x}}$，对公式 43 求逆可知，$S_N \le s$ 对应于 $|B_N(\mathbf{x})| \gtrsim \sqrt{(2\sigma_\eta^2/\tau_{\mathbf{x}})\log(c_\delta\sqrt{\tau_{\mathbf{x}}}/s)}$。由于 $B_N(\mathbf{x})$ 是方差为 $\mathrm{Var}[B_N(\mathbf{x})]$ 的次高斯变量，标准尾界可推出

$$\mathbb{P}\big(S_N \le s \mid \tau_{\mathbf{x}}\big) \approx s^{\,c(\tau_{\mathbf{x}})/\mathrm{Var}[B_N(\mathbf{x})]},$$

至缓变因子，其中 $c(\tau_{\mathbf{x}})$ 在 $\tau_{\mathbf{x}} \in [\tau_{\mathrm{lo}}, \tau_{\mathrm{hi}}]$ 上有正的下界和有限的上界。对这一中等范围内的 $\tau_{\mathbf{x}}$ 积分，可得第二项 $B(N)s^{\gamma(N)}$，其中 $\gamma(N) \asymp 1/$方差$[B_N(\mathbf{x})]$。合并各项贡献即可得到公式 44；例如，可参见正则变差情形下关于拉普拉斯–斯蒂尔切斯变换的陶伯定理。□

E.4 PASS@k 的混合律

回顾一下，在我们的独立性假设下，

$$\mathcal{L}_{\mathrm{inf}}(k; N) = \mathbb{E}\big[(1 - S_N)^k\big].$$

这是在 k 处取值的拉普拉斯–斯蒂尔切斯变换。对这类变换应用标准陶伯理论，并结合公式 44 中的尾部形式，可得：

推论 E.4 ($\mathcal{L}_{\mathrm{inf}}(k;\ N)$)。在命题 *E.3* 的假设下，存在常数 $C_1 > 0$ 和 $C_2(N) > 0$，使得当 $k \to \infty$ 时，

$$\mathcal{L}_{inf}(k; N) = C_1\, k^{-\beta} + C_2(N)\, k^{-\gamma(N)}\big(1 + o(1)\big), \tag{46}$$

如公式 *45* 中的 γ（N），以及

$$C_1 = \frac{2\,\Gamma(\beta)}{\Gamma(\beta/2)}\, c_\delta^{-\beta}, \qquad c_\delta = \sqrt{\frac{2}{\pi}}\,\frac{\delta}{\sigma_\eta}.$$

推论 E.4 是正文中定理 4.3 的更详细版本。

E.5 有效双对数斜率

设 $\beta_{\text{eff}}(N;\,[k_1,k_2])$ 为方程 23 中定义的局部双对数斜率。代入方程 46，并在固定 k_2/k_1 的情况下取 $k_1, k_2 \to \infty$，可以看出，方程 46 中占主导地位的项决定了该斜率。这给出：

推论 E.5 (依赖训练的有效指数，精确形式**)**。 在推论 ***E.4*** 的条件下，固定一个窗口 $[k_1, k_2]$，使得对于每个 N， $k^{-\beta}$ 项或 $k^{-\gamma(N)}$ 项在该窗口内占主导。于是，当 $k_1 \to \infty$ 时，

$$\beta_{\text{eff}}(N;[k_1,k_2]) = \min\{\beta, \gamma(N)\} + o(1).$$

特别地，如果 $\mathcal{L}_{gen}(N)$ 随 N 单调递减，那么 $\gamma(N)$ 递增，且 $\beta_{\text{eff}}(N)$ 关于 N 单调不减，并收敛到 β。

推论 E.5 是正文中推论 4.4 的形式化对应，并为经验饱和拟合方程 17 提供了依据。

E.6 摘要

摘要：有限-N 修正与依赖训练的指数

单次试验成功（小窗口）。对于容差 $\delta > 0$ 和偏差 $B_N(\mathbf{x})$，

$$1 - p(\mathbf{x}, \tau_{\mathbf{x}}) = c_\delta \sqrt{\tau_{\mathbf{x}}}\, \exp\Big(-\frac{B_N(\mathbf{x})^2\, \tau_{\mathbf{x}}}{2\sigma_\eta^2}\Big)\big(1 + o(1)\big), \quad c_\delta = \sqrt{\tfrac{2}{\pi}}\, \frac{\delta}{\sigma_\eta}.$$

