arXiv:2603.01293v1 [cs.LG] 1 Mar 2026

预训练与后训练推理模型中数据质量与协同效应的理论视角

Adel Javanmard[1,3],Baharan Mirzasoleiman[2,3],和 Vahab Mirrokni[3]

[1]南加州大学[2]加州大学洛杉矶分校[3]Google Research

摘要

大语言模型在海量数据集上进行预训练，随后通过监督微调（SFT）或强化学习（RL）进行指令调优。最佳实践强调大规模、多样化的预训练数据，而后训练的运作方式则有所不同：监督微调依赖于规模更小、质量更高的数据集，而强化学习则更受益于规模，更多的反馈往往比标签质量更为重要。然而，为何预训练与强化学习需要大规模数据集、为何监督微调在较小数据集上表现更优，以及何为高质量监督微调数据，仍不明确。在本文中，我们从理论上分析了针对线性回归的上下文内权重预测任务所训练的 Transformer。我们的分析揭示了若干关键发现：（i）均衡的预训练数据可以诱发潜在能力，这些能力随后在后训练中被激活；（ii）监督微调从对预训练模型具有挑战性的少量样本中学习效果最佳，而过大的 SFT 数据集可能会稀释富含信息的预训练信号。相比之下，强化学习在对预训练模型而言并非过于困难的大规模数据上最为有效。我们通过在大型非线性 Transformer 架构上的实验验证了这些理论洞见。

1 引言

在大规模语言数据集上进行预训练，随后进行后训练，对于解锁并塑造大语言模型（LLMs）的能力至关重要。虽然预训练赋予模型广泛的语言知识和一般性的世界理解，但后训练将这些潜在能力转化为可通过指令可靠引出的可用技能。这种转变通常通过监督微调（SFT）或强化学习（RL）实现：前者训练模型模仿高质量示范，后者利用标量反馈优化模型行为，以改进推理质量和偏好对齐等全局属性。尽管二者处于核心地位，预训练数据与后训练数据之间的相互作用——以及这种相互作用如何决定最终的模型能力——仍未得到充分理解。

在实践中，预训练通常依赖大规模且多样化的数据混合，而后训练则遵循多种方案。例如，OpenAI o1（OpenAI，2024年）和 DeepSeek R1（Guo 等人，2025年）通过对大规模数据集应用强化学习，取得了最先进的推理表现；而 s1（Muennighoff 等人，2025年）则在一小套人工精选的困难且多样的示例上使用监督微调，展现出可比的数学推理表现。最近，Llama 4（Meta，2025年）在难度逐步提高的数据上采用监督微调与强化学习的迭代轮次。然而，预训练数据的哪些特征能够解锁更优的后训练性能，以及要将预训练模型提升至最优性能，对后训练数据的质量和规模有何要求，仍不明确。

在这项工作中，我们通过研究线性回归的上下文内权重预测任务来回答上述问题。该任务的目标是根据输入提示序列预测线性权重向量。此前已有研究使用这一框架分析训练思维链的内在机制（Huang 等人，

2025b；Javanmard 等人，2025）。在这项工作中，我们提出了一种新颖的流程：预训练期间，模型直接进行上下文学习，并输出对权重向量的预测；后训练期间，变换器通过监督微调或强化学习进行思维链推理，并在得出最终的权重向量预测之前生成多个中间步骤。我们在预训练任务和后训练任务的组合上测试该模型。

尽管我们的理论设置捕捉到了**结果监督**（强化学习，即对最终答案给予奖励）与**过程监督**（监督微调，即监督中间步骤）之间的关键区别，但它对涉及采样、优势估计和策略梯度的标准强化学习算法进行了大幅抽象。在这里，我们将强化学习建模为变换器的上下文预测任务上的结果监督回归。这一简化有助于进行清晰的理论分析，但限制了其对大语言模型中完整基于人类反馈的强化学习实现的直接适用性。

我们的分析有助于阐明以下几个问题：

(i) 哪些预训练数据特征有助于模型形成潜在能力，使其能在后训练中得到有效释放?

(ii) 对于给定的预训练模型，什么样的监督微调数据能够有效促进新技能的适应，同时最大限度地减少对预训练期间习得能力的干扰?

(iii) 对于给定的预训练模型，强化学习数据的哪些属性最为关键？强化学习的优化景观与监督微调的优化景观有何不同？强化学习在何种情况下能够取得与监督微调相当的结果?

我们的分析有助于严谨地理解文献中报告的一些经验现象。具体而言，对于我们的上下文设置，分析表明：**(i)** 有效的预训练数据包含来自所有类别且比例均衡的数据混合。这类数据能够诱发潜在能力，并在后训练期间将其激活。**(ii)** 对预训练模型而言，使用少量具有挑战性的示例进行监督微调，最有利于后训练；而更多的监督微调数据反而可能损害性能。**(iii)** 强化学习需要大规模的数据，这些数据应具有信息量，但又不能让预训练模型觉得过于困难。

我们通过在仅含单个线性自注意力（LSA）的变换器上进行线性回归的上下文内权重预测任务，以及在大型非线性 Transformer 架构（即 GPT-2（Radford 等，2019））上进行实验，验证了我们的发现。

2 相关工作

近期工作强调了若干与我们的研究相关的现象。

预训练。对于大语言模型的预训练，通常的做法是使用大规模的语言数据混合。近期研究主要聚焦于数据过滤（Li 等，2024年）、数据选择（Nguyen 等，2024年；Yang 等，2024年）以及混合数据重加权（Xie 等，2023年）。从实证角度看，高质量预训练数据应当规模庞大且具有多样性。这类高质量预训练数据能够诱发潜在能力，这些能力在预训练之后不一定会被观察到，但会在后训练期间被激活（Akter 等，2025年）。

后训练。对于后训练，近期研究主要聚焦于比较监督微调和强化学习（Aminian 等，2025年；Xiong 等；Zhao 等，2025年）这两种后训练方法。从理论上看，监督微调具有模式覆盖特性：通过最小化相对于示范数据的前向KL散度，它鼓励模型为所有合理的响应分配概率质量。相比之下，强化学习（RL）具有模式寻优特性：通过优化奖励（通常在 KL 约束下进行），它将概率集中于高奖励响应，并抑制排名较低的替代响应。因此，监督微调定义了可接受行为的空间，而强化学习则在该空间内选择并强化最受偏好的行为。从实证角度看，监督微调数据应当规模较小且质量较高，即具有挑战性和多样性（Guha 等，2025年；Huang 等，2025年b；Muennighoff 等，2025年），而规模更大的监督微调数据会抵消高质量预训练数据带来的益处（Akter 等，2025年）。相比之下，强化学习受益于规模更大的数据，这些数据仍应具有挑战性，但不应让预训练模型感到过于困难（Meta，2025年；Yue 等，2025年；Zeng 等，2025年）。

然而，预训练数据的某些特征为何能带来更优的后训练性能，SFT 为何能从少量高难度且多样的示例中受益、而更大的数据集反而会降低其效果，以及在强化学习中，数据规模为何比表面质量更重要，这些问题的原因仍不明晰。

我们的理论框架揭开了这些现象背后的奥秘，弥合了经验结果与对数据动态的原则性理解之间的鸿沟。

3 问题设定

我们聚焦于上下文学习（ICL）设定：模型接收一个上下文数据集 $D = \{(x_i, y_i)\}_{i=1}^n$，其中每个 (x_i, y_i) 对都从某个底层分布 P 中独立采样。这里，输入向量 $\{x_i\}_{i=1}^n$ 属于 R^d，对应的标签 $\{y_i\}_{i=1}^n$ 可以是实数（用于回归任务），也可以是二元值，例如 $\{0, 1\}$（用于分类任务）。随后，模型会收到一个新的测试输入 $x_{n+1} \sim P_x$，并需要预测其对应的标签或相应的上下文权重预测器。换言之，上下文学习作用于由输入-输出对组成的序列（称为提示），即 $(x_1, y_1, \ldots, x_n, y_n, x_{n+1})$，且每个提示都可能有其自身的分布。

线性自注意力（lSA）令 Z 为由提示构成的嵌入（具体构造将在后文讨论）。softmax自注意力模块以一个嵌入矩阵为输入，并输出一个大小相同的矩阵，

$$\begin{aligned} &f_{\mathrm{Attn}}(Z; W_K, W_Q, W_V, W_P) \\ &= Z + W_P W_V Z \cdot \mathrm{softmax}\left(\frac{(W_K Z)^\top W_Q Z}{\lambda}\right) \end{aligned}$$

其中 softmax 按列应用。在 线性自注意力（LSA）中，softmax非线性被移除。定义 $W := W_K^\top W_Q$，$V = W_P W_V$和 $\theta = (W, V)$ 后，我们得到：

$$f_{\mathrm{LSA}}(Z; \theta) = Z + VZ \cdot \frac{Z^\top W Z}{\lambda} \tag{3.1}$$

我们将重点关注上下文线性预测器。每个提示的形式为 $P_\tau = (x_{\tau,1}, y_{\tau,1}, \ldots, x_{\tau,n}, y_{\tau,n}, x_{\tau,n+1})$，其中 $y_{\tau,i} = w_\tau^\top x_{\tau,i}$，且 $w_\tau \sim \mathrm{N}(0, I_d)$。

监督微调与结果监督。我们首先介绍使用 k 步思维链推理的结果监督（OS）训练。正如引言中所述，这种表述通过将标准强化学习（RL）——涉及采样、优势估计和策略梯度——建模为对最终答案给予奖励的结果监督回归，简化了标准强化学习，同时仍保留了其与过程监督的监督微调之间的核心区别。

假设给定提示 $P_\tau = (x_{\tau,1}, y_{\tau,1}, \ldots\ldots, x_{\tau,n}, y_{\tau,n})$。我们构造嵌入

$$\hat{Z}_{\tau,0} = \begin{bmatrix} x_{\tau,1} & \ldots & x_{\tau,n} & 0 \\ y_{\tau,1} & \ldots & y_{\tau,n} & 0 \\ 0 & \ldots & 0 & w_{\tau,0} \\ 0 & \ldots & 0 & 1 \end{bmatrix}, \tag{3.2}$$

并迭代定义 $\hat{}Z_{\tau,i+1} = [\hat{}Z_{\tau,i}, f_{\mathrm{LSA}}(\hat{}Z_{\tau,i})_{[:,-1]}]$。我们初始化 $w_{\tau,0} = 0_{d\times 1}$ 并设置 $\hat{}w_{\tau,i+1} := f_{\mathrm{LSA}}(\hat{}Z_{\tau,i})_{[d+2:2d+1,-1]}$。由此得到

$$\hat{Z}_{\tau,i} = \begin{bmatrix} x_{\tau,1} & \ldots & x_{\tau,n} & 0 & * & \ldots & * \\ y_{\tau,1} & \ldots & y_{\tau,n} & 0 & * & \ldots & * \\ 0 & \ldots & 0 & w_{\tau,0} & \hat{w}_{\tau,1} & \ldots & \hat{w}_{\tau,i} \\ 0 & \ldots & 0 & 1 & 1 & \ldots & 1 \end{bmatrix}, \tag{3.3}$$

设 w_τ^* 为提示 P_τ 的真实权重，对于 $\tau \in [B]$。结果监督（OS）损失为

$$\mathcal{L}^{\mathrm{OS}}(V, W) = \frac{1}{2B} \sum_{\tau=1}^{B} \|\hat{w}_{\tau,k} - w_\tau^*\|_{\ell_2}^2, \tag{3.4}$$

即，结果监督（OS）仅惩罚 k 步推理过程的最后一步。

对于监督微调（SFT），我们使用真实思维链（CoT）序列

$$Z_{i,\tau} = \begin{bmatrix} x_1 & \dots & x_n & 0 & * & \dots & * \\ y_1 & \dots & y_n & 0 & * & \dots & * \\ 0 & \dots & 0 & w_{0,\tau} & w_{1,\tau} & \dots & w_{i,\tau} \\ 0 & \dots & 0 & 1 & 1 & \dots & 1 \end{bmatrix}, \tag{3.5}$$

其中$w_{i,\tau} = (1-(1-\eta)^i)w_\tau^*$，且$w_{0,\tau} = 0$ 提供指数收敛的中间目标，速率 η任意但固定。模型被训练用于预测下一个词元$Z_{i+1,\tau}[:,-1] := (0_d, 0, w_{i+1,\tau}, 1)$，给定$Z_{i,\tau}$。在 B 个训练提示上，监督微调损失为

$$\mathcal{L}^{\mathrm{SFT}}(V,W) := \frac{1}{2B}\sum_{\tau=1}^{B}\sum_{i=0}^{k}\left\| f_{\mathrm{LSA}}(Z_{i,\tau})_{[:,-1]} - (0,0,w_{i+1,\tau},1)\right\|_{\ell_2}^2 .$$

流程：预训练、后训练、后测试。 我们的流程有三个由数据协方差区分的阶段：在 Σ_0上预训练，在 $\Sigma = \Sigma_0 + \Delta$ （低秩 Δ）上后测试，以及在选定的中间分布上后训练（稍后讨论，以获得最优后测试性能）。输入 $x \in \mathrm{R}^d$ 始终服从高斯分布。

假设预训练提示无限，（Huang 等人，2025a）的总体分析表明，在适当的初始化下，预训练参数由下式给出：

$$\hat{V}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\Gamma_0^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \hat{W}_0 = \begin{bmatrix} 0 & 0 & I & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.6}$$

其中

$$\Gamma_0 := \left(1+\frac{1}{n}\right)\Sigma_0 + \frac{1}{n}\mathrm{tr}(\Sigma_0)I_d \in \mathbb{R}^{d\times d}, \tag{3.7}$$

其中 n 为提示长度。后训练从 ($\hat{}V_0$、$\hat{}W_0$) 初始化，并通过最小化监督微调损失或 OS 损失来更新 Transformer 权重。

总体情形启发的稀疏结构。 (Huang 等人，2025a) 表明，在总体情形下（$B \to \infty$ 在 d、n之前）使用思维链进行训练（与我们的监督微调损失相对应）会保留初始化时权重中的稀疏性 (3.6)。具体而言，(Huang 等人，2025a) 中的引理 C.2 证明，梯度流轨迹会保留以下稀疏结构：

$$V(t) = \begin{bmatrix} 0 & 0\,0\,0 \\ 0 & 0\,0\,0 \\ V_{31}(t) & 0\,0\,0 \\ 0 & 0\,0\,0 \end{bmatrix}, \quad W(t) = \begin{bmatrix} 0\,0 & W_{13}(t) & 0 \\ 0\,0 & 0 & -1 \\ 0\,0 & 0 & 0 \\ 0\,0 & 0 & 0 \end{bmatrix}, \tag{3.8}$$

其中，$V_{31}(t)$ $W_{13}(t) \in \mathrm{R}^{d\times d}$ 表示时间 t的参数。虽然他们的分析假设数据服从单位协方差高斯分布，且中间权重 $w_{i,\tau}$ 由标准梯度下降得到，但 (Huang 等人，2025a) 中引理 C.2 的证明仅依赖于 $w_\tau^* \sim \mathrm{N}(0, I)$ 的对称性，以及 $w_{i,\tau}$ 是关于 w_τ^*的奇函数这一事实。因此，这一结构性结果可推广到我们的一般协方差和监督序列设定。尽管我们的分析超越了总体情形，这些发现促使我们将变换器模型约束为遵循类似的稀疏模式。全文中，我们使用简写 $\tilde{V}$和 $\tilde{W}$ 分别表示 V和 W的非零块。

4 监督微调损失分析

令 $S_\tau := \frac{1}{n}\sum_{i=1}^{n} x_{i,\tau} x_{i,\tau}^\top$ 为 $\tau \in [B]$的经验特征协方差。我们还定义以下矩阵：

$$\Omega := [w_1^*, \ldots, w_B^*] \in \mathbb{R}^{d\times B}$$
$$\Phi := [S_1 w_1^*, \ldots, S_B w_B^*] \in \mathbb{R}^{d\times B}, \quad M := \Phi\Phi^\top \in \mathbb{R}^{d\times d}$$

下一个定理 刻画了距离初始化最近的监督微调损失极小化器 $(-\Gamma_0^{-1}, I)$.

定理 4.1 定义

$$(\widetilde{V}_\lambda, \widetilde{W}_\lambda) = \arg\min_{(\widetilde{V},\widetilde{W})} \mathcal{L}^{\mathrm{SFT}}(\widetilde{V},\widetilde{W}) + \lambda\left\|\widetilde{V}+\Gamma_0^{-1}\right\|_F^2 + \lambda\left\|\widetilde{W}-I\right\|_F^2 .$$

那么我们有 $\lim_{\lambda\to 0^+}(\widetilde{V}_\lambda, \widetilde{W}_\lambda) = (\widetilde{V}_*, \widetilde{W}_*)$，其中

$$\widetilde{W}_* = I, \quad \widetilde{V}_* = -\eta\Omega\Phi^\dagger - \Gamma_0^{-1}(I - \Phi\Phi^\dagger) \tag{4.1}$$

下一个定理表明，可以通过从 $(-\Gamma_0^{-1}, I)$ 初始化的梯度下降达到解 $(\widetilde{V}_*, \widetilde{W}_*)$，并给出保证收敛的步长条件及其收敛速率。

定理4.2 固定 $W = I$。考虑由梯度下降更新 $\widetilde{V}_{t+1} = \widetilde{V}_t - \gamma\nabla_{\widetilde{V}}\mathcal{L}^{\mathrm{SFT}}(\widetilde{V}_t, I)$ 生成的权重序列 $\{V_t\}_{t\ge 0}$，其初始化为 $\widetilde{V}_0 = -\Gamma_0^{-1}$，步长为常数 $0 < \gamma$。定义 $\rho := 1-\eta$ 和 $c_k := \sum_{i=0}^{k}\rho^{2i} < \frac{1}{1-\rho^2} = \frac{1}{2\eta-\eta^2}$。若 $\gamma < \frac{2B}{c_k\lambda_{\max}(M)}$，则 *GD* 更新以如下收敛速率收敛至 $\widetilde{V}_*$：

$$\left\|\widetilde{V}_t - \widetilde{V}_*\right\|_F \le \alpha^t \|\Gamma_0^{-1} + \widetilde{V}_*\|_F ,$$

$$\alpha := \max\left(\left|1 - \frac{\gamma c_k}{B}\lambda_{\max}(M)\right|, \left|1 - \frac{\gamma c_k}{B}\lambda_{\min}^+(M)\right|\right)$$

其中，$\lambda_{\max}(M)$ 和 $\lambda_{\min}^+(M)$ 分别表示 M 的最大特征值和最小非零特征值。特别地，令 $\gamma = \frac{B}{c_k\lambda_{\max}(M)}$，可得

$$\left\|\widetilde{V}_t - \widetilde{V}_*\right\|_F \le \left(1 - \frac{\lambda_{\min}^+(M)}{\lambda_{\max}(M)}\right)^t \|\Gamma_0^{-1} + \widetilde{V}_*\|_F$$

注释 4.1 注意，由 (4.1) 给出的损失极小化器 $(\widetilde{V}_*, \widetilde{W}_*)$ 取决于 n （提示长度）和 B（提示数量），以及监督权重路径中的步长 η，但不取决于 k （推理路径长度）。然而，如果固定梯度下降步长 $\gamma < \frac{2B}{(2\eta-\eta^2)\lambda_{\max}(M)}$，根据定理*4.2*，较大的 k 意味着较大的 c_k，从而带来更快的收敛速率。

值得推导总体情形下 V_* 的极限，其中 $B\to\infty$，而 n，d 保持不变。

命题 4.3 假设特征按如下方式生成：$x_{i,\tau} \sim \mathsf{N}(0, A)$ 其中$A \in \mathbb{R}^{d\times d}$为半正定矩阵。假设 n和d 固定，但提示数量为 $B\to\infty$。那么 $\widetilde{V}_*$ 将收敛到极限 $\widetilde{V}_\infty$，其定义如下：

$$\widetilde{V}_\infty = -\eta\left(\frac{n+1}{n}A + \frac{\mathrm{tr}(A)}{n}AA^\dagger\right)^\dagger - \Gamma_0^{-1}(I - AA^\dagger) \tag{4.2}$$

5 通过 SFT 进行后训练的数据选择

命题 5.1 考虑一个参数为(V, W)的 *LSA* 模型。我们固定 $W = I$，并假设测试提示的形式为 $P = (x_1, \langle w, x_1\rangle \ldots, x_m, \langle w, x_m\rangle)$。使用 $w_0 = 0$初始化上下文学习后，预测权重为$\hat{w} = -\frac{1}{n}\widetilde{V}XX^\top w^*$，其中 $X = [x_1|\ldots|x_n] \in \mathbb{R}^{d\times n}$。此外，如果 $x_i \sim \mathsf{N}(0, \Sigma)$，则有

$$\begin{aligned}&\mathbb{E}_{X,w^*}[\|\hat{w} - w^*\|^2] = \mathbb{E}_X[\left\|I + \widetilde{V}\widehat{\Sigma}\right\|_F^2] = \\ &= \left\|I + \widetilde{V}\Sigma\right\|_F^2 + \frac{1}{n}\left(\mathrm{tr}(\widetilde{V}\Sigma^2\widetilde{V}^\top) + \mathrm{tr}(\widetilde{V}\Sigma\widetilde{V}^\top)\mathrm{tr}(\Sigma)\right)\end{aligned} \tag{5.1}$$

其中，期望是针对 X 和 $w^* \sim \mathsf{N}(0,\ I_d)$ 中的随机性而言的。

在测试误差（5.1）中，对于较大的提示长度 n，我们关注主导项 $\left\|I + \widetilde{V}\Sigma\right\|_F$。假设 $\widetilde{}$ 后训练特征独立同分布于 $\mathsf{N}(0,\ A)$，其中 $A \geq 0$，则后训练权重 $V_*(A)$ 通过 (4.1) 中的 Φ 依赖于协方差 A。因此，最优数据选择归结为选择能使后测试误差最小化的协方差 A。

5.1 最优数据分配

为分析预训练与后训练之间的相互作用，我们考虑测试时协方差 $\Sigma = \Sigma_0 + \Delta$，其中 Σ_0 表示预训练期间所见的分布，Δ 表示适应任务偏移。现在，我们刻画后训练协方差 A 的选择如何影响不同子空间上的后测试误差。

令 $U = \mathrm{range}(A)$。根据 (4.1)，项 Φ 与值域 U 相同，而在正交补空间 $U^\perp$ 上，权重矩阵 $\tilde{V}_*$ 仅作为预训练逆矩阵 $-\Gamma_0^{-1}$ 起作用。此外，在适应偏移 Δ 的值域之外，测试时协方差 Σ 与预训练协方差 Σ_0 一致。由于根据 (3.7) 中 Γ_0 的定义，$\Gamma_0^{-1}\Sigma_0 \approx I$，当我们使 U 与 $\mathrm{range}(\Delta)$ 对齐时，$U^\perp$ 上的残差误差 $I + \tilde{V}\Sigma$ 将变得可忽略不计。此对齐确保后训练资源专门集中于预训练模型存在不足的子空间。

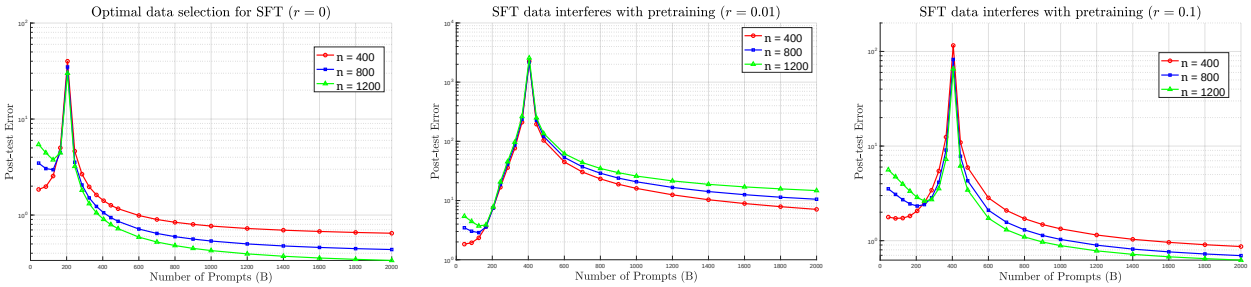
限制在适应子空间 $\mathrm{U} = \mathrm{range}(\Delta)$ 上时，总体极限误差算子表示为：

$$\begin{aligned}&P_U(I + V_\infty\Sigma)P_U \\ &= I - \eta\left(\frac{n+1}{n}A + \frac{\mathrm{tr}(A)}{n}I\right)^{-1}(P_U\Sigma_0P_U + \Delta)\end{aligned}$$

在高维情形（n 较大）下，迹项和 $1/n$ 缩放因子变得次要，这意味着后训练协方差的最优选择近似为 $A \approx \eta($$P_U\Sigma_0P_U + \Delta)$。

与样本难度的联系。 在实践中，后训练常被用于弥补模型中的“缺口”——具体而言，即预训练期间缺失或代表性不足的技能或主题。为刻画此类情形，我们假设预训练协方差 Σ_0 的值域与适应偏移 Δ 的值域具有较小的内积（即二者近乎正交）。因此，$P_U\Sigma_0P_U$ 仅构成 Σ_0的一个很小分量。我们认为，在这些情形下，最有效的策略是选择预训练模型认为“困难”的后训练样本。具体而言，命题 5.1 表明，预训练模型在提示为 $x_{i,\tau} \sim \mathsf{N}(0, A)$的任务上的误差约为 $\mathrm{L}_{pre} \approx \|I - \Gamma_0^{-1}A\|_F^2$。由于 Σ_0 的支撑集在值域(Δ)上很小，算子 Γ_0^{-1}——其本质上充当预训练密度的逆——在该空间上取最大值。因此，协方差由值域(Δ)张成的样本代表了预训练模型置信度最低、残差误差最高的方向。这引出了我们的第一个关键洞见：

> **洞见 1：** 选择对预训练模型而言“困难”的示例（即与适应偏移 Δ 相符的示例），是后训练最有效的策略。



(a) 监督微调的最优数据选择（$r = 0$）(b) 存在干扰时的监督微调数据选择（$r = 0.01$）(c) 存在干扰时的监督微调数据选择（$r = 0.1$）。

图1：后测试误差随提示数量 B 的变化。此处，$d = 400$，$m = 200$ 提示长度不同（n）。预训练协方差为 $\Sigma_0 = \mathrm{diag}(\rho 1_m, 1_{d-m})$，$\Delta = \mathrm{diag}(1_m, 0_{n-m})$。左图表示协方差为 $A = \mathrm{diag}(\eta(\rho+1)1_m, 0_{n-m})$ 的最优监督微调数据分配，其中 $\rho = 0.1$。右图表示监督微调数据分布干扰预训练分布的情况。此处，$A = \mathrm{diag}(\eta(\rho+1)1_m, r1_{n-m})$，其中 $r = 0.01$。

5.2 监督微调中的数据扩展

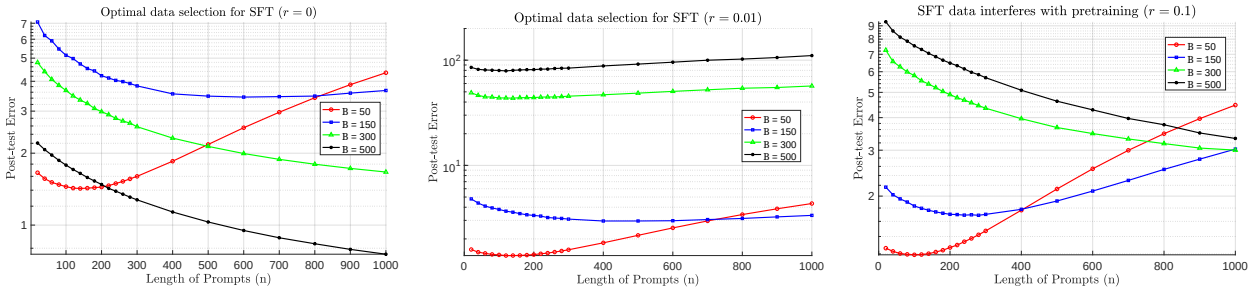
我们通过分析后测试提示 $\mathrm{N}(0, \Sigma)$ 上的期望误差（命题 5.1），研究监督微调数据规模如何影响后训练性能。我们考察该误差如何随监督微调期间的提示数量 B 和提示长度 n 而变化。

我们先介绍实验，然后给出支持这些结论的理论。预训练分布为 $\mathrm{N}(0, \Sigma_0)$，其中 $\Sigma_0 = \mathrm{diag}(\rho 1_m, 0_{n-m})$、$d = 400$ 和 $m = 200$。后测试分布使用 $\Sigma = \Sigma_0 + \Delta$，其中 $\Delta = \mathrm{diag}(1_m, 0_{d-m})$。在后训练期间，数据从 $\mathrm{N}(0, A)$ 中抽取，其中 $A = \mathrm{diag}(\eta(\rho+1)1_m, r1_{n-m})$；前 m 个坐标与 $\eta\Sigma$ 相匹配，其余坐标则使用 r。我们将 ρ 和 r 设为较小值，使得前 m 个方向在预训练中代表不足，并可在后训练期间得到加强。当 $r = 0$ 时，后训练分布与第 5.1 节的最优分配相匹配。然而，非零的 r 会导致后训练数据与预训练数据之间产生干扰，这在实际中很常见。根据式 (4.1)，变换器参数取决于经验协方差的伪逆，因此，非零的 r 越小，干扰越强。

在第一个实验中，我们将提示数量 B 从 50 变化到 2000，对于提示长度 $n \in \{400, 800, 1200\}$，固定 $\rho = 0.1$，并考察干扰水平 $r \in \{0, 0.01, 0.1\}$。图 1 显示，误差呈现双下降，在 $r = 0$ 时于 $B = m$ 处出现超调，在 $r \neq 0$ 时于 $B = d$ 处出现超调。误差起初随 B 增加而下降，随后再次上升，且交叉点随提示长度 n 增加而增大。当干扰较强时，即使在大 B 极限下，误差仍高于最优 B 时的误差值（图 1b）。在第二个实验中，我们将提示长度 n 从 20 变化到 1000，并在 $B \in \{50, 150, 300, 500\}$ 处评估后测试误差。如图 2 所示，误差趋势因 B 的选择而异。在存在干扰且 B 较小时或处于中等水平时，误差先随 n 增加而下降，随后转为单调递增，形成 U 形曲线，这表明存在一个使测试误差最小化的最优提示长度。

这些结果表明，在存在干扰的情况下，增加监督微调数据量——无论是增加提示数量 B 还是延长提示长度 n——反而可能降低性能。关键的权衡在于，更多的监督微调数据有助于模型学习预训练中代表不足的维度，但也会加剧干扰，进而削弱预训练能力。因此，我们的研究结果表明，存在一个能够平衡这些相互竞争效应的最佳数据规模。这也进一步印证了经验上对小规模、高质量数据集的偏好：它们较高的信息密度能够实现有效适应，同时避免过度参数化和干扰带来的灾难性代价。我们将这一观察形式化如下：

> **洞见 2：** 为减轻预训练与后训练之间干扰的影响，应精心筛选 *SFT* 数据集，使其规模相对较小、质量较高。



(a) 监督微调的最优数据选择（$r = 0$） (b) 存在干扰时监督微调的数据选择（$r = 0.01$）。(c) 存在干扰时监督微调的数据选择（$r = 0.1$）。

图2：在与图1相同的设置下，随着提示长度 n 变化，后测试误差的变化情况。

在附录 B 中，我们分析了后测试误差。与我们的实验结果一致，该分析预测，在存在干扰时，测试误差将按 $B \to d$ 发散（$r \neq 0$）；而当 $r = 0$ 时，测试误差将按 $B \to m$ 发散。我们还刻画了以下缩放区间内后测试误差的渐近极限： d， m， 且 $B \to \infty$，同时它们的相对比率保持不变。该分析揭示了不同因素对测试误差行为的定量影响。

6 OS 损失分析

We 首先推导结果监督（OS）损失的一个更直接的刻画 .

命题 6.1 对于在后训练期间进行 k 步思考的 *LSA* 模型，*OS* 损失可写为

$$\mathcal{L}^{\mathrm{OS}}(\widetilde{V},\widetilde{W})=\frac{1}{2B}\sum_{\tau=1}^{B}\left\|\left(I+\sum_{i=0}^{k-1}(\widetilde{V}S_\tau\widetilde{W}+I)^i\widetilde{V}S_\tau\right)w_\tau^*\right\|_{\ell_2}^2$$

参数（$\widetilde{V}$， $\widetilde{W}$）在预训练阶段从（$-\Gamma_0^{-1}$， I）初始化。接下来，我们研究 OS 损失的优化景观，以揭示通过 OS 进行后训练的一些有趣特性，以及它与 SFT 后训练的比较结果。为简化讨论和推导，我们固定 $W = I$，仅通过梯度下降更新 V。不过，我们预计讨论也适用于同时更新两个参数的一般情况，只是推导会更复杂。在实验中，我们更新了所有 Transformer 权重，分析中得到的见解也在实验中得到了验证。

固定 $\widetilde{W} = I$后，OS 损失可简化为：

$$\mathcal{L}^{\mathrm{OS}}(\widetilde{V},I)=\frac{1}{2B}\sum_{\tau=1}^{B}\left\|(I+\widetilde{V}S_\tau)^k w_\tau^*\right\|_{\ell_2}^2 .$$

~令 $M_\tau = I + VS_\tau$。如附录 E 中推导所示，OS 损失关于算子 V的梯度为：

$$\nabla_V\mathcal{L}^{\mathrm{OS}}=\frac{1}{B}\sum_{\tau=1}^{B}\sum_{j=0}^{k-1}(M_\tau^T)^j M_\tau^k w_\tau^*(w_\tau^*)^T(M_\tau^T)^{k-1-j}S_\tau^T .$$

OS 损失中的梯度消失与增长。 梯度包含项 M_τ^k，它起到强缩放因子的作用。在稳定区域（$\rho(M_\tau) < 1$）中，项 M_τ^k会随着链长 k 的增加而以指数速度将梯度收缩至零。在该状态下，模型在该任务上已经稳定，但~梯度消失使得越来越难以将矩阵 V“微调”到最优子空间以进行进一步优化。相反，若 $\rho(M_\tau) > 1$，梯度会随 k呈指数增长。这会形成一个

损失景观中靠近稳定性边界（$\rho \approx 1$）处的陡峭“悬崖”，并且训练需要无穷小的步长，以防止数值发散。

优化景观的尖锐度与曲率。由于 OS 损失实际上是一个 $2k$ 次多项式，海森矩阵 $\nabla^2 \mathbb{L}$ 对算子的谱性质高度敏感。如附录 E 所示，在接近全局最小值且 $M_\tau^k w_\tau^* \approx 0$ 的位置，海森矩阵谱范数 $\lambda_{\max}$ 的变化规律为：

$$\lambda_{\max}(H) \propto \frac{1}{B} \sum_{\tau=1}^{B} k^2 \cdot \rho(M_\tau)^{2k-2} \tag{6.1}$$

这表明，在稳定性边界附近，曲率会随着迭代次数 k 的增加而呈二次增长。如果梯度下降的运行时间不足，模型就会停留在这个高曲率的“悬崖”附近。在这种状态下，有限的 n、B 或测试后评估期间的样本噪声所引起的微小变化，都可能使模型重新进入不稳定区域，导致“过度思考”，即使它在训练期间曾被拉入稳定区域。

> **洞见 3：对样本变化高度敏感**。$\rho \approx 1$ 附近的尖锐曲率表明，除非使用大量数据（n，B）和许多梯度步进行训练，否则结果监督（OS）容易出现不稳定。训练不足会使模型停留在一个“尖锐”的最小值处，轻微的分布偏移就会导致较大的误差。

预训练与泛化。作为 OS 损失初始化的预训练模型，对 OS 的稳定性起着至关重要的作用。考虑一个从测试时协方差 $\Sigma = \Sigma_0 + \Delta$ 中抽取的新任务，其中 Σ_0 表示预训练协方差，Δ 表示适应偏移。在初始化附近，并假设提示长度 n 足够大，使得 $S_\tau \to \Sigma$，学习到的算子 V 主要由先验 $V_0 \approx -\Gamma_0^{-1}$ 决定。因此，我们有 $VS_\tau \approx -\Gamma_0^{-1}(\Sigma_0 + \Delta) \approx -I - \Gamma_0^{-1}\Delta$。由此，转移矩阵变为：

$$M_\tau = I + VS_\tau \approx -\Gamma_0^{-1}\Delta \implies \rho(M_\tau) \approx \rho(\Gamma_0^{-1}\Delta).$$