双尾混合。当 $k \to \infty$时，

$$\mathcal{L}_{\text{inf}}(k; N) = C_1\, k^{-\beta} + C_2(N)\, k^{-\gamma(N)}\big(1 + o(1)\big), \qquad \gamma(N) = \Theta\Big(\tfrac{1}{\mathcal{L}_{\text{gen}}(N)}\Big),$$

其中 $C_1\, \mathrm{g}^{\text{定给出 ex}}$ p式地y 以上。

有效斜率。 在任何由单一尾部主导的固定 k 窗口中，

$$\beta_{\text{eff}}(N) = \min\{\beta, \gamma(N)\} + o(1) \quad \text{as } k_1 \to \infty,$$

因此，随着模型训练得越来越充分， $\beta_{\text{eff}}(N)$ 是 N 的递增函数，并会在 β 处趋于饱和。

F CIFAR-10H LID：设置与评估详情

F.1 实验设置

特征提取。我们将每张图像输入在 ImageNet 上预训练的 ResNet-18，并提取倒数第二层的向量 $\mathbf{z}_i \in \mathbb{R}^d$，其中 $d = 512$。这些冻结特征在我们的 LID 分析中充当 $\mathbf{x}$，我们仅拟合线性头。

来自人工标签的随机目标。 CIFAR-10H 数据集为每张图像提供了多位人工标注者所赋标签的分布。对于每张图像$\mathbf{x}_i$（映射到特征$\mathbf{z}$

$_i$ $m_i = \sum_{c=0}^{9} c \cdot P$ $= c|$ $_i$ $v_i = \sum_{c=0}^{9} c - m_i$ $^2 \cdot$），我们计算这P

些人工标注分布的均值（标签$\mathbf{x}$）和方差（）（标签 $= c|\mathbf{x}_i$）。与我们的 LID 模型的关键联系在于，将线性回归器的目标视为本质上具有随机性。为了控制信噪比，我们引入一个缩放因子 s，以考虑噪声方差的大小；在 LID 设定中，该大小是可控的。随后，学习问题定义在重新缩放后的量上

$$\text{Rescaled features: } \mathbf{z}_i^{\text{scaled}} = \mathbf{z}_i/s, \qquad \text{Rescaled mean target: } m_i^{\text{scaled}} = m_i^{\text{orig}}/s. \tag{47}$$

训练期间，对于每个重新缩放后的特征向量$\mathbf{z}_i^{\text{scaled}}$，目标标签 y_i 从一个高斯分布中采样，该分布以重新缩放后的均值为中心，并使用人工标注方差

$$y_i \sim \mathcal{N}(\text{mean} = m_i^{\text{scaled}}, \text{variance} = v_i + \epsilon_v). \tag{48}$$

其中，m_i^{scaled}是该样本经重新缩放后的人类标签均值，v_i 是人类标签的方差，ϵ_v 是一个较小的常数（例如 10^{-6}），用于确保方差非零。在这里，v_i 的作用类似于合成 LID 模型中的 $1/\tau_{\mathbf{x}_i}$；人类标签方差较高表示该样本“较难”（有效精度较低）。

模型与训练。 我们通过训练线性回归模型 $\hat{}y = \mathbf{z}^{\text{scaled}}$, $_T \hat{}\boldsymbol{\theta}$ 来进行微调，以根据从预训练 ResNet 中提取的重新缩放特征 $\mathbf{z}^{\text{scaled}}$ 预测一个标量值。给定 $\hat{}$ 二次损失和线性模型，我们使用岭回归的解解析地计算最优权重 $\boldsymbol{\theta}$，其中正则化参数 $\lambda_{\text{eff}} = \lambda/s^2$经过了有效缩放

$$\hat{\boldsymbol{\theta}}_\lambda = (N^{-1}\mathbf{Z}_{\text{scaled}}^T\mathbf{Z}_{\text{scaled}} + \lambda_{\text{eff}}\mathbf{I})^{-1}N^{-1}\mathbf{Z}_{\text{scaled}}^T\mathbf{y}, \tag{49}$$