这种关系揭示了两种截然不同的优化机制，它们取决于预训练分布与适应偏移之间的谱对齐程度：

- **情况 1：渐进式适应（谱对齐）**。当 Γ_0 在 Δ 显著的方向上较大时——这意味着预训练分布有效地覆盖了这种偏移——谱半径 $\rho(M_\tau)$ 会保持较小。在这种机制下，模型在初始化时处于稳定区域（$\rho < 1$）内，因此可以安全地逐步优化模型参数。

- **情形 2：新任务适应（谱失配）**。如果任务涉及新子空间，其中Γ_0 较小而 Δ 较大，那么谱半径就会变大，即 $\rho(\Gamma_0^{-1}\Delta) \gg 1$。模型一开始就处于不稳定区域的深处，因此必须大幅减小步长 η 以维持稳定：

$$\eta < \frac{2}{\lambda_{\max}(H)} \propto \frac{C}{k^2 \rho(M_\tau)^{2k-2}},$$

根据（6.1）。这些观察总结如下：

> **洞见 4：预训练与结果监督的协同作用**。在线合成在提升预训练期间已部分学会的任务的性能方面最为有效。对于新任务，较高的初始谱半径意味着训练过程必须缓慢进行，而且可能不稳定。

训练的实践启示。稳定性要求对结果监督与强化学习施加了若干限制。为确保特征值保持在稳定区域内，必须根据海森矩阵中曲率最大的方向仔细调整学习率。这造成了优化景观中的鲜明反差：不稳定方向迫使步长 η 小到趋近于无穷小，以至于在数据对齐方向上也难以取得实质进展。此外，虽然强化学习不像监督微调那样需要高质量的人工筛选标签，但它需要以海量的数据多样性和规模作为补偿。要克服“平坦”方向上的缓慢进展，需要大量梯度步；而充足的数据规模则能确保模型在广泛的任务中深入稳定区域，从而降低推理时“过度思考”的风险。

7 预训练中的数据多样性与分布平衡

在我们的分析中，预训练模型对后训练模型的影响，在数学上被概括为初始化 $V_0 = -\Gamma_0^{-1}$；根据定义 (3.7)，$\Gamma_0 \approx \Sigma_0$ 即为预训练协方差。命题 5.1 所刻画的后测试误差由乘积 $V\Sigma = V(\Sigma_0 + \Delta)$ 决定；在初始化时，该乘积为 $V\Sigma \approx -I - \Gamma_0^{-1}\Delta$。因此，失衡的预训练分布——其特征是奇异或条件不良的 Γ_0——会严重阻碍模型在新方向上的适应，因为在这些方向上， Γ_0 较小，而 Δ较大。尽管监督微调可以通过监督信号的稳定作用，部分缓解先验不匹配问题，但在线合成与强化学习的优化严格受限于 Γ_0 与 Δ之间的谱对齐。如果 Γ_0 缺乏足够的多样性，即使是新子空间中的微小偏移，也会触发海森矩阵谱范数的指数级增长，其尺度为 $k^2\rho^{2k-2}$。这种谱发散要求步长小到趋近于无穷小，并使模型在训练过程中对样本提示的变化十分敏感。这种不稳定性常表现为推理时的“过度思考”。因此，预训练必须优先保证分布平衡和数据多样性，将其作为实现优化稳定性的关键机制。宽广的谱先验可确保模型初始化于稳定区域（$\rho < 1$），从而有效地将强化学习景观中高曲率的“峭壁”平滑为便于下游适应的平缓区域。

8 实验

在本节中，我们进行实验以验证我们的理论结果。

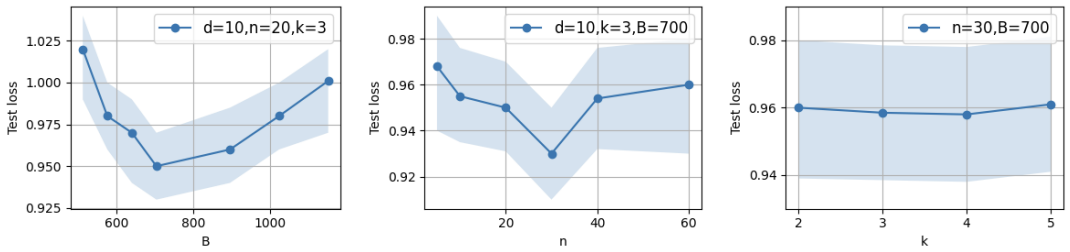
设定。我们在两种设定下进行实验。首先，我们考虑一个具有单个线性自注意力（线性自注意力）的变换器，以确认我们定理的结果。然后，我们考虑大型非线性变换器架构，即GPT-2，以验证我们结论的普适性。

在两组实验中，数据分布均遵循第3节中的上下文内权重预测任务：在预训练中，数据的协方差为 Σ_0；在使用监督微调或在线合成的后测试中，我们有$\Sigma = \Sigma_0 + \Delta$。在后训练期间，我们让模型在返回最终预测权重向量之前输出多个步数，即在每一步 i ，我们将嵌入与 $[0_d, \hat{w}_i, 1]$ 进行拼接，如式(3.3)所示，并将拼接后的嵌入矩阵输入模型。估计的$\hat{w}_k$ 将在 k 步思维链（思维链）后返回。我们报告10次运行的平均结果和误差棒。

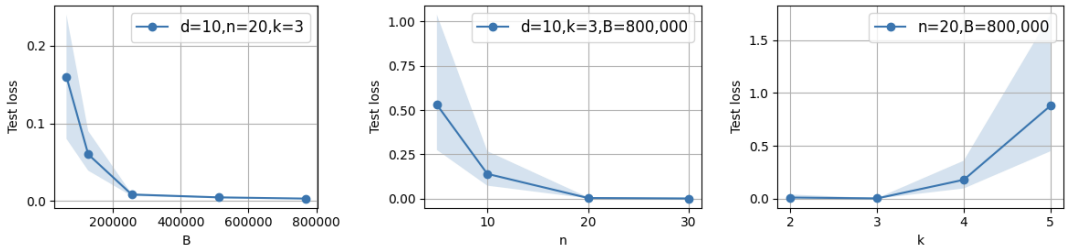
预训练、后训练和测试数据。我们使用 Σ_0 生成预训练数据，其中 $\Sigma_{i,i} = 0.1$ 取值为$i \in \{1, \ldots\ldots, d/5\}$，而 $\Sigma_{i,i} = 1$ 取值为 $i \in \{d/5, \ldots\ldots, d\}$。随后，我们使用 Δ生成的合成数据对变换器进行后训练，其中 Δ 是一个低秩半正定矩阵，且 $\Delta_{i,i} = 1$。测试模型时，我们使用 $\Sigma = \Sigma_0 + \Delta$。

大型非线性 Transformer 架构。我们采用 GPT-2 系列（Radford 等，2019）中的仅解码器 Transformer 架构（Vaswani 等，2017），该架构由 12 层、8 个注意力头和一个 256 维嵌入空间构成。模型总共包含 950 万个参数。该架构以嵌入空间中的向量序列为输入，并预测同一空间中的权重向量。我们按以下方式将此架构应用于形式为 $(x_{\tau,1}, y_{\tau,1}, \cdots, x_{\tau,m}, y_{\tau,m}, w_0, 1)$ 的提示。与 Garg 等（2022）的方法一致，我们通过补零，将每个 $y_{\tau,i}$ 映射到与 $x_{\tau,i}$ 相同的维度；并通过一个（可学习的）线性变换，将 $x_{\tau,i}$，$y_{\tau,i}$ 映射到 Transformer 的潜在嵌入空间。我们将预测的 w_τ 作为模型输出。同样，我们通过另一个（可学习的）线性变换，将模型输出（即 w_τ ）从 Transformer 的潜在嵌入空间映射到一个 d 维向量。训练时，监督微调（SFT）采用批量大小为 64 的批次，共训练 100 步；在线合成（OS）则训练 $12k$ 步。模型首先以 CoT 长度 $k = 8$进行预训练。在训练和测试期间，我们均采用长度为 $k = 3$的思维链。我们采用课程学习（Garg 等，2022）来加快训练。

图 3 (a)-(c) 展示了使用监督微调损失进行后训练时的结果。图 3a、3b 表明，增加样本量（B）或上下文长度（n）起初会降低测试损失，但继续增加样本量或上下文长度则会提高测试损失。图 3c 表明，测试损失相对稳健，不受后训练思维链长度（k）的显著影响。图 3 (d)-(f) 展示了使用 OS 损失进行后训练时的结果。与监督微调不同，我们发现在线合成受益于更大的样本量（B）和上下文长度



(a) 监督微调 (b) 监督微调 (c) 监督微调



(d) 结果监督 (e) 结果监督 (f) 结果监督

图3：GPT-2 实验：(a)–(c) 为采用监督微调进行后训练的测试损失，(d)–(f) 为采用结果监督（OS）进行后训练的测试损失。对于监督微调，存在一个转折点：更大的样本量（B）和上下文长度（n）会损害性能。相比之下，对于 OS，更大的 B，n 会提升性能。

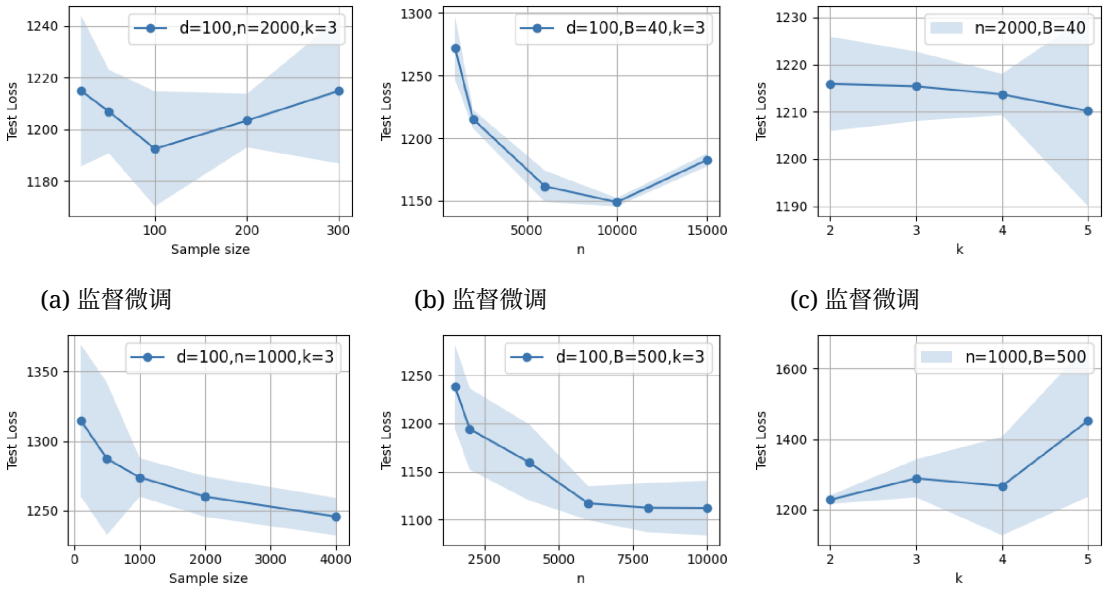
（n）。此外，后训练期间更长的思维链（k）会提高测试损失并降低性能，这证实了第6节中的洞见 4。

线性自注意力（LSA）实验。接下来，我们展示采用单层线性自注意力（LSA）的 Transformer 的实验结果。我们选择词元维度 $d = 100$，并使用 Adam、以学习率 $\eta = 0.001$ 对模型进行 130 个 epoch 的后训练。推理时，我们直接返回最终的预测权重向量，不使用思维链（CoT）；也就是说，测试时我们使用 $k = 1$。

图4(a)–(c) 展示了采用监督微调损失进行后训练的结果。图4a、4b 表明，增大样本量（B）或上下文长度（n）起初会降低测试损失，但继续增大样本量或上下文长度会使测试损失上升。图4c 表明，测试损失相对稳健，不易受后训练思维链（CoT）长度（k）影响。图4(d)–(f) 展示了采用 OS 损失进行后训练的结果。与监督微调不同，图4d、4e 表明 OS 能从更大的样本量（B）和上下文长度（n）中获益；图4f 表明，后训练期间更长的思维链（k）会增加测试损失并降低性能。

9 结论

我们的工作为大语言模型的预训练与后训练协同设计提供了理论与实证框架。均衡的预训练能够形成潜在能力，而通过少量经过精心挑选、与目标偏移相契合的困难样本进行监督微调，最能激活这些能力。扩大监督微调数据规模会引入干扰，侵蚀预训练结构，因此小规模、高质量的数据集更为理想。结果监督与强化学习的优化景观曲率极大且不稳定，因此对数据需求量大，但有助于精炼已部分习得的预训练能力。这些见解为如何协同使用各种方法提供了指导：针对困难样本进行定向监督微调，以高效实现适应；并辅以大规模强化学习（结果监督），实现稳健的技能精炼。



(a) 监督微调 (b) 监督微调 (c) 监督微调

(d) 结果监督 (e) 结果监督 (f) 结果监督

图 4：线性自注意力实验：测试损失分别对应 (a)-(c) 使用监督微调进行后训练，以及 (d)-(f) 使用结果监督（OS）进行后训练。对于监督微调，存在一个转折点，超过该点后，较大的样本量（B）和上下文长度（n）会损害性能。相比之下，对于结果监督（OS），较大的 B，n 会提升性能。

致谢

AJ 的研究部分由美国国家科学基金会奖 DMS-2311024、亚马逊教职研究奖、奥多比教职研究奖以及南加州大学马歇尔商学院的一项 iORB 资助支持。BM 的研究部分由美国国家科学基金会职业奖 2146492、美国国家科学基金会-西蒙斯宇宙起源人工智能研究所（CosmicAI）以及美国国家科学基金会机器学习基础人工智能研究所（IFML）支持。

A 定理与技术引理的证明

A.1 定理 4.1 的证明

~ ~ ~ ~ ~ ~由于 $\lambda \to 0^+$，极小化器 (V_λ, W_λ) 必须收敛到 $L(V, W)$ 的零损失流形上的一点 (V_*, W_*)，该点在弗罗贝尼乌斯范数下距初始化 $(-\Gamma_0^{-1}, I)$ 最近。

我们首先将线性自注意力的动态过程简化为对估计权重 $\hat{}w_i$ 的循环更新。我们有 我们有

$$
\begin{aligned}
f_{\mathrm{LSA}}(Z_i, \theta^*)_{[:,-1]} &= \begin{bmatrix} 0_{d\times 1} \\ 0 \\ \hat{w}_i \\ 1 \end{bmatrix} + VZ_i \cdot \frac{Z_i^\top W Z_{i[:,-1]}}{n} \\
&= \begin{bmatrix} 0_{d\times 1} \\ 0 \\ \hat{w}_i \\ 1 \end{bmatrix} + \frac{1}{n} V Z_i Z_i^\top \begin{bmatrix} \widetilde{W}\hat{w}_i \\ -1 \\ 0 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0_{d\times 1} \\ 0 \\ \hat{w}_i \\ 1 \end{bmatrix} + \frac{1}{n} \begin{bmatrix} 0_{d\times n} & 0_{d\times 1} & 0_{d\times 1} & 0_{d\times 1} \\ 0_{1\times n} & 0 & 0 & 0 \\ \widetilde{V}X & 0_{d\times 1} & 0_{d\times 1} & 0_{d\times 1} \\ 0_{1\times n} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X & 0 & 0 & \dots & 0 \\ y & 0 & 0 & \dots & 0 \\ 0_{d\times n} & w_0 & \hat{w}_1 & \dots & \hat{w}_i \\ 0_{1\times n} & 1 & 1 & \dots & 1 \end{bmatrix}^\top \begin{bmatrix} \widetilde{W} w_0 \\ -1 \\ 0 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0_{d\times 1} \\ 0 \\ \hat{w}_i \\ 1 \end{bmatrix} + \frac{1}{n} \begin{bmatrix} 0_{d\times n} & 0_{d\times 1} & 0_{d\times 1} & 0_{d\times 1} \\ 0_{1\times n} & 0 & 0 & 0 \\ \widetilde{V}X & 0_{d\times 1} & 0_{d\times 1} & 0_{d\times 1} \\ 0_{1\times n} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X^\top \widetilde{W}\hat{w}_i - y^\top \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0_{d\times 1} \\ 0 \\ \hat{w}_i \\ 1 \end{bmatrix} + \frac{1}{n} \begin{bmatrix} 0_{d\times 1} \\ 0 \\ \widetilde{V}XX^\top(\widetilde{W}\hat{w}_i - w^*) \\ 0 \end{bmatrix}.
\end{aligned}
$$

因此，我们得到每个提示权重向量的以下递推关系：

$$
\hat{w}_{i+1,\tau} = \hat{w}_{i,\tau} + \widetilde{V} S_\tau(\widetilde{W}\hat{w}_{i,\tau} - w_\tau^*). \tag{A.1}
$$

现在请注意，在监督微调损失中，我们在每一步都向模型提供思维链真实序列 $(w_{1,\tau}, \dots, w_{i,\tau})$，并计算误差 $\|w_{i+1,\tau} - \hat{}w_{i,\tau}\|_{\ell_2}^2$。令 $\rho = 1 - \eta$。给定 $w_{i,\tau} = (1-\rho^i)w_\tau^*$，我们定义残差 $R_{i,\tau}$，其中$i = 0, .., k$ 以及 $\tau = 1,..., B$，如下所示：

$$
\begin{aligned}
R_{i,\tau} &= w_{i,\tau} + \widetilde{V} S_\tau(\widetilde{W} w_{i,\tau} - w_\tau^*) - w_{i+1,\tau} \\
&= (1-\rho^i)w_\tau^* + \widetilde{V} S_\tau(\widetilde{W}(1-\rho^i)w_\tau^* - w_\tau^*) - (1-\rho^{i+1})w_\tau^* \\
&= \widetilde{V} S_\tau(\widetilde{W} - I)w_\tau^* - \rho^i(\widetilde{V} S_\tau \widetilde{W} + \eta I)w_\tau^*
\end{aligned}
$$

我们通过分析每个块 τ和迭代 $i \in \{0, \dots, k\}$ 的残差 $R_{i,\ \tau}$来刻画该流形。损失函数可写为

$$
\mathcal{L}^{\mathrm{SFT}}(\widetilde{V}, \widetilde{W}) = \frac{1}{2B} \sum_{\tau=1}^{B} \sum_{i=0}^{k} \|R_{i,\tau}\|_{\ell_2}^2 .
$$

~ ~为刻画零损失流形，注意到对于 $L(V, W) = 0$，我们要求$R_{i,\tau} = 0$ 对所有 i成立。由于1与 ρ^i 在 $i \neq 0$ 上线性无关， ρ^i 中多项式的系数必须各自为零：

1. $\widetilde{V} S_\tau(\widetilde{W} - I)w_\tau^* = 0$

$\widetilde{w}_\tau^* = 0 \implies \widetilde{V}S_\tau W w_\tau^* = -\eta w_\tau^*$ 2. $(\widetilde{V}S_\tau W + \eta I)$

将第二个条件代入第一个条件，我们得到：

$$\widetilde{V}S_\tau \widetilde{W} w_\tau^* - \widetilde{V}S_\tau w_\tau^* = 0 \implies -\eta w_\tau^* - \widetilde{V}S_\tau w_\tau^* = 0 \implies \widetilde{V}S_\tau w_\tau^* = -\eta w_\tau^*,$$

对所有 $\tau = 1$、……、B均成立。令 $\Omega = [w_1^*, \ldots, w_B^*]$和 $\Phi = [S_1 w_1^*, \ldots, S_B w_B^*]$。该系统可表示为 $V\Phi = -\eta\Omega$。极限 $\widetilde{V}_*$ 在满足 $\widetilde{V}\Phi = -\eta\Omega$ 的条件下使 $\left\|\widetilde{V} + \Gamma_0^{-1}\right\|_F^2$最小，可通过摩尔-彭罗斯伪逆求解：

$$\widetilde{V}_* = -\eta\Omega\Phi^\dagger - \Gamma_0^{-1}(I - \Phi\Phi^\dagger).$$

项（$I - \Phi\Phi^\dagger$）是到 $\Phi^\top$ 的零空间的正交投影，确保 $\widetilde{V}$沿着数据未张成的方向遵循初始化 $-\Gamma_0^{-1}$。

既然 V_* 已得到刻画，我们接下来证明 $W_* = I$。注意，这一选择 V_*，W_*满足上述梯度条件 (1) 和 (2)。此外，由于惩罚项 $\lambda\left\|\widetilde{W} - I\right\|_F^2$，我们得到$W_* = I$ 作为唯一极小值点。

A.2 定理4.2的证明

设 $\rho = 1 - \eta$ 与 $c_k = \sum_{i=0}^k \rho^{2i}$。给定 $\widetilde{W} = I$ 与 $w_{i,\tau} = (1 - \rho^i)w_\tau^*$，残差为 $R_{i,\tau} = -\rho^i(\widetilde{V}S_\tau + \eta I)w_\tau^*$，损失可写为

$$\mathcal{L}^{\mathrm{SFT}}(\widetilde{V}, I) = \frac{c_k}{2B}\left\|\widetilde{V}\Phi + \eta\Omega\right\|_F^2,$$

其中我们回顾 $\Phi = [S_1 w_1^*, \ldots, S_B w_B^*]$与 $\Omega = [w_1^*, \ldots, w_B^*]$。损失的梯度由下式给出

$$\nabla_{\widetilde{V}}\mathcal{L}_{SF} = \frac{c_k}{B}(\widetilde{V}\Phi + \eta\Omega)\Phi^\top$$

定义 $\Delta_t = V_t - V_*$ 并注意到 $V_*\Phi = -\eta\Omega$，梯度下降更新 $V_{t+1} = V_t - \gamma\nabla_{\widetilde{V}}\mathrm{L}_{SF}(V_t, I)$ 得到：

$$\Delta_{t+1} = \Delta_t\left(I - \frac{\gamma c_k}{B}M\right), \quad M = \Phi\Phi^\top$$

误差范数按$\|\Delta_{t+1}\|_F \le \|\Delta_t\|_F \cdot \left\|I - \frac{\gamma c_k}{B}M\right\|_{\mathrm{op}}$演化，其中 $\|\cdot\|_{\mathrm{op}}$ 表示算子范数。

注意，条件 $\gamma < \frac{2B}{c_k\lambda_{\max}(M)}$ 确保了 $\left\|I - \frac{\gamma c_k}{B}M\right\|_{\mathrm{op}} < 1$，因此梯度下降更新会收敛到$V_*$。具体而言，收缩因子由误差 Δ_t 在数据 Φ张成的子空间中所见到的最极端特征值决定。在 Φ的值域上，收缩因子为

$$\alpha := \max\left(\left|1 - \frac{\gamma c_k}{B}\lambda_{\max}(M)\right|, \left|1 - \frac{\gamma c_k}{B}\lambda_{\min}^+(M)\right|\right)$$

取 $\gamma = \frac{B}{c_k\lambda_{\max}(M)}$后，收敛速率可简化为

$$\alpha = 1 - \frac{\lambda_{\min}^+(M)}{\lambda_{\max}(M)}$$

代入 $\Delta_0 = \widetilde{V}_0 - \widetilde{V}_* = -\Gamma_0^{-1} - \widetilde{V}_*$，即可得到所需的界：

$$\|\widetilde{V}_t - \widetilde{V}_*\|_F \le \left(1 - \frac{\lambda_{\min}^+(M)}{\lambda_{\max}(M)}\right)^t \|\Gamma_0^{-1} + \widetilde{V}_*\|_F,$$

至此完成证明。

A.3 命题 4.3 的证明

~回顾（4.1），V_* 满足系统 $V\Phi = -\eta\Omega$。为求出当 $B \to \infty$时的显式极限，我们分析正规方程：

$$\widetilde{V}\left(\frac{1}{B}\Phi\Phi^\top\right) = -\frac{\eta}{B}\Omega\Phi^\top$$

回顾 $w_\tau^* \sim \mathrm{N}(0, I)$，且 S_τ为来自 $\mathrm{N}(0, A)$ 的 n 个样本的经验协方差。此外，w_τ^*与S_τ 相互独立。

我们有

$$\mathbb{E}\left[\frac{1}{B}\Omega\Phi^\top\right] = \mathbb{E}\left[\frac{1}{B}\sum_{\tau=1}^{B} w_\tau^*(S_\tau w_\tau^*)^\top\right] = \mathbb{E}[w^* w^{*\top} S_\tau^\top]$$

根据独立性以及 $\mathrm{E}[w^* w^{*\top}] = I$ 与 $\mathrm{E}[S_\tau] = A$这一事实，我们得到

$$\mathbb{E}\left[\frac{1}{B}\Omega\Phi^\top\right] = A$$

此外，

$$\mathbb{E}\left[\frac{1}{B}\Phi\Phi^\top\right] = \mathbb{E}\left[\frac{1}{B}\sum_{\tau=1}^{B}(S_\tau w_\tau^*)(S_\tau w_\tau^*)^\top\right] = \mathbb{E}[S_\tau w^* w^{*\top} S_\tau^\top] = \mathbb{E}[S_\tau^2]$$

利用 $S_\tau = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$ 在 $x_i \sim \mathrm{N}(0, A)$ 下的威沙特分布性质–（参见（Javanmard 等人，2025）中的引理 A.2），我们有

$$\mathbb{E}[S_\tau^2] = \frac{n+1}{n}A^2 + \frac{1}{n}\mathrm{tr}(A)A$$

首先考虑 A 可逆的情形。由斯卢茨基定理以及样本协方差的一致性，~当 $B \to \infty$ 时，学习到的算子 V 依概率收敛到：

$$\widetilde{V}_\infty = -\eta A\left(\mathbb{E}[S_\tau^2]\right)^{-1}$$

代入 $\mathrm{E}[S_\tau^2]$ 的显式形式：

$$\widetilde{V}_\infty = -\eta A\left(\frac{n+1}{n}A^2 + \frac{\mathrm{tr}(A)}{n}A\right)^{-1} = -\eta\left(\frac{n+1}{n}A + \frac{\mathrm{tr}(A)}{n}I\right)^{-1}$$

~当 A 为奇异矩阵时，相同的推导在 A 的值域内成立。在 A 的零空间中，V_∞ 保持在其初始化值 $-\Gamma_0^{-1}$。这两种情况可以统一表示如下：

$$\widetilde{V}_\infty = -\eta\left(\frac{n+1}{n}A + \frac{\mathrm{tr}(A)}{n}AA^\dagger\right)^\dagger - \Gamma_0^{-1}(I - AA^\dagger),$$

至此，证明完成。

A.4 命题 5.1 的证明

将递推关系 (A.1) 特化到 $i = 0$ 和 $\widetilde{W} = I$，可得 $\hat{w} = w_0 + \frac{1}{n}\widetilde{V}XX^\top(w_0 - w^*)$。选择初始化 $w_0 = 0$ 后，可得 $\hat{w} = -\frac{1}{n}\widetilde{V}XX^\top w^*$。

令 $\widehat{\Sigma} = \frac{1}{n}XX^\top$，我们有

$$\mathbb{E}[\|\hat{w} - w^*\|_{\ell_2}^2] = \mathbb{E}[\left\|I + \widetilde{V}\widehat{\Sigma}\right\|_F^2] = \left\|I + \widetilde{V}\Sigma\right\|_F^2 + \frac{1}{n}\left(\mathrm{tr}(\widetilde{V}\Sigma^2\widetilde{V}^\top) + \mathrm{tr}(\widetilde{V}\Sigma\widetilde{V}^\top)\mathrm{tr}(\Sigma)\right)$$

其中最后一步由下文的引理 A.1 得出。

引理 A.1 设 $X = [x_1|\dots|x_n]^\top$ 满足 $x_i \sim \mathsf{N}(0,\Sigma)$ 且 $\Sigma \in \mathbb{R}^{d\times d}$。定义 $\widehat{\Sigma} := \frac{1}{n}X^\top X$。那么，对于任意矩阵 $A \in \mathbb{R}^{d\times d}$，有

$$\mathbb{E}[\left\|I + A\widehat{\Sigma}\right\|_F^2] = \|I + A\Sigma\|_F^2 + \frac{1}{n}\left(\mathrm{tr}(A\Sigma^2A^\top) + \mathrm{tr}(A\Sigma A^\top)\mathrm{tr}(\Sigma)\right) \tag{A.2}$$

证明（引理 A.1 的证明）我们写作

$$\mathbb{E}[\left\|I + A\widehat{\Sigma}\right\|_F^2] = d + \mathbb{E}[\left\|A\widehat{\Sigma}\right\|_F^2] - 2\,\mathbb{E}[\mathrm{tr}(A\Sigma)] \tag{A.3}$$

根据 Javanmard 等人，2025（引理 A.2），我们有

$$\mathbb{E}[\widehat{\Sigma}(A^\top A)\widehat{\Sigma})] = \frac{n-1}{n}\Sigma(A^\top A)\Sigma + \frac{1}{n}\left(2\Sigma(A^\top A)\Sigma + \mathrm{tr}(\Sigma A^\top A)\Sigma\right).$$

因此，对等式两边取迹，并交换期望与迹的顺序（因为迹是线性算子），可得

$$\mathbb{E}[\left\|A\widehat{\Sigma}\right\|_F^2] = \frac{n+1}{n}\mathrm{tr}(A\Sigma^2A^\top) + \frac{1}{n}\mathrm{tr}(A\Sigma A^\top)\mathrm{tr}(\Sigma).$$

这里我们还用到了同阶方阵的恒等式 $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ e.

代回 (A.3) 后，我们得到

$$\begin{aligned}\mathbb{E}[\left\|I + A\widehat{\Sigma}\right\|_F^2] &= d + \mathrm{tr}(A^\top\Sigma^2A) - 2\,\mathbb{E}[\mathrm{tr}(A\Sigma)] + \frac{1}{n}\left(\mathrm{tr}(A\Sigma^2A^\top) + \mathrm{tr}(A\Sigma A^\top)\mathrm{tr}(\Sigma)\right)\\ &= \|I + A\Sigma\|_F^2 + \frac{1}{n}\left(\mathrm{tr}(A\Sigma^2A^\top) + 2\mathrm{tr}(A\Sigma A^\top)\mathrm{tr}(\Sigma)\right)\end{aligned}$$

至此，引理得证。 ■

A.5 命题 6.1 的证明

我们首先回顾递推关系 (A.1):

$$\begin{aligned}\hat{w}_{i+1,\tau} &= \hat{w}_{i,\tau} + \widetilde{V}S_\tau(\widetilde{W}\hat{w}_{i,\tau} - w_\tau^*)\\ &= (I + \widetilde{V}S_\tau\widetilde{W})\hat{w}_{i,\tau} - \widetilde{V}S_\tau w_\tau^*\end{aligned}$$

求解该递推关系，我们得到

$$\hat{w}_{k,\tau} = (I + \widetilde{V}S_\tau\widetilde{W})^k\hat{w}_0 - \sum_{i=0}^{k-1}(I + \widetilde{V}S_\tau\widetilde{W})^i\widetilde{V}S_\tau w_\tau^*. \tag{A.4}$$

接下来，利用 $\hat{}w_0 = w_0 = 0$，我们得到

$$\begin{aligned}\mathcal{L}^{\mathrm{OS}}(V,W) &= \frac{1}{2B}\sum_{\tau=1}^{B}\|\hat{w}_{\tau,k} - w_\tau^*\|_{\ell_2}^2\\ &= \frac{1}{2B}\sum_{\tau=1}^{B}\left\|\left(I + \sum_{i=0}^{k-1}(\widetilde{V}S_\tau\widetilde{W} + I)^i\widetilde{V}S_\tau\right)w_\tau^*\right\|_{\ell_2}^2,\end{aligned}$$

由此完成证明。

B 监督微调后训练的渐近分析

我们回顾第4节中的记号。令 $S_\tau := \frac{1}{n}\sum_{i=1}^{n} x_{i,\tau} x_{i,\tau}^\top$ 为$\tau = 1, \ldots, B$的经验特征协方差。我们还定义以下矩阵：

$$\Omega := [w_1^*, \ldots, w_B^*] \in \mathbb{R}^{d\times B}, \quad \Phi := [S_1 w_1^*, \ldots, S_B w_B^*] \in \mathbb{R}^{d\times B}, \tag{B.1}$$

另请回忆，监督微调数据是按 $x_i \sim \mathsf{N}(0, A)$ 生成的，其中 $A = \eta(P_U \Sigma_0 P_U + \Delta) + r P_{U^\perp}$，且 $U = \mathsf{range}(\Delta)$。当 $r = 0$ 时，这对应于第 5.1 节讨论的最优数据分配，而 $r \neq 0$ 则刻画了监督微调数据与预训练模型之间的干扰。

我们考虑预训练协方差 Σ_0 和分布偏移协方差 Δ 的以下特定结构，与第5.2节中的实验类似，即

$$\Sigma_0 = \mathrm{diag}(\rho 1_m, 1_{d-m}), \qquad \Delta = \mathrm{diag}(1_m, 0_{d-m}).$$

在后训练期间，监督微调数据由 $\mathsf{N}(0, A)$ 生成，其中

$$A = \mathrm{diag}(\eta(\rho+1)1_m, r 1_{d-m}), \tag{B.2}$$

后测试分布由协方差 $\Sigma = \Sigma_0 + \Delta$ 给出。值得注意的是，只要这些矩阵的经验谱分布弱收敛到 $\mathsf{R}_{\geq 0}$ 上具有有限二阶矩的概率测度，我们的渐近框架便可推广至任意协方差结构 Γ_0，Δ 和 A。在这一~平均场机制下，学习到的算子 V_* 的宏观行为由数据矩阵和偏移矩阵的谱密度决定，而非由它们具体的坐标级实现决定。

分解 $\widetilde{V}_*$： 从 $\widetilde{V}_* = -\eta\Omega\Phi^\dagger - \Gamma_0^{-1}(I - \Pi_\Phi)$ 出发，并采用投影 $\Pi_\Phi = \Phi\Phi^\dagger$。

令 $\Phi = M + \mathsf{E}$，其中 $M = A\Omega$，而 E 是由于经验特征协方差 S_τ 中的随机性，导致 Φ 相对于其期望 $A\Omega$ 的扰动，适用于 $\tau \in [B]$。利用伪逆的一阶展开：

$$\Pi_\Phi \approx (M + \mathcal{E})(M^\dagger - M^\dagger \mathcal{E} M^\dagger + \ldots)$$

将其展开并仅保留 E 的一阶项：

$$\Pi_\Phi \approx \underbrace{MM^\dagger}_{\Pi_\Omega} + \underbrace{\mathcal{E}M^\dagger - MM^\dagger\mathcal{E}M^\dagger}_{\text{First-order correction}}$$

提取因子 $MM^\dagger$，即可简化修正项：

$$\Pi_\Phi \approx MM^\dagger + (I - MM^\dagger)\mathcal{E}M^\dagger,$$
$$(I - \Pi_\Phi) \approx (I - MM^\dagger) - (I - MM^\dagger)\mathcal{E}M^\dagger.$$

~将展开后的投影代回 V_* 的定义，可得以下一阶近似：

$$\widetilde{V}_* \approx -\eta\Omega(M^\dagger - M^\dagger\mathcal{E}M^\dagger) - \Gamma_0^{-1}\left[(I - MM^\dagger) - (I - MM^\dagger)\mathcal{E}M^\dagger\right] \tag{B.3}$$

现在，将各项分为确定性分量（V_S）和随机分量（V_N）。零阶分量为：

$$V_S = -\eta\Omega M^\dagger - \Gamma_0^{-1}(I - MM^\dagger)$$

一阶项为：

$$V_N = -V_S\mathcal{E}M^\dagger$$

公式 (B.3) 可写为

$$\widetilde{V}_* \approx \widetilde{V} := V_S + V_N. \tag{B.4}$$

接下来，我们利用 (5.1) 刻画测试误差的极限。为方便起见，下面重新写出期望测试误差的刻画：

$$\mathsf{Err}(\widetilde{V}) = \underbrace{\frac{1}{d}\mathbb{E}[\left\|I + \widetilde{V}\Sigma\right\|_F^2]}_{\text{Term I}} + \underbrace{\frac{1}{nd}\mathbb{E}\left[\mathrm{tr}(\widetilde{V}\Sigma^2\widetilde{V}^\mathsf{T}) + \mathrm{tr}(\widetilde{V}\Sigma\widetilde{V}^\mathsf{T})\mathrm{tr}(\Sigma)\right]}_{\text{Term II}} \tag{B.5}$$

其中，期望是对训练数据和测试数据同时取期望。我们还将测试误差按维度 d 进行了归一化。

比例渐近区间。 我们考虑比例渐近区域，其中 $d, m, n, B \to \infty$，其中 n 为提示长度， B 为提示数量。此外，对于某些任意但固定的常数 β, μ_1, γ，有 $B/d \to \beta$, $m/d \to \mu_1$, $d/n \to \gamma$。我们还令 $\mu_2 = 1 - \mu_1$。