其中，$\mathbf{Z}_{\text{scaled}}$ 是重新缩放后的训练特征矩阵，$\mathbf{y}$ 是采样得到的噪声标签向量（来自公式 (48)）。

评估指标。 泛化损失 $\mathcal{L}_{gen}$。 在留出集上，我们计算 $\hat{}y_{\text{val}} = \mathbf{z}_{\text{val}}^{\text{scaled}\,T}\hat{}\boldsymbol{\theta}$ 与重新缩放后的人类均值 $m_{\text{val}}^{\text{scaled}}$之间的均方误差。（为比较不同的 s，将结果乘以 s^2。）

推理失败 $\mathcal{L}_{inf}(k; N)$。 对于每个验证点，我们抽取 k 个独立同分布的样本 $y_j \sim \mathcal{N}(m_{\text{val}}^{\text{scaled}}, v_{\text{val}} + \epsilon_v)$；若对于任意 j，$\lceil y_{\text{val}} - y_j| \le \delta_{\text{eff}}$，且满足 $\delta_{\text{eff}} = \delta/s$，则判定成功。我们对该集合上的全失败指示变量取平均。这实现了第 4.2 节中的完美验证假设。

G CIFAR-10H 的更多示例

这里，我们提供 CIFAR-10H 数据集中的一些额外示例，以展示数据中固有的随机标签类型。

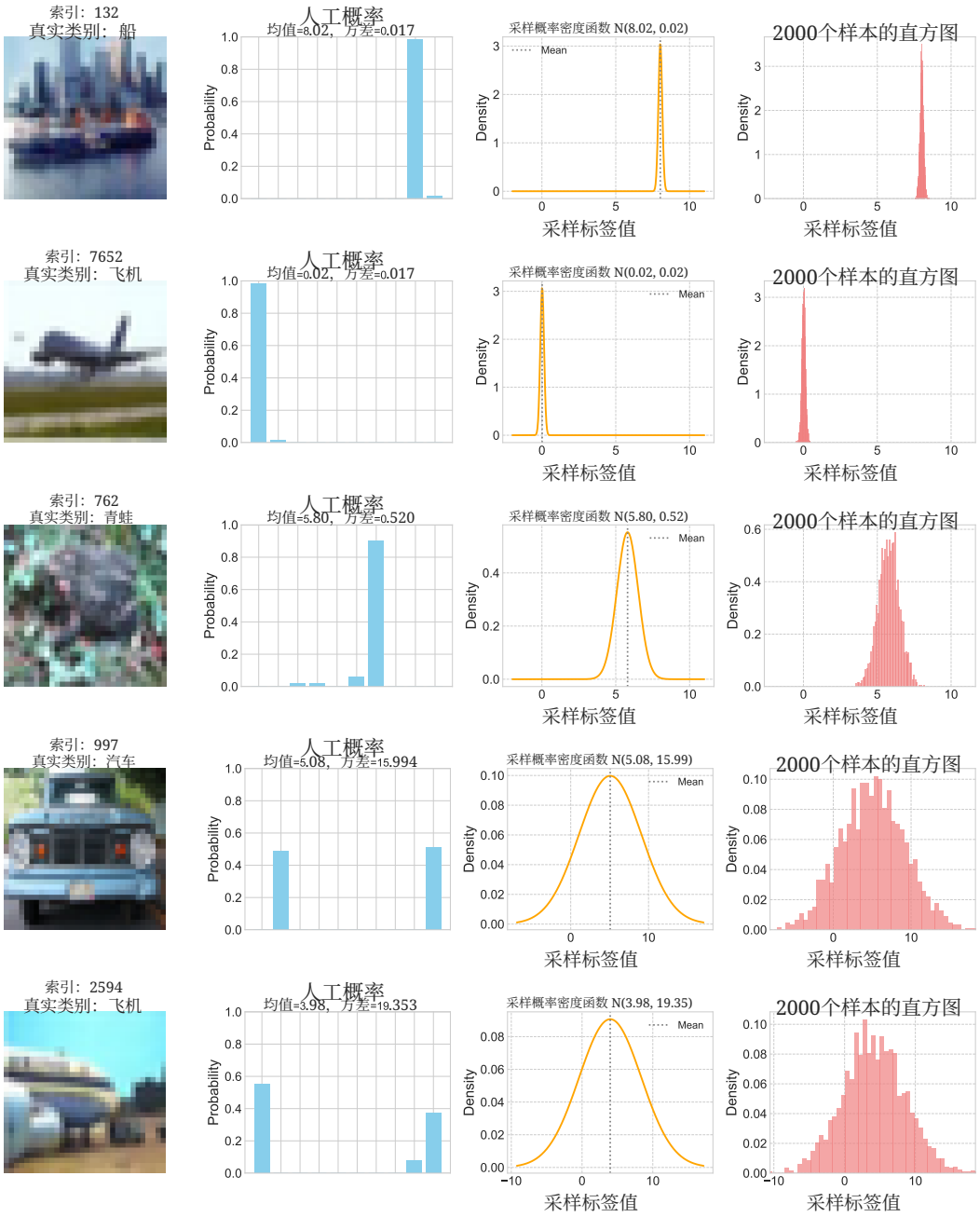

图 5：**CIFAR-10H 中低方差和高方差图像的示例**。 从上到下： 低方差样本更容易预测（更集中在平均预测附近），而高方差样本则难以预测。

H GSM8K 教师–学生蒸馏：设置与评估详情

数据与模型。 教师：Flan-T5-XL。学生：使用 LoRA（秩 r=8,α=16，dropout 0.05）的 Flan-T5-small，训练一个轮次，批量大小为 4，学习率为 5×10^{-5}。我们为每个训练问题采样一个教师解答，并在子集 $N \in \{10, \ldots, 6309\}$（按对数间隔取样）上进行微调。评估采用标准 GSM8K 测试集划分。