记号。 确定性特征协方差 A 是对角矩阵，其块对角元素为 a_1, a_2。测试协方差Σ 也是对角矩阵，其块对角元素为 Σ_1, Σ_2，即

$$\begin{aligned} a_1 &:= \eta(\rho+1), \qquad & a_2 &:= r \\ \Sigma_1 &:= \rho + 1, & \Sigma_2 &:= 1 \end{aligned} \tag{B.6}$$

令 $D_1 = I - \Gamma_0^{-1}\Sigma$ 和 $D_{pre} = \Gamma_0^{-1} - \eta A^{-1}$。二者均为块对角确定性矩阵；在~比例渐近区域中，令 α_k 和 δ_k 分别表示它们在块 $k \in \{1,2\}$ 上的标量值，并令 $\tilde{\delta}_k = \delta_k \Sigma_k$。简单计算可得，对于 $\kappa := \gamma(\mu\rho + 1 - \mu)$，有

$$\begin{aligned} \alpha_1 &= \frac{\kappa - 1}{\rho + \kappa}, \qquad & \alpha_2 &= \frac{\kappa}{\kappa + 1} \\ \delta_1 &= \frac{1-\kappa}{\rho+\kappa}, \qquad & \tilde{\delta}_1 &= \frac{\delta_1}{\Sigma_1} = \frac{1-\kappa}{(\rho+\kappa)(\rho+1)} \\ \delta_2 &= \frac{1}{1+\kappa} - \frac{\eta}{r}, \qquad & \tilde{\delta}_2 &= \frac{\delta_2}{\Sigma_2} = \delta_2 \end{aligned} \tag{B.7}$$

此外， $-\Gamma_0^{-1}$也是一个块对角确定性矩阵；在渐近区间内，我们令 g_k 表示其在块 $k \in \{1, 2\}$上的相应标量值。容易看出，

$$g_1 = -\frac{1}{\rho+\kappa}, \quad g_2 = -\frac{1}{1+\kappa} \tag{B.8}$$

矩阵 $\Sigma_A := \frac{1}{d}(\mathrm{tr}(A)A + A^2)$ 也是块对角矩阵；在比例渐近区间内，其在块 $k \in \{1, 2\}$上的相应标量值收敛于

$$\begin{aligned} s_1 &= (\mu_1 a_1 + \mu_2 a_2)a_1 = (\mu_1\eta(\rho+1) + \mu_2 r)\eta(\rho+1) \\ s_2 &= (\mu_1 a_1 + \mu_2 a_2)a_2 = (\mu_1\eta(\rho+1) + \mu_2 r)r \end{aligned} \tag{B.9}$$

~ ~**定理 B.1** 设 $V = V_S + V_N$ 为(B.4)中的 V_* 的一阶近似。在比例渐近区域下，以下结论成立：

$$\lim_{d\to\infty} \mathsf{Err}(\widetilde{V}) = Bias + \gamma\mathcal{T}_{inv}\mathcal{T}_{var} + \gamma\bar{\Sigma}\mathcal{T}_{var,\Sigma} + \gamma^2\bar{\Sigma}\mathcal{T}_{inv,\Sigma}\mathcal{T}_{var} \tag{B.10}$$

其中，各项根据(B.6)、(B.7)、(B.8)和(B.9)中定义的记号定义如下：

- 偏差：对于 $\beta < 1$，令 q 为以下方程的非负解：

$$\beta = \sum_{k=1}^{2} \mu_k \frac{a_k^2 q}{1 + a_k^2 q} \tag{B.11}$$

并定义 $w_k = \frac{a_k^2 q}{1+a_k^2 q}$，$v_k = w_k(1-w_k)$，其中 $k \in \{1、2\}$，以及

$$T_{12} = \frac{\mu_1\mu_2 v_1 v_2}{\mu_1 v_1 + \mu_2 v_2} \tag{B.12}$$

对于 $\beta \geq 1$，设定 $w_k = 1$、$v_k = 0$ 和 $T_{12} = 0$。于是我们有

$$Bias := \sum_{k=1}^{2} \mu_k \left[\alpha_k^2(1-w_k) + (\alpha_k+\delta_k)^2 w_k\right] + T_{12}(\tilde{\delta}_2^2 - \tilde{\delta}_1^2)(\Sigma_1^2 - \Sigma_2^2) \tag{B.13}$$

●项 $\mathcal{T}_{inv}$ 和 $\mathcal{T}_{inv,\,\Sigma}$ 定义为

$$\mathcal{T}_{inv} = \left\{ q\frac{\sum_{k=1}^2 \mu_k \frac{\Sigma_k^2}{a_k^2} w_k^2}{\sum_{k=1}^2 \mu_k \frac{1}{a_k^2} w_k^2} \right\} \mathbf{1}(\beta<1) + \left\{ \frac{1}{\beta-1} \sum_{k=1}^{2} \mu_k \frac{\Sigma_k^2}{a_k^2} \right\} \mathbf{1}(\beta>1) \tag{B.14}$$

$$\mathcal{T}_{inv,\Sigma} = \left\{ q\frac{\sum_{k=1}^2 \mu_k \frac{\Sigma_k}{a_k^2} w_k^2}{\sum_{k=1}^2 \mu_k \frac{1}{a_k^2} w_k^2} \right\} \mathbf{1}(\beta<1) + \left\{ \frac{1}{\beta-1} \sum_{k=1}^{2} \mu_k \frac{\Sigma_k}{a_k^2} \right\} \mathbf{1}(\beta>1) \tag{B.15}$$

●项 $\mathcal{T}_{var}$、$\mathcal{E}$ 和 $\mathcal{T}_{var,\,\Sigma}$ 定义为

$$\mathcal{T}_{var} = \sum_{k=1}^{2} \mu_k s_k \left[g_k^2(1-w_k) + (g_k+\tilde{\delta}_k)^2 w_k\right] - T_{12}(\tilde{\delta}_1 - \tilde{\delta}_2)(\tilde{\delta}_1 s_1 - \tilde{\delta}_2 s_2) \tag{B.16}$$

$$\mathcal{T}_{var,\Sigma} = \sum_{k=1}^{2} \mu_k \Sigma_k \left[g_k^2(1-w_k) + (g_k+\tilde{\delta}_k)^2 w_k\right] - T_{12}(\tilde{\delta}_1 - \tilde{\delta}_2)(\tilde{\delta}_1 \Sigma_1 - \tilde{\delta}_2 \Sigma_2) \tag{B.17}$$

~接下来，我们将误差的预测渐近极限与数值实验进行比较。回顾一下，V_* 是由式（4.1）给出的监督微调损失~ ~ ~极小化器，V 是其一阶近似，由式（B.4）给出。在图 5 中，我们绘制了误差(V_*)、误差(V)以及理论曲线（B.10）。可以看到，我们的理论预测与（误差(V)）的~ ~模拟结果非常吻合。此外，它对误差(V_*)也给出了较好的近似，并且随着提示长度（n）增加，近似效果更加准确（图 5b 中 $n = 5000$ 的近似效果优于图 5a 中 $n = 1000$ 的近似效果）。

利用定理 B.10，我们证明了渐近误差的若干性质，并表明在存在干扰时，其最小值出现在 $\beta < 1$ 区域。这印证了正文中的洞见 2：监督微调数据集的规模应相对较小。

我们将预测理论误差（即 (B.10) 右侧）记为 $F(\beta)$，视为 β 的函数，因为我们希望了解它随 β 变化时的表现。

命题 B.2 以下结论成立：

(i) $\lim_{\beta\to 1} F(\beta) = \infty$。当 $\beta > 1$ 时，$F(\beta)$ 严格递减。随着 $\beta \to \infty$，它收敛于一个有限的渐近下限：

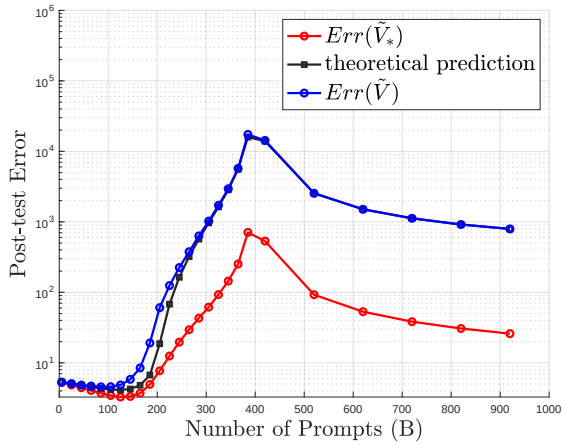
$$F(\uparrow\infty) := \lim_{\beta\to\infty} F(\beta) = \sum_{k=1}^{2} \mu_k(\alpha_k+\delta_k)^2 + \gamma\bar{\Sigma} \sum_{k=1}^{2} \mu_k \Sigma_k (g_k+\tilde{\delta}_k)^2$$

(ii)我们有

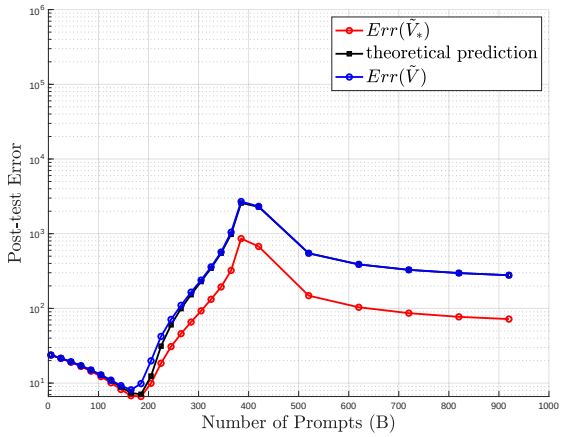
$$F(0) = \sum_{k=1}^{2} \mu_k \alpha_k^2 + \gamma\bar{\Sigma} \sum_{k=1}^{2} \mu_k \Sigma_k g_k^2$$

此外，$F(\uparrow\infty) - F(0)$ 的渐近规模为 $O(1/r^2)$。因此，当 $r > 0$ 足够小时，$F(\uparrow\infty) > F(0)$。这保证了 $F(\beta)$ 的全局最小值严格出现在过参数化区域（$\beta < 1$）。

(iii) 假设 $\mu_1 \geq \frac{\rho^2}{1+\rho^2}$。对于足够小的 r 和 γ，初始导数严格为负（$F'(0) < 0$）。因此，立即引入少量提示会严格降低测试误差。



(a) 提示长度（$n = 1000$）



(b) 提示长度（$n = 5000$）

图 5：渐近误差 $\mathrm{Err}(V)$ 的理论预测、$\mathrm{Err}(V)$ 和 $\mathrm{Err}(V_*)$ 的仿真结果的比较。我们看到理论预测与仿真结果高度吻合。这里，$d = 600$、$m = 300$、$n = 600$（提示规模），$\rho = 0.1$、$\eta = 0.2$、$r = 0.1$（干扰参数）。仿真结果取 10 次实现的平均值。

C 定理 B.1 的证明

C.1 第一项的分析

我们首先分析第一项。我们有

$$\mathbb{E}[\|I + \widetilde{V}\Sigma\|_F^2] = \mathbb{E}[\|I + V_S\Sigma + V_N\Sigma\|_F^2] = \mathbb{E}[\|I + V_S\Sigma\|_F^2] + \mathbb{E}[\|V_N\Sigma\|_F^2]$$

因为在给定 Ω 的条件下，$\mathrm{E} = [(S_1 - A)w_1^*, \ldots, (S_B - A)w_B^*]$ 的均值为零，且与 V_S 相互独立。因此，

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}[\|I + \widetilde{V}\Sigma\|_F^2] = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}[\|I + V_S\Sigma\|_F^2] + \lim_{d\to\infty} \frac{1}{d}\mathbb{E}[\|V_N\Sigma\|_F^2]. \tag{C.1}$$

偏差项分析。 测试误差的确定性部分（偏差）由矩阵$M_S = I + V_S\Sigma$决定。我们首先用正交投影矩阵 $\Pi_M = MM^\dagger$表示 V_S。利用恒等式 $A^{-1}\Pi_M = \Omega M^\dagger$，可得：

$$V_S = -\eta A^{-1}\Pi_M - \Gamma_0^{-1}(I - \Pi_M) \tag{C.2}$$

$$M_S = (I - \Gamma_0^{-1}\Sigma) + (\Gamma_0^{-1} - \eta A^{-1})\Pi_M\Sigma \tag{C.3}$$

令 $D_1 = I - \Gamma_0^{-1}\Sigma$ 和 $D_{pre} = \Gamma_0^{-1} - \eta A^{-1}$。因此，$M_S = D_1 + D_{pre}\Pi_M\Sigma$。注意，$D_1$ 和 D_{pre} 是块对角确定性矩阵；在渐近区间内，它们在块。 $k \in \{1,2\}$上的相应标量值收敛至由（B.7）给出的 α_k 和 δ_k。令 $\delta_k = \delta_k\Sigma_k$。

展开归一化平方弗罗贝尼乌斯范数，可得：

$$\frac{1}{d}\|M_S\|_F^2 = \frac{1}{d}\mathrm{tr}(D_1^2) + \frac{2}{d}\mathrm{tr}(D_1\Sigma\Pi_M D_{pre}) + \frac{1}{d}\mathrm{tr}(D_{pre}^2\Pi_M\Sigma^2\Pi_M) \tag{C.4}$$

接下来我们注意到

$$\frac{1}{d}\mathrm{tr}(D_1^2) = \sum_{k=1}^{2}\frac{d_k}{d}\alpha_k^2 = \sum_{k=1}^{2}\mu_k\alpha_k^2 \tag{C.5}$$

此外，我们还

$$\frac{2}{d}\mathrm{tr}(D_1\Sigma\Pi_M D_{pre}) = \frac{2}{d}\mathrm{tr}(D_{pre}D_1\Sigma\Pi_M) = \frac{2}{d}\sum_{k=1}^{2}\alpha_k\delta_k\mathrm{tr}(\Pi_{kk}) \tag{C.6}$$

为计算二次迹 Quad $:= \frac{1}{d}\mathrm{tr}(D_{pre}^2\Pi_M\Sigma^2\Pi_M)$，我们将投影矩阵分块为 Π_{ij}，其中 $i,\ j \in \{1,\ 2\}$，Π_{11} 的大小为 m，而 $\Pi_{2,\ 2}$ 的大小为 $d-m$。令 $T_{ij} = \frac{1}{d}\mathrm{tr}(\Pi_{ij}\Pi_{ji})$。按块展开迹可得:

$$\mathrm{Quad} = \delta_1^2 T_{11} + \delta_2^2 T_{22} + \left(\tilde{\delta}_2^2\Sigma_1^2 + \tilde{\delta}_1^2\Sigma_2^2\right)T_{12}\,. \tag{C.7}$$

由于 Π_M 是真正的正交投影，$\Pi_M^2 = \Pi_M$。考察该恒等式的对角块可得

$$\Pi_{kk}^2 + \Pi_{kj}\Pi_{jk} = \Pi_{kk}, \tag{C.8}$$

其中 $k \neq j \in \{1,\ 2\}$。

接下来的引理利用随机矩阵理论中的斯蒂尔切斯变换和 Silverstein 方程，刻画 Π_{kk} 的归一化迹的极限。

引理 C.1 令 $w_k := \lim_{d_k\to\infty}\frac{1}{d_k}tr(\Pi_{kk})$。则对于 $\beta < 1$，有:

$$w_k = \frac{a_k^2 q}{1 + a_k^2 q}\,, \tag{C.9}$$

其中 q 为以下方程的非负解:

$$\beta = \sum_{k=1}^{2}\mu_k\frac{a_k^2 q}{1 + a_k^2 q} \tag{C.10}$$

对于 $\beta \geq 1$，我们有 $w_k = 1$。

对 (C.8) 取归一化迹，得到:

$$T_{kk} = \mu_k w_k - T_{12}, \quad k \in \{1, 2\} \tag{C.11}$$

将其代入二次项并化简:

$$\mathrm{Quad} = \mu_1\delta_1^2 w_1 + \mu_2\delta_2^2 w_2 + T_{12}(\tilde{\delta}_2^2 - \tilde{\delta}_1^2)(\Sigma_1^2 - \Sigma_2^2) \tag{C.12}$$

此外，回顾 (C.6) 可得

$$\lim_{d\to\infty}\frac{2}{d}\mathrm{tr}(D_1\Sigma\Pi_M D_{pre}) = \frac{2}{d}\mathrm{tr}(D_1\Sigma\Pi_M D_{pre}) = \lim_{d\to\infty} 2\sum_{k=1}^{2}\left(\frac{d_k}{d}\right)\alpha_k\delta_k\left(\frac{1}{d_k}\mathrm{tr}(\Pi_{kk})\right) = 2\sum_{k=1}^{2}\mu_k\alpha_k\delta_k w_k \tag{C.13}$$

结合 (C.5)、(C.13) 和 (C.12) 给出的线性迹与二次迹，可得完整且严格的偏差:

$$\lim_{d\to\infty}\frac{1}{d}\|M_S\|_F^2 = \sum_{k=1}^{2}\mu_k\left[\alpha_k^2(1 - w_k) + (\alpha_k + \delta_k)^2 w_k\right] + T_{12}(\tilde{\delta}_2^2 - \tilde{\delta}_1^2)(\Sigma_1^2 - \Sigma_2^2)\,. \tag{C.14}$$

I在下一个引理中，我们刻画 T_{12}，从而完成对偏差项的分析。

引理 C.2 设 Π_M 是投影到 $M = A\Omega$ 的列空间上的正交投影矩阵，其中$\Omega \in \mathbb{R}^{d\times B}$的元素独立同分布，方差为$1/d$；$A$ 是一个确定性分块对角矩阵，其块维数为$d_k = \mu_k d$，对应的特征值平方为 a_k^2，其中 $k \in \{1, 2\}$。设 Π_{ij}表示 Π_M 的子块。当

$d, B \to \infty$ 满足 $B/d \to \beta$ 时，归一化的跨子空间泄漏迹 $T_{12} = \lim_{d\to\infty} \frac{1}{d}\mathrm{tr}(\Pi_{12}\Pi_{21})$ 几乎必然为：

$$T_{12} = \frac{\mu_1\mu_2 v_1 v_2}{\mu_1 v_1 + \mu_2 v_2} \tag{C.15}$$

其中 $v_k = w_k(1 - w_k)$ 是块 k 上投影的方差因子，$w_k = \frac{a_k^2 q}{1+a_k^2 q}$ 是由不动点根 q 定义的斯蒂尔切斯权重。

噪声项分析。我们将维度比记为 $\mu_1 = m/d$ 和 $\gamma = d/n$。作用于测试协方差的噪声算子严格定义为 $V_N\Sigma = -V_S\mathbf{E}M^\dagger\Sigma$。我们寻求归一化期望平方弗罗贝尼乌斯范数的极限：

$$\frac{1}{d}\mathbb{E}_{\mathcal{E}}\left[\|V_N\Sigma\|_F^2\right] = \frac{1}{d}\mathbb{E}_{\mathcal{E}}\left[\mathrm{tr}\Big(V_S\mathcal{E}M^\dagger\Sigma^2(M^\dagger)^T\mathcal{E}^T V_S^T\Big)\right] \tag{C.16}$$

令 $Q = M^\dagger\Sigma^2(M\dagger)^T$。由于 $M = \mathbf{E}[\Phi]$ 是确定性的，Q 相对于噪声实现 $\mathbf{E}$ 为常数。

令 $\varepsilon_\tau = (S_\tau - A)w_\tau^*$ 为 $\mathbf{E}$ 的第 τ 列。我们首先在给定权重矩阵 Ω 的条件下，计算关于特征样本 $x_{i,\ \tau}$ 的期望。由于特征样本在不同权重 τ 之间相互独立，$\mathbf{E}$ 的各列相互独立且均值为零：

$$\mathbb{E}[\varepsilon_\tau\varepsilon_\gamma^\top \mid \Omega] = 0 \quad \text{for } \tau \neq \gamma$$

对于对角项，我们使用威沙特二次型协方差的标准恒等式。对于任意固定向量 u 和 $S \sim W_d(n,\ \frac{1}{n}A)$：

$$\mathbb{E}[(S-A)uu^\top(S-A)] = \frac{1}{n}\left((u^\top Au)A + Auu^\top A\right)$$

对 Q 的各个元素求和：

$$\mathbb{E}_{\mathcal{E}|\Omega}[\mathcal{E}Q\mathcal{E}^\top] = \sum_{\tau\neq\gamma} Q_{\tau\gamma}\mathbb{E}[\varepsilon_\tau\varepsilon_\gamma^\top \mid \Omega] = \frac{1}{n}\sum_{\tau=1}^{B} Q_{\tau\tau}\left((w_\tau^{*\top}Aw_\tau^*)A + Aw_\tau^* w_\tau^{*\top}A\right)$$

现在我们对 Ω 取期望。根据高斯分布的旋转不变性，任何在正交变换下具有等变性的 Ω 函数，其期望必定是各向同性的。特别地，项 $Z = \mathbf{E}\left[\sum_\tau Q_{\tau\tau}w_\tau^* w_\tau^{*\top}\right]$ 必须满足 $Z = cI_d$。取迹：

$$cd = \mathbb{E}\left[\sum_{\tau=1}^{B} Q_{\tau\tau}\|w_\tau^*\|^2\right] = d\mathbb{E}[\mathrm{tr}(Q)] \implies c = \mathbb{E}[\mathrm{tr}(Q)]$$

在高维极限下，权重范数二次型（$w_\tau^{*\top}Aw_\tau^*$）与核对角线 $Q_{\tau\tau}$ 之间的相关性消失，由此得到

$$\mathbb{E}\left[\sum_{\tau=1}^{B} Q_{\tau\tau}(w_\tau^{*\top}Aw_\tau^*)\right] = \mathbb{E}[\mathrm{tr}(Q)]\mathrm{tr}(A)$$

综合以上结果，得到：

$$\mathbb{E}[\mathcal{E}Q\mathcal{E}^\top] = \frac{\mathbb{E}[\mathrm{tr}(Q)]}{n}\left(\mathrm{tr}(A)A + A^2\right) \tag{C.17}$$

我们令 $\Sigma_A := \frac{1}{d}(\mathrm{tr}(A)A + A^2)$。将上述恒等式代入 (C.16)，得到：

$$\frac{1}{d}\mathbb{E}_{\mathcal{E}}\left[\|V_N\Sigma\|_F^2\right] = \frac{1}{d}\mathrm{tr}\Big(V_S\left[\gamma\mathrm{tr}(Q)\Sigma_A\right]V_S^T\Big) \tag{C.18}$$

由于 $\mathrm{tr}(Q)$ 是一个标量，它可以完全从矩阵乘积中提出。利用性质 $(M\dagger)^T M^\dagger = (MM^T)^\dagger$，期望严格地分解为两个独立的归一化迹泛函的乘积：

$$\frac{1}{d}\mathbb{E}[\|V_N \Sigma\|_F^2] = \gamma \underbrace{\left[\mathrm{tr}(\Sigma^2 (MM^T)^\dagger)\right]}_{\mathcal{T}_{inv}} \cdot \underbrace{\left[\frac{1}{d}\mathrm{tr}(V_S \Sigma_A V_S^T)\right]}_{\mathcal{T}_{var}} \tag{C.19}$$

伪逆迹的推导（T_{inv}）。我们必须计算目标迹 $\mathrm{T}_{inv} = \mathrm{tr}(\Sigma^2 (MM^T)^\dagger)$，它决定了方差泛函。由于特征矩阵 $M = A\Omega$ 由 $\Omega \sim \mathrm{N}(0, 1)$ 的独立同分布元素构成，未缩放矩阵 $G = MM^T$的特征值按 $O(d)$ 的尺度增长。为严格应用斯蒂尔切斯变换，我们定义归一化矩阵 $\hat{G} = \frac{1}{d}G$，其特征值为 $O(1)$。目标迹的缩放关系为：

$$\mathcal{T}_{inv} = \mathrm{tr}\big(\Sigma^2 (d\hat{G})^\dagger\big) = \frac{1}{d}\mathrm{tr}\big(\Sigma^2 \hat{G}^\dagger\big) \tag{C.20}$$

● **过参数化区间**（$\beta < 1$）。由于 $\hat{G}$ 在过参数化区间（$B < d$）内严格奇异，直接求逆是无效的。为了对任意纵横比 β严格计算该迹，我们引入严格为正的正则化参数 $z > 0$，并定义扰动后的预解式：

$$R(z,t) = (\hat{G} + tA\Sigma^2 A + zI_d)^{-1} \tag{C.21}$$

令 $m(z, t) = \frac{1}{d}\mathrm{tr}(R(z, t))$ 为其归一化迹。由于 $z > 0$， $R(z, t)$ 对所有 B 都是无条件可逆且有界的。对 $m(z, t)$ 关于连续扰动 t 求导，并在 $t = 0$ 处取值得：

$$\frac{\partial}{\partial t} m(z,t)\Big|_{t=0} = -\frac{1}{d}\mathrm{tr}\big(R(z,0)(A\Sigma^2 A)R(z,0)\big) \tag{C.22}$$

由于 $R(z, 0)$、 A 和 Σ 都是定义良好的有限维 $d \times d$ 矩阵，因此我们可以对迹进行循环置换。我们将最右侧的 $R(z, 0)$ 移到左侧，并利用对角矩阵 A 和 Σ^2 可交换这一事实（$A\Sigma^2 A = \Sigma^2 A^2$）：

$$\frac{\partial}{\partial t} m(z,t)\Big|_{t=0} = -\frac{1}{d}\mathrm{tr}\big((A\Sigma^2 A)R(z,0)^2\big) = -\frac{1}{d}\mathrm{tr}\big(\Sigma^2 A^2 R(z,0)^2\big) \tag{C.23}$$

现在，我们将目标斯蒂尔切斯导数 $q'(0)$ 定义为该正则化导数在 $z \to 0^+$ 时的极限。通过将算子极限 $\lim_{z\to 0^+} R(z, 0)^2 \equiv (\hat{G}^\dagger)^2$ 严格定义在非零子空间上，它在所有参数化机制下都等价于目标方差迹 T_{inv}：

$$-q'(0) = \lim_{z\to 0^+} \frac{1}{d}\mathrm{tr}\big(\Sigma^2 A^2 R(z,0)^2\big) \equiv \frac{1}{d}\mathrm{tr}\big(\Sigma^2 A^2 (\hat{G}^\dagger)^2\big) = \mathcal{T}_{inv} \tag{C.24}$$

为了解析求得 $q'(0)$，我们对扰动后预解式的不动点方程求导。扰动后的确定性包络的特征值为 $a_k^2(1 + t\Sigma_k^2)$。利用 Silverstein 方程，可得以下不动点方程：

$$\beta = \sum_{k=1}^{2} \mu_k \frac{a_k^2 (1 + t\Sigma_k^2) q(t)}{1 + a_k^2 (1 + t\Sigma_k^2) q(t)} \tag{C.25}$$

对等式两边关于 t 求导，并在 $t = 0$ 处代入（其中 $q(0) = q$），得到：

$$0 = \sum_{k=1}^{2} \mu_k \frac{a_k^2 \Sigma_k^2 q + a_k^2 q'(0)}{(1 + a_k^2 q)^2} \tag{C.26}$$

分离各项并注意到有效块权重为 $w_k = \frac{a_k^2 q}{1 + a_k^2 q}$，我们得到代数恒等式 $\frac{a_k^2}{(1+a_k^2 q)^2} = \frac{w_k^2}{a_k^2 q^2}$。将其代入微分方程，得到：

$$-q'(0) \sum_{k=1}^{2} \mu_k \frac{w_k^2}{a_k^2 q^2} = q \sum_{k=1}^{2} \mu_k \Sigma_k^2 \frac{w_k^2}{a_k^2 q^2} \tag{C.27}$$

两边乘以 q^2 并分离出 $-q'(0)$，即可得到精确的闭式极限：

$$\mathcal{T}_{inv} = q\frac{\sum_{k=1}^{2}\mu_k\frac{\Sigma_k^2}{a_k^2}w_k^2}{\sum_{k=1}^{2}\mu_k\frac{1}{a_k^2}w_k^2} \tag{C.28}$$

● **欠参数化情形**（$\beta > 1$）。微分斯蒂尔切斯方法依赖不动点根q 为有限值，而这严格成立于过参数化区间（$\beta < 1$）。当 $\beta > 1$时，样本数¯超过环境维数（$B > d$），导致秩比例饱和于 $\beta = 1$，从数学上使得 $q \to \infty$。

然而，在这一过参数化区间中，未缩放的特征协方差矩阵 $G = MM^T$几乎必然变为严格满秩。因此，归一化矩阵 $\hat{}G = \frac{1}{d}G$ 严格可逆，其伪逆退化为标准逆 $\hat{}G^{-1}$。我们略去扰动导数，直接利用广义样本协方差矩阵逆的确定性等价物计算迹。注意，$\hat{}G^{-1} = A^{-1}W^{-1}A^{-1}$ 在 $W = \frac{1}{d}\Omega\Omega^{\mathsf{T}}$ 时是一个大小为 $d \times B$的标准威沙特矩阵，因此根据马尔可夫-帕斯图尔定律的逆矩，其确定性等价物为 $W \asymp \frac{1}{\beta-1}I_d$，这意味着

$$\hat{G}^{-1} \asymp \frac{1}{\beta-1}(A^2)^{-1} \tag{C.29}$$

将该确定性等价物直接代入目标迹泛函，可得 $\beta > 1$ 的精确闭式极限：

$$\mathcal{T}_{inv} = \frac{1}{d}\mathrm{tr}\left(\Sigma^2\left[\frac{1}{\beta-1}A^{-2}\right]\right) = \frac{1}{\beta-1}\sum_{k=1}^{2}\mu_k\frac{\Sigma_k^2}{a_k^2} \tag{C.30}$$

方程 (C.28) 和 (C.30) 在插值阈值 ($\beta = 1$) 处均发散。

我们将两个方程合并为一个统一关系：

$$\mathcal{T}_{inv} = \left\{q\frac{\sum_{k=1}^{2}\mu_k\frac{\Sigma_k^2}{a_k^2}w_k^2}{\sum_{k=1}^{2}\mu_k\frac{1}{a_k^2}w_k^2}\right\}\mathbf{1}(\beta < 1) + \left\{\frac{1}{\beta-1}\sum_{k=1}^{2}\mu_k\frac{\Sigma_k^2}{a_k^2}\right\}\mathbf{1}(\beta > 1) \tag{C.31}$$

迹项 (T_{var}) 的推导。我们计算迹项 $\mathrm{T}_{var} = \lim_{d\to\infty}\frac{1}{d}\mathrm{tr}(V_S\Sigma_A V_S^T)$。回顾确定性测试算子 $V_S = -\Gamma_0^{-1} + D_{pre}\Pi_M$，其中 $D_{pre} = \Gamma_0^{-1} - \eta A^{-1}$。注意， D_{pre}和 $-\Gamma_0^{-1}$ 均为块对角确定性矩阵。此外，在˜比例渐近区域中，它们在块 $k \in \{1,2\}$ 上各自的标量值分别收敛到 δ_k 和 g_k ，由 (B.7) 和 (B.8) 给出。展开该迹可得：

$$\mathcal{T}_{var} = \frac{1}{d}\mathrm{tr}(\Gamma_0^{-2}\Sigma_A) + \frac{2}{d}\mathrm{tr}(-\Gamma_0^{-1}\Sigma_A D_{pre}\Pi_M) + \frac{1}{d}\mathrm{tr}(D_{pre}\Pi_M\Sigma_A D_{pre}\Pi_M) \tag{C.32}$$

令 $T_{ij} = \frac{1}{d}\mathrm{tr}(\Pi_{ij}\Pi_{ji})$。线性迹仅在对角块上计算。与推导 (C.5) 和 (C.13) 类似，我们有

$$\lim_{d\to\infty}\mathrm{tr}(\Gamma_0^{-2}\Sigma_A) = \sum_{k=1}^{2}\mu_k g_k^2 s_k$$

其中， s_1 和 s_2 是由（B.9）给出的 Σ_A 各块上的标量的极限。此外，

$$\lim_{d\to\infty}\frac{2}{d}\mathrm{tr}(-\Gamma_0^{-1}\Sigma_A D_{pre}\Pi_M) = 2\sum_{k=1}^{2}\mu_k g_k\tilde{\delta}_k s_k w_k$$

该量 块分区 a 上的二次迹Quad $:= \frac{1}{d}\mathrm{tr}(D_{pre}\Pi_M\Sigma_A D_{pre}\Pi_M)$ 展开为 2×2 s:

$$\mathrm{Quad} = \tilde{\delta}_1^2 s_1 T_{11} + \tilde{\delta}_2^2 s_2 T_{22} + \tilde{\delta}_1\tilde{\delta}_2(s_1+s_2)T_{12} \tag{C.33}$$

应用式 (C.11)，可得 $T_{kk} = \mu_k w_k - T_{12}$。将这些约束代入二次展开，得到：

$$\text{Quad} = \sum_{k=1}^{2} \mu_k \tilde{\delta}_k^2 s_k w_k - T_{12}\left[\tilde{\delta}_1^2 s_1 + \tilde{\delta}_2^2 s_2 - \tilde{\delta}_1\tilde{\delta}_2 s_1 - \tilde{\delta}_1\tilde{\delta}_2 s_2\right] \tag{C.34}$$

针对 T_{12} 的括号乘数可解析地分解为$(\delta_1 - \delta_2)(\delta_1 s_1 - \delta_2 s_2)$。将线性分量与二次分量重新组合，即可对对角元素配方，得到：

$$\mathcal{T}_{var} = \sum_{k=1}^{2} \mu_k s_k\left[g_k^2(1-w_k) + (g_k + \tilde{\delta}_k)^2 w_k\right] - T_{12}(\tilde{\delta}_1 - \tilde{\delta}_2)(\tilde{\delta}_1 s_1 - \tilde{\delta}_2 s_2) \tag{C.35}$$

回顾(C.19)，噪声极限 $\frac{1}{d}\mathrm{E}_{\mathcal{E}}\left[\|V_N\Sigma\|_F^2\right]$由方程(C.28)与(C.35)的乘积给出。

C.2 第二项分析

由于$\mathrm{tr}(\Sigma)$的量级为 $O(d)$，在高维极限下，$\mathrm{tr}(\widetilde{V}\Sigma\widetilde{V}\intercal)\mathrm{tr}(\Sigma)$项主导$\mathrm{tr}(\widetilde{V}\Sigma^2\widetilde{V}\intercal)$项。令 $\bar{\Sigma} = \lim_{d\to\infty}\frac{1}{d}\mathrm{tr}(\Sigma)$，第二项的主导分量求值为：

$$\text{Term II} = \gamma\bar{\Sigma}\cdot\frac{1}{d}\mathbb{E}\left[\mathrm{tr}(\tilde{V}\Sigma\tilde{V}^T)\right] \tag{C.36}$$

回顾 $\tilde{V} = V_S + V_N$，其中 V_N 的均值为零。由于交叉项为零，我们得到如下分解：

$$\text{Term II} = \underbrace{\gamma\bar{\Sigma}\frac{1}{d}\mathrm{tr}(V_S\Sigma V_S^T)}_{\text{Term II Signal}} + \underbrace{\gamma\bar{\Sigma}\frac{1}{d}\mathbb{E}_{\mathcal{E}}\left[\mathrm{tr}(V_N\Sigma V_N^T)\right]}_{\text{Term II Noise}} \tag{C.37}$$

项 II 信号的推导。我们计算 $\mathrm{T}_{var,\Sigma} = \frac{1}{d}\mathrm{tr}(V_S\Sigma V_S^T)$。确定性算子为 $V_S = -\Gamma_0^{-1} + D_{pre}\Pi_M$，其中 $D_{pre} = \Gamma_0^{-1} - \eta A^{-1}$ 是块对角的。展开该迹可得：

$$\mathcal{T}_{var,\Sigma} = \frac{1}{d}\mathrm{tr}(\Gamma_0^{-2}\Sigma) + \frac{2}{d}\mathrm{tr}(-\Gamma_0^{-1}\Sigma D_{pre}\Pi_M) + \frac{1}{d}\mathrm{tr}(D_{pre}\Pi_M\Sigma D_{pre}\Pi_M) \tag{C.38}$$

观察 $\mathrm{T}_{var,\Sigma}$ 的表达式，它与 T_{var} 相同，只是将 Σ_A 替换为 Σ。因此，通过与 (C.35) 类似的推导，我们得到

$$\mathcal{T}_{var,\Sigma} = \sum_{k=1}^{2} \mu_k \Sigma_k\left[g_k^2(1-w_k) + (g_k + \tilde{\delta}_k)^2 w_k\right] - T_{12}(\tilde{\delta}_1 - \tilde{\delta}_2)(\tilde{\delta}_1\Sigma_1 - \tilde{\delta}_2\Sigma_2) \tag{C.39}$$