训练损失的严格贪心代理指标。 对于每个测试问题，我们解码单个学生输出，并仅在数字出现在明确的"最终答案："栏位中时才进行解析。令 V_N 为有效解析的测试条目集合。容差为 $\delta = 10^{-3}$ 时，我们报告

$$\mathcal{L}_{\text{gen}}^{(\text{greedy})}(N) = \frac{1}{|V_N|} \sum_{x \in V_N} \big(\hat{y}(x) - y^\star(x)\big)^2,$$

通过裁去 $\lceil \hat{} y - y^\star|$ 的最高 1% 进行温莎化，以减小罕见格式或尺度离群值的影响。（我们会跟踪覆盖率 $|V_N|/|\text{test}|$，但不会在正文中绘制。）

推理指标。 我们使用保存的学生样本和一个具有相同容差 δ 的完美$\hat{}$数值验证器，计算 $\mathcal{L}_{\text{inf}}(k;\ N) = 1 - \text{pass@}k$。有效斜率 $\beta_{\text{eff}}(N)$ 是在固定的中心 k 窗口内，$\mathcal{L}_{\text{inf}}(\cdot;\ N)$ 的局部双对数斜率（参见 §4.2）。

解读及其与 LID 的联系。 蒸馏对每个训练输入使用一个带噪的教师样本，因此改进最初可能表现为正确答案所占比例增加，而非点估计大幅偏移。因此，相较于严格的 pass@1 式贪心指标，$\mathcal{L}_{\text{inf}}(k;\ N)$ 更为敏感。随着 $N\hat{}$ 增大，有限-N偏差分量减小，$\mathcal{L}_{\text{inf}}(k;\ N)$ 曲线变陡，而 $\beta_{\text{eff}}(N)$ 则趋近于内在难度指数 β（定理 4.3）。

局限性。 (i) 贪心指标有意设得较为严格（仅接受明确给出的最终答案），因此会低估未以规范形式输出的改进；为此，我们计算仅有效样本均方误差。(ii) 每个训练样本仅使用一个教师样本，并不会直接以教师均值为目标；训练侧均方误差的改善幅度不大，但与理论强调推理时行为的观点一致。(iii) 前因子和饱和程度 N 取决于 LoRA 容量以及教师/学生模型选择；训练与推理指数之间的定性耦合关系则较为稳健。