项 II 噪声的推导。我们必须计算 $\frac{1}{d}\mathrm{E}\left[\mathrm{tr}(V_N\Sigma V_N^T)\right]$。按照与 (C.19) 类似的推导，将 Σ 替换为 $\Sigma^{1/2}$，我们得到

$$\frac{1}{d}\mathbb{E}[\|V_N\Sigma^{1/2}\|_F^2] = \gamma\underbrace{\left[\mathrm{tr}(\Sigma(MM^T)^\dagger)\right]}_{\mathcal{T}_{inv,\Sigma}}\cdot\underbrace{\left[\frac{1}{d}\mathrm{tr}(V_S\hat{\Sigma}_{\mathcal{E}}V_S^T)\right]}_{\mathcal{T}_{var}} \tag{C.40}$$

注意，我们已在项 I 的分析中刻画了 T_{var}。

接下来，我们评估 $\mathrm{T}_{inv,\,\Sigma} = \lim_{d\to\infty}\frac{1}{d}\mathrm{tr}(\Sigma G^\dagger)$，其中 $G = MM^\dagger = A\Omega\Omega^T A$。请注意，$\mathrm{T}_{inv,\,\Sigma}$的表达式与 T_{inv} 相同，只需将 Σ^2 替换为 Σ。按照 (C.31) 的相同推导过程，我们得到

$$\mathcal{T}_{inv,\Sigma} = \left\{q\frac{\sum_{k=1}^2 \mu_k \frac{\Sigma_k}{a_k^2}w_k^2}{\sum_{k=1}^2 \mu_k \frac{1}{a_k^2}w_k^2}\right\}\mathbf{1}(\beta<1) + \left\{\frac{1}{\beta-1}\sum_{k=1}^{2}\mu_k\frac{\Sigma_k}{a_k^2}\right\}\mathbf{1}(\beta>1) \tag{C.41}$$

Comb结合上述刻画，第二项各分量的极限为 y:

$$\text{Term II Signal} = \gamma\bar{\Sigma}\cdot\mathcal{T}_{var,\Sigma} \tag{C.42}$$

$$\text{Term II Noise} = \gamma^2\bar{\Sigma}\cdot\mathcal{T}_{inv,\Sigma}\cdot\mathcal{T}_{var} \tag{C.43}$$

其中，$\mathrm{T}_{var,\ \Sigma}$ 由公式（C.39）给出，$\mathrm{T}_{inv,\ \Sigma}$ 由公式（C.41）给出，而 T_{var} 由第一项推导中的（C.35）给出。将（B.5）中针对第一项和第二项推导出的刻画结合起来，即完成证明。

C.2.1 引理 C.1 的证明

为计算 Π_M 的渐近迹，我们将经验特征矩阵 M 的列空间上的正交投影算子表示为正则化参数 $z\to 0^+$ 下岭正则化逆矩阵的极限：

$$\Pi_M = \lim_{z\to 0^+} M(M^TM + zI_B)^{-1}M^T = I_d - \lim_{z\to 0^+} z(G + zI_d)^{-1} \tag{C.44}$$

其中，$G = MM^T = A\Omega\Omega^TA$ 是广义样本协方差矩阵，$R(z) = (G + zI_d)^{-1}$ 是其预解式。

根据 Bai-Silverstein 定理，当 $d,\ B\to\infty$ 且 $B/d\to\beta$ 时，随机预解式 $R(z)$ 与确定性对角矩阵 $T(z)$ 渐近等价。对于任意有界确定性矩阵 D，其归一化迹几乎处处收敛：

$$\lim_{d\to\infty}\frac{1}{d}\mathrm{tr}(DR(z)) - \frac{1}{d}\mathrm{tr}(DT(z)) \xrightarrow{a.s.} 0 \tag{C.45}$$

其中 $T(z)$ 定义为

$$T(z) = \left(zI_d + v(z)A^2\right)^{-1},$$

且 $v(z)$ 是伴随矩阵 $\tilde{G} = \Omega^TA^2\Omega$ 的斯蒂尔切斯变换。

我们将第 k 个块中保留的有效秩比例定义为限制在该子空间上的投影矩阵的归一化迹：

$$w_k = \lim_{d\to\infty}\frac{1}{d_k}\mathrm{tr}(\Pi_{kk}) \tag{C.46}$$

代入预解式的极限及其确定性等价物 $T(z)$：

$$\begin{aligned} w_k &= 1 - \lim_{z\to 0^+}\frac{1}{d_k}\sum_{i\in\text{Block }k} zT_{ii}(z) \\ &= 1 - \lim_{z\to 0^+}\frac{z}{z + a_k^2v(z)} \\ &= 1 - \lim_{z\to 0^+}\frac{1}{1 + a_k^2\frac{v(z)}{z}} \end{aligned} \tag{C.47}$$

我们将严格的斯蒂尔切斯不动点根 q 定义为该比值在原点附近的极限：

$$q = \lim_{z\to 0^+}\frac{v(z)}{z} \tag{C.48}$$

将 q 代入极限，得到块权重的以下关系：

$$w_k = 1 - \frac{1}{1 + a_k^2q} = \frac{a_k^2q}{1 + a_k^2q} \tag{C.49}$$

为确定不动点根 q，我们利用 $d\times d$ 广义样本协方差矩阵 G 及其 $B\times B$ 伴随矩阵 $\tilde{G} = \Omega^TA^2\Omega$ 的预解式之间的迹恒等式。由于它们的非零特征值完全相同，特征预解式的归一化迹 $m(z) = \frac{1}{d}\mathrm{tr}(R(z)) = \frac{1}{d}\mathrm{tr}[(G + zI_d)^{-1}]$ 可表示为：

$$zm(z) = 1 - \beta + \beta zv(z)\,. \tag{C.50}$$

根据 Bai-Silverstein 定理， $m(z)$ 渐近等价于确定性矩阵$T(z)$的迹。代入该确定性等价物后，得到：

$$m(z) = \sum_{k=1}^{K} \mu_k \frac{1}{z + a_k^2 v(z)} \tag{C.51}$$

乘 以 z 相乘，并将其与迹恒等式 (C.50) 相等，即可得到精确关系 ：

$$1 - \beta + \beta z v(z) = \sum_{k=1}^{K} \mu_k \frac{z}{z + a_k^2 v(z)} = \sum_{k=1}^{K} \mu_k \frac{1}{1 + a_k^2 \frac{v(z)}{z}} \tag{C.52}$$

我们求该方程在 $z \to 0^+$时的严格极限。在右侧，代入根 $q = \lim_{z\to 0^+} v(z)$ 的定义。在左侧， $zv(z)$ 的极限由 伴随矩阵 G的零空间维度决定。 G 的最大秩受限于 d。若 $B > d$ （即 $\beta > 1$），伴随矩阵恰有 $B - d$ 个严格为零的特征值，且预解式的迹与 $\frac{B-d}{B}\frac{1}{z}$ 成正比。因此，我们可精确地计算该极限：

$$\lim_{z\to 0^+} zv(z) = \max\left(1 - \frac{1}{\beta}, 0\right) \tag{C.53}$$

将这些极限代入迹恒等式两侧，得到：

$$1 - \beta + \beta \max\left(1 - \frac{1}{\beta}, 0\right) = \sum_{k=1}^{K} \mu_k \frac{1}{1 + a_k^2 q} \tag{C.54}$$

左侧在数学上恰好化简为 $\max(1-\beta, 0)$。在右侧，我们代入块权重 $w_k = \frac{a_k^2 q}{1+a_k^2 q}$的定义，并利用恒等式 $\frac{1}{1+a_k^2 q} = 1 - w_k$：

$$\max(1 - \beta, 0) = \sum_{k=1}^{K} \mu_k (1 - w_k) = 1 - \sum_{k=1}^{K} \mu_k w_k \tag{C.55}$$

立即重新整理各项可得：

$$\sum_{k=1}^{K} \mu_k w_k = 1 - \max(1 - \beta, 0) = \min(\beta, 1) = \bar{\beta} \tag{C.56}$$

这一推导对所有参数化情形均成立。在欠参数化情形($\beta > 1$)中，有效秩比例饱和于 $\beta = 1$，这在数学上必然意味着 $w_k \to 1$ 和 $q \to \infty$。方程 (C.49) 和 (C.56) 完全且确定地参数化了随机投影 Π_M 的有限维迹。

C.2.2 引理 C.2 的证明

我们将正交 投影矩阵 Π_M 表示为正则化预解式 $R(z) = (A\Omega\Omega^T A + zI_d)^{-1}$ 表示为 $z \to 0^+$：

$$\Pi_M = I_d - \lim_{z\to 0^+} zR(z) \tag{C.57}$$

令 D_1 和 D_2 为子空间 1 和 2 的正交分块指示矩阵，满足 $D_1 D_2 = 0$。具体而言，

$$D_1 = \begin{bmatrix} I_m & 0_{m\times(d-m)} \\ 0_{(d-m)\times m} & 0_{d-m} \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0_m & 0_{m\times(d-m)} \\ 0_{(d-m)\times m} & I_{d-m} \end{bmatrix} \tag{C.58}$$

交叉迹可写为 $\mathrm{tr}(\Pi_{12}\Pi_{21}) = \mathrm{tr}(D_1\Pi_M D_2\Pi_M)$。将预解式极限代入迹的定义可得：

$$T_{12} = \lim_{z\to 0^+}\lim_{d\to\infty}\frac{1}{d}\mathrm{tr}\left(D_1(I_d - zR(z))D_2(I_d - zR(z))\right) \tag{C.59}$$

由于 $D_1D_2 = 0$ ，展开乘积后，所有低于 $R(z)^2$ 阶的项都会精确地消失 ：

$$T_{12} = \lim_{z\to 0^+} z^2\left[\lim_{d\to\infty}\frac{1}{d}\mathrm{tr}\left(D_1R(z)D_2R(z)\right)\right] \tag{C.60}$$

在下一个引理中，我们刻画内层极限。

引理 C.3 令 $R(z) = (A\Omega\Omega^TA + zI_d)^{-1}$ 为广义样本协方差矩阵的预解式，令 $T(z) = (zI_d + v(z)A^2)^{-1}$ 为其确定性等价物。令 D_1 和 D_2 为满足 $D_1D_2 = 0$的 $d\times d$ 对角正交块指示矩阵。在高维极限 $d, B\to\infty$ 且 $B/d\to\beta$时，两个预解式乘积的归一化迹几乎必然收敛至：

$$\lim_{d\to\infty}\frac{1}{d}tr(D_1R(z)D_2R(z)) = \frac{v(z)^2}{\beta}\frac{\Psi_1(z)\Psi_2(z)}{\Delta(z)} \tag{C.61}$$

其中， Ψ_k 和 Δ_k 定义如下：

$$\Psi_k(z) = \lim_{d\to\infty}\frac{1}{d}tr\left(D_kA^2T(z)^2\right),$$
$$\Delta(z) = 1 - \frac{v(z)^2}{\beta}\lim_{d\to\infty}\frac{1}{d}tr\left(A^4T(z)^2\right).$$

利用引理 C.3，我们有

$$\Psi_k(z) = \lim_{d\to\infty}\frac{1}{d}\mathrm{tr}\left(D_kT(z)A^2T(z)\right) = \frac{\mu_ka_k^2}{(z + a_k^2v(z))^2} \tag{C.62}$$

$$\Delta(z) = 1 - \frac{v(z)^2}{\beta}\lim_{d\to\infty}\frac{1}{d}\mathrm{tr}\left(A^4T(z)^2\right) = 1 - \frac{v(z)^2}{\beta}\sum_{k=1}^{2}\frac{\mu_ka_k^4}{(z + a_k^2v(z))^2} \tag{C.63}$$

现在我们计算 $z\to 0^+$ 时的极限。根据斯蒂尔切斯固定点定义 $q = \lim_{z\to 0^+} v(z)\,z$ ，有$v(z) = qz + o(z)$。

首先，我们计算缩放块迹 $z^2\Psi_k(z)$ 的极限：

$$\lim_{z\to 0^+} z^2\Psi_k(z) = \lim_{z\to 0^+}\frac{z^2\mu_ka_k^2}{z^2(1 + a_k^2q)^2} = \frac{\mu_ka_k^2}{(1 + a_k^2q)^2} \tag{C.64}$$

回想 $w_k = \frac{a_k^2q}{1+a_k^2q}$, 这意味着方差因子为 $v_k = w_k(1 - w_k) = \frac{a_k^2q}{(1+a_k^2q)^2}$。除以 q 后，我们便能将块迹精确地对应到方差因子：

$$\lim_{z\to 0^+} z^2\Psi_k(z) = \frac{\mu_kv_k}{q} \tag{C.65}$$

其次，我们使用 (C.61) 来计算由 (C.60) 给出的 T_{12}。将来自投影极限的 z^2 乘数与 $v(z)^2/z^2\to q^2$ 收敛项一并展开，可得：

$$\lim_{z\to 0^+}\frac{1}{\beta}\left(\frac{v(z)^2}{z^2}\right)\left[z^2\Psi_1(z)\right]\left[z^2\Psi_2(z)\right] = \frac{1}{\beta}(q^2)\left(\frac{\mu_1v_1}{q}\right)\left(\frac{\mu_2v_2}{q}\right) = \frac{\mu_1\mu_2v_1v_2}{\beta} \tag{C.66}$$

第三，我们将分母 $\Delta(0)$ 计算为 $z\to 0^+$：

$$\Delta(0) = \lim_{z\to 0^+}\left[1 - \frac{1}{\beta}\left(\frac{v(z)^2}{z^2}\right)\sum_{k=1}^{2}\frac{\mu_ka_k^4}{(1 + a_k^2q)^2}\right] = 1 - \frac{1}{\beta}\sum_{k=1}^{2}\mu_k\frac{a_k^4q^2}{(1 + a_k^2q)^2} \tag{C.67}$$

注意到平方权重 $w_k^2 = \left(\frac{a_k^2 q}{1+a_k^2 q}\right)^2$，，我们得到 $\Delta(0) = 1 - \frac{1}{\beta}\sum_{k=1}^2 \mu_k w_k^2$。我们应用由 (C.56) 给出的恒等式 $\beta = \sum_{k=1}^2 \mu_k w_k$，来替换开头的 1：

$$\Delta(0) = \frac{\sum_{k=1}^2 \mu_k w_k - \sum_{k=1}^2 \mu_k w_k^2}{\beta} = \frac{1}{\beta}\sum_{k=1}^2 \mu_k w_k(1-w_k) = \frac{\mu_1 v_1 + \mu_2 v_2}{\beta} \tag{C.68}$$

最后，将计算得到的分子与分母相除，可得

$$T_{12} = \frac{\frac{\mu_1\mu_2 v_1 v_2}{\beta}}{\frac{\mu_1 v_1 + \mu_2 v_2}{\beta}} = \frac{\mu_1\mu_2 v_1 v_2}{\mu_1 v_1 + \mu_2 v_2} \tag{C.69}$$

至此，证明完毕。

C.2.3 引理 C.3 的证明

最后，通过向预解式引入连续的确定性扰动 t，来计算交叉迹。我们将扰动后的预解矩阵定义为 $R(z,\ t) = (A\Omega\Omega^T A + tD_2 + zI_d)^{-1}$。设 $m_1(z,\ t) = \frac{1}{d}\mathrm{tr}(D_1 R(z,\ t))$ 为其在第一子空间上的归一化迹。

对随机迹 $m_1(z,\ t)$ 关于扰动 t 在 $t=0$ 处求导，便可直接得到目标交叉迹。利用矩阵求导恒等式 $\frac{\partial}{\partial t}M^{-1} = -M^{-1}\frac{\partial M}{\partial t}M^{-1}$：

$$\frac{\partial}{\partial t} m_1(z,t)\Big|_{t=0} = -\frac{1}{d}\mathrm{tr}\Big(D_1 R(z,0) D_2 R(z,0)\Big) = -\frac{1}{d}\mathrm{tr}\Big(D_1 R(z) D_2 R(z)\Big) \tag{C.70}$$

根据 Bai-Silverstein 定理，$R(z,\ t)$ 渐近等价于扰动后的确定性矩阵$T(z,\ t)$。由于扰动 tD_2 仅仅平移对角线，扰动后的斯蒂尔切斯根 $v(z,\ t)$ 使确定性等价物具有以下精确的结构形式：

$$T(z,t) = (zI_d + v(z,t)A^2 + tD_2)^{-1} \tag{C.71}$$

对确定性迹 $\bar{m}_1(z,\ t) = \frac{1}{d}\mathrm{tr}(D_1 T(z,\ t))$ 在 $t=0$ 处求导，得到：

$$\bar{m}_1'(0) = -\frac{1}{d}\mathrm{tr}\Big(D_1 T(z)\big[v'(0)A^2 + D_2\big]T(z)\Big) \tag{C.72}$$

由于 D_1 和 D_2 严格正交（$D_1 D_2 = 0$），且 $T(z)$ 为对角矩阵，各项可交换，因此 D_2交叉项为零（$D_1 T(z) D_2 T(z) = 0$）。因此，

$$\bar{m}_1'(0) = -v'(0)\left[\frac{1}{d}\mathrm{tr}(D_1 A^2 T(z)^2)\right] = -v'(0)\Psi_1(z) \tag{C.73}$$

为计算标量导数 $v'(0)$，我们必须构造扰动根 $v(z,\ t)$ 的不动点方程。根据 Silverstein 方程（Silverstein, 1995年），我们还有

$$\frac{1}{v(z,t)} = z + \frac{1}{\beta d}\mathrm{tr}\Big(A^2 T(z,t)\Big) \tag{C.74}$$

我们对该不动点方程两边关于 t 在 $t=0$ 处求导：

$$-\frac{v'(0)}{v(z)^2} = \frac{1}{\beta d}\mathrm{tr}\left(A^2\frac{\partial}{\partial t}T(z,t)\Big|_{t=0}\right) = -\frac{1}{\beta d}\mathrm{tr}\Big(A^2 T(z)\big[v'(0)A^2 + D_2\big]T(z)\Big) \tag{C.75}$$

将迹算子线性地分配到求和式中，可得：

$$-\frac{v'(0)}{v(z)^2} = -\frac{v'(0)}{\beta d}\mathrm{tr}\big(A^4 T(z)^2\big) - \frac{1}{\beta d}\mathrm{tr}\big(D_2 A^2 T(z)^2\big) \tag{C.76}$$

相乘 两边乘以 $-v(z)^2$，并代入定义 $\Psi_2(z) = \frac{1}{d}\mathrm{tr}(D_2A^2T(z)^2)$，可得 – s:

$$v'(0) = v'(0)\frac{v(z)^2}{\beta d}\mathrm{tr}\left(A^4T(z)^2\right) + \frac{v(z)^2}{\beta}\Psi_2(z) \tag{C.77}$$

将左侧的 $v'(0)$ 项归并后，可显露出精确的宏观涨落分母 $\Delta(z)$:

$$v'(0)\underbrace{\left[1 - \frac{v(z)^2}{\beta d}\mathrm{tr}\left(A^4T(z)^2\right)\right]}_{\Delta(z)} = \frac{v(z)^2}{\beta}\Psi_2(z) \implies v'(0) = \frac{v(z)^2}{\beta}\frac{\Psi_2(z)}{\Delta(z)} \tag{C.78}$$

由于随机迹导数 (C.70) 的渐近极限等于确定性迹导数 (C.73)，我们将 $v'(0)$ 的解析解代入等式 $-\lim \frac{1}{d}$ $\mathrm{tr}(D_1RD_2R) = -v'(0)\Psi_1(z)$。负号相互抵消，得到精确的闭式极限: –

$$\lim_{d\to\infty}\frac{1}{d}\mathrm{tr}\left(D_1R(z)D_2R(z)\right) = \frac{v(z)^2}{\beta}\frac{\Psi_1(z)\Psi_2(z)}{\Delta(z)} \tag{C.79}$$

由此完成证明。

D 命题 B.2 的证明

(*i*) 当 $\beta \to 1^+$ 时，由于 (C.28)、(C.41) 中的项 $1/(\beta-1)$，T_{inv} 和 $\mathrm{T}_{inv,\Sigma}$ 显然发散。同样，当 $\beta \to 1^-$ 时，$q \to \infty$，因此 T_{inv} 和 $\mathrm{T}_{inv,\Sigma}$ 发散。这表明 $\lim_{\beta\to1} F(\beta) = \infty$。

接下来注意，对于 $\beta \geq 1$，模型定义规定了 $w_k = 1$、$v_k = 0$ 和 $T_{12} = 0$。我们将这些常数代入 $F(\beta)$ 的各个组成部分:

$$\begin{aligned}
\mathrm{Bias}(\beta) &= \sum_{k=1}^{2}\mu_k(\alpha_k + \delta_k)^2 \\
\mathcal{T}_{var}(\beta) &= \sum_{k=1}^{2}\mu_k s_k(g_k + \tilde{\delta}_k)^2 \\
\mathcal{T}_{var,\Sigma}(\beta) &= \sum_{k=1}^{2}\mu_k\Sigma_k(g_k + \tilde{\delta}_k)^2 \\
\mathcal{T}_{inv}(\beta) &= \frac{1}{\beta-1}\sum_{k=1}^{2}\mu_k\frac{\Sigma_k^2}{a_k^2} \\
\mathcal{T}_{inv,\Sigma}(\beta) &= \frac{1}{\beta-1}\sum_{k=1}^{2}\mu_k\frac{\Sigma_k}{a_k^2}
\end{aligned}$$

通过代 入这些组成部分到目标函数 $F(\beta)$ 中，我们可以将其写成以下形式 :

$$F(\beta) = C_1 + \frac{C_2}{\beta-1}$$

其中，C_1 和 C_2 是有限的严格正常数，且与 β 无关。导数为 $F'(\beta) = -\ ^{C_2}{}_{(\beta-1)^2}$。由于 $C_2 > 0$，我们有 $F'(\beta) < 0$，这证明了 $F(\beta)$ 在 $\beta > 1$ 上严格递减。当 $\beta \to \infty$ 时，项 $\frac{C_2}{\beta-1} \to 0$。该函数收敛到常数 $\mathcal{C}_1$，其定义如下: $F(\uparrow\infty)$:

$$F(\uparrow\infty) = \sum_{k=1}^{2}\mu_k(\alpha_k + \delta_k)^2 + \gamma\bar{\Sigma}\sum_{k=1}^{2}\mu_k\Sigma_k(g_k + \tilde{\delta}_k)^2$$

(*ii*) 当 $\beta = 0$时，隐变量 $q = 0$，这意味着 $w_k = 0$、 $v_k = 0$和 $T_{12} = 0$。此外，T_{inv} 和 $\mathrm{T}_{inv,\,\Sigma}$ 中的首项q 乘子使两个逆迹项都恰好为零。将这些代入$F(\beta)$后，所有交叉项均被消去，得到：

$$F(0) = \sum_{k=1}^{2} \mu_k \alpha_k^2 + \gamma \bar{\Sigma} \sum_{k=1}^{2} \mu_k \Sigma_k g_k^2$$

接下来，我们评估差值 $\Delta F = F(\uparrow \infty) - F(0)$：

$$\Delta F = \sum_{k=1}^{2} \mu_k \left[(\alpha_k + \delta_k)^2 - \alpha_k^2 \right] + \gamma \bar{\Sigma} \sum_{k=1}^{2} \mu_k \Sigma_k \left[(g_k + \tilde{\delta}_k)^2 - g_k^2 \right]$$

根据参数定义， $a_2 = r$ 和 $\delta_2 = \tilde{\delta}_2 = \frac{1}{1+\kappa} - \frac{\eta}{r}$。展开 $k = 2$的平方扰动可得：

$$(\alpha_2 + \delta_2)^2 - \alpha_2^2 = \left(1 - \frac{\eta}{r}\right)^2 - \alpha_2^2 = \frac{\eta^2}{r^2} - \frac{2\eta}{r} + 1 - \alpha_2^2$$

$$(g_2 + \tilde{\delta}_2)^2 - g_2^2 = \left(-\frac{\eta}{r}\right)^2 - g_2^2 = \frac{\eta^2}{r^2} - g_2^2$$

代入 入 ΔF后，当 $r \to 0^+$ 时，主导阶行为由 $1/r^2$ 项决定：

$$\Delta F = \frac{\eta^2}{r^2} \mu_2 \left(1 + \gamma \bar{\Sigma}\right) + \mathcal{O}\left(\frac{1}{r}\right)$$

由于 $\mu_2(1+\gamma\bar{\Sigma})\eta^2 > 0$，当 $r \to 0^+$时，差值趋于正无穷。因此，存在一个足够小的 $r > 0$，使得 $\Delta F > 0$，或者 $F(\uparrow \infty) > F(0)$。

(*iii*) 接下来，我们计算 $F'(0) = \frac{dF}{d\beta}\Big|_{\beta=0}$。 $\beta < 1$ 情形下的渐近测试误差为：

$$F(\beta) = \mathrm{Bias} + \gamma \mathcal{T}_{inv} \mathcal{T}_{var} + \gamma \bar{\Sigma} \mathcal{T}_{var,\Sigma} + \gamma^2 \bar{\Sigma} \mathcal{T}_{inv,\Sigma} \mathcal{T}_{var}$$

对 β 应用乘积法则后，完整的导数为：

$$F'(\beta) = \mathrm{Bias}' + \gamma \left(\mathcal{T}'_{inv} \mathcal{T}_{var} + \mathcal{T}_{inv} \mathcal{T}'_{var} \right) + \gamma \bar{\Sigma} \mathcal{T}'_{var,\Sigma} + \gamma^2 \bar{\Sigma} \left(\mathcal{T}'_{inv,\Sigma} \mathcal{T}_{var} + \mathcal{T}_{inv,\Sigma} \mathcal{T}'_{var} \right)$$

要在 $\beta = 0$ 处计算该值，我们必须考察逆迹项。T_{inv} 和 $\mathrm{T}_{inv,\,\Sigma}$ 的定义中都含有一个前置因子 q。当 $\beta \to 0$ 时，隐式根 $q \to 0$。由于 q 后面的分数在 $q \to 0$ 时收敛到一个有限常数，我们可以精确地得到：

$$\mathcal{T}_{inv}(0) = 0 \quad \text{and} \quad \mathcal{T}_{inv,\Sigma}(0) = 0$$

将这些零值代入乘积法则后，$\mathrm{T}'_{var}(0)$ 项便全部消去。导数可简化为：

$$F'(0) = \mathrm{Bias}'(0) + \gamma \mathcal{T}'_{inv}(0) \mathcal{T}_{var}(0) + \gamma \bar{\Sigma} \mathcal{T}'_{var,\Sigma}(0) + \gamma^2 \bar{\Sigma} \mathcal{T}'_{inv,\Sigma}(0) \mathcal{T}_{var}(0)$$

我们将最右侧的各项统称为方差惩罚 $V(\gamma)$：

$$V(\gamma) := \gamma \left(\mathcal{T}'_{inv}(0) \mathcal{T}_{var}(0) + \bar{\Sigma} \mathcal{T}'_{var,\Sigma}(0) \right) + \gamma^2 \left(\bar{\Sigma} \mathcal{T}'_{inv,\Sigma}(0) \mathcal{T}_{var}(0) \right)$$

因此， $F'(0) =$ 偏差$'(0) + V(\gamma)$。由于对于任何严格为正的 $r > 0$， $V(\gamma)$ 中的所有项都是有限的，因此在 $\beta = 0$ 处计算的导数都是有限常数。由于 $V(\gamma)$ 中的每一项都乘以 γ 或 γ^2，我们有：

$$\lim_{\gamma \to 0} V(\gamma) = 0$$

接下来我们推导 Bias′(0)。回顾 Bias 的定义：

$$\mathrm{Bias}(\beta) = \sum_{k=1}^{2} \mu_k \left[\alpha_k^2(1-w_k) + (\alpha_k+\delta_k)^2 w_k\right] + T_{12}(\tilde{\delta}_2^2 - \tilde{\delta}_1^2)(\Sigma_1^2 - \Sigma_2^2)$$

展开内层括号并合并 w_k 项后，我们得到：

$$\alpha_k^2 - \alpha_k^2 w_k + (\alpha_k^2 + 2\alpha_k\delta_k + \delta_k^2)w_k = \alpha_k^2 + w_k(2\alpha_k\delta_k + \delta_k^2)$$

由此得到重新表述后的偏差方程：

$$\mathrm{Bias}(\beta) = \sum_{k=1}^{2} \mu_k \alpha_k^2 + \sum_{k=1}^{2} \mu_k w_k(2\alpha_k\delta_k + \delta_k^2) + T_{12}(\tilde{\delta}_2^2 - \tilde{\delta}_1^2)(\Sigma_1^2 - \Sigma_2^2)$$

为对此关于 β 求导，我们通过隐变量 q 应用链式法则。首先，定义常数 $c = \mu_1 a_1^2 + \mu_2 a_2^2$。根据定义方程 $\beta(q) = \sum_{k=1}^{2} \mu_k \frac{a_k^2 q}{1+a_k^2 q}$，我们对 q 求导：

$$\frac{d\beta}{dq} = \sum_{k=1}^{2} \mu_k \frac{a_k^2}{(1+a_k^2 q)^2}$$

在 $q=0$ 处求值得到 $\left.\frac{d\beta}{dq}\right|_0 = \mu_1 a_1^2 + \mu_2 a_2^2 = c$。根据反函数定理，$q'(0) = \left.\frac{dq}{d\beta}\right|_0 = \frac{1}{c}$。现在，我们依次计算子分量 w_k 和 T_{12} 的初始导数：

由于 $w_k = \frac{a_k^2 q}{1+a_k^2 q}$，链式法则给出 $w_k'(0) = a_k^2 q'(0) = \frac{a_k^2}{c}$。此外，当 q 较小时，变量 $v_k = w_k(1-w_k)$ 的一阶展开为 $v_k = a_k^2 q + O(q^2)$。将其代入 T_{12} 的定义可得：

$$T_{12}(q) = \frac{\mu_1\mu_2(a_1^2 q)(a_2^2 q)}{\mu_1(a_1^2 q) + \mu_2(a_2^2 q)} + \mathcal{O}(q^2) = q\frac{\mu_1\mu_2 a_1^2 a_2^2}{c} + \mathcal{O}(q^2)$$

对 β 求导得到 $T_{12}'(0) = q'(0)\frac{\mu_1\mu_2 a_1^2 a_2^2}{c} = \frac{\mu_1\mu_2 a_1^2 a_2^2}{c^2}$。最后，我们将 $w_k'(0)$ 和 $T_{12}'(0)$ 直接代入求导后的偏差方程：

$$\begin{aligned}\mathrm{Bias}'(0) &= \sum_{k=1}^{2} \mu_k w_k'(0)(2\alpha_k\delta_k + \delta_k^2) + T_{12}'(0)(\tilde{\delta}_2^2 - \tilde{\delta}_1^2)(\Sigma_1^2 - \Sigma_2^2) \\ &= \frac{1}{c}\sum_{k=1}^{2} \mu_k a_k^2(2\alpha_k\delta_k + \delta_k^2) + \frac{\mu_1\mu_2 a_1^2 a_2^2}{c^2}(\tilde{\delta}_2^2 - \tilde{\delta}_1^2)(\Sigma_1^2 - \Sigma_2^2)\end{aligned}$$

对于第一类（$k=1$），由定义可得 $\delta_1 = -\alpha_1$，从而得到 $2\alpha_1\delta_1 + \delta_1^2 = -\alpha_1^2$。对于第二类（$k=2$），由于 $r \to 0^+$，项 $a_2^2(2\alpha_2\delta_2 + \delta_2^2) \to r^2(\eta^2/r^2) = \eta^2$。交叉项收敛到 $\frac{\mu_2}{\mu_1 a_1^2}\eta^2((\rho+1)^2 - 1)$。将这些渐近分量相加可得：

$$\begin{aligned}\lim_{r\to 0^+} \mathrm{Bias}'(0) &= \frac{1}{\mu_1 a_1^2}\left[\mu_1 a_1^2(-\alpha_1^2) + \mu_2\eta^2\right] + \frac{\mu_2\eta^2}{\mu_1 a_1^2}\left((\rho+1)^2 - 1\right) \\ &= -\alpha_1^2 + \frac{\mu_2\eta^2}{\mu_1 a_1^2}(\rho+1)^2\end{aligned}$$

由于 $a_1^2 = \eta^2(\rho+1)^2$,，因此可简化为：

$$\lim_{r\to 0^+} \mathrm{Bias}'(0) = -\alpha_1^2 + \frac{\mu_2}{\mu_1}$$

回想一下，$\alpha_1 = (\kappa - 1)/(\rho + \kappa)$ 和 $\kappa = \gamma(\mu\rho + 1 - \mu)$。因此，$\frac{d}{d\gamma}\alpha_1^2(\gamma)\Big|_0 < 0$。此外，

$$\alpha_1^2(0) = \frac{1}{\rho^2} \geq \frac{1 - \mu_1}{\mu_1} = \frac{\mu_2}{\mu_1},$$

根据我们的假设。由连续性可知，对于足够小的 γ，我们有 $\alpha_1^2 \geq \mu_2/\mu_1$，因此有 $\lim_{r\to 0^+}$ 偏差′(0) <0。由于对于足够小的 r，偏差′(0) 严格为负，且当 γ 足够小时，方差惩罚项 $V(\gamma)$ 可以任意小，因此必定存在常数 $r_0 > 0$ 和 $\gamma_0 > 0$，使得对于所有 $r < r_0$ 和 $\gamma < \gamma_0$，我们都有 $F'(0) < 0$。至此，命题得证。

E 结果监督（OS）损失的梯度和海森矩阵计算

为清晰起见，令 $M = I + VS$（对单个批次省略索引 τ），并将损失定义为 $f(V) = \frac{1}{2}\|M^k w^*\|^2$。我们采用微分法。令 dV 为 V 中的一个小扰动。那么 $dM = (dV)S$。损失的微分为：

$$df = \langle M^k w^*, d(M^k w^*)\rangle = (w^*)^T (M^k)^T d(M^k) w^*$$

根据微分幂法则，$d(M^k) = \sum_{j=0}^{k-1} M^j (dM) M^{k-1-j}$。代入 $dM = (dV)S$：

$$df = \sum_{j=0}^{k-1} (w^*)^T (M^k)^T M^j (dV) S M^{k-1-j} w^*$$

利用性质 $\mathrm{tr}(A^T BC) = \mathrm{tr}(CA^T B)$，我们将 dV 单独列出：

$$df = \mathrm{tr}\left(dV \sum_{j=0}^{k-1} S M^{k-1-j} w^* (w^*)^T (M^k)^T M^j \right)$$

梯度 $\nabla_V \mathrm{L}$ 是与 dV 相乘的矩阵的转置：

$$\nabla_V \mathcal{L} = \sum_{j=0}^{k-1} (M^T)^j M^k w^* (w^*)^T (M^T)^{k-1-j} S^T$$

接下来计算损失的海森矩阵。梯度可以视为多个项的乘积：$G(V) = \sum_{j=0}^{k-1} A_j(V) M^k(V) B_j(V)$，其中 $A_j(V) = (M\tau)^j$，且 $B_j(V) = w^*(w^*)^T (M^T)^{k-1-j} S^T$。对微分应用乘积法则 dG：

$$dG = \sum_{j=0}^{k-1} \left((dA_j) M^k B_j + A_j (dM^k) B_j + A_j M^k (dB_j) \right)$$

在全局最小值附近，项 $M^k w^* \approx 0$。在此情形下，包含 M^k （外部因子）的项会消失，只留下微分直接作用于 M^k 的项。因此，

$$dG \approx \sum_{j=0}^{k-1} (M^T)^j (dM^k) w^* (w^*)^T (M^T)^{k-1-j} S^T$$

代入 $d(M^k) = \sum_{j=0}^{k-1} M^j (dM) M^{k-1-j}$，我们得到

$$H[dM] \approx \sum_{j=0}^{k-1} \sum_{l=0}^{k-1} (M^T)^j \left(M^l (dM) M^{k-1-l} \right) w^* (w^*)^T (M^T)^{k-1-j} S^T$$

其中，对于方向 E，我们有 $H[E] = \frac{d}{dt}\nabla_V \mathrm{L}(V + tE)\big|_{t=0}$。

接下来，我们对海森矩阵的谱范数给出上界：

$$\begin{aligned}\|H\|_{\mathrm{op}} &\le \sum_{j=0}^{k-1}\sum_{l=0}^{k-1} \left\|M^j\right\|_{\mathrm{op}} \left\|M^l\right\|_{\mathrm{op}} \left\|M^{k-1-l}\right\|_{\mathrm{op}} \left\|M^{k-1-j}\right\|_{\mathrm{op}} \|w^*\|_{\ell_2}^2 \|S\|_{\mathrm{op}} \\ &\le \sum_{j=0}^{k-1}\sum_{l=0}^{k-1} \|M\|_{\mathrm{op}}^j \|M\|_{\mathrm{op}}^l \|M\|_{\mathrm{op}}^{k-1-l} \|M\|_{\mathrm{op}}^{k-1-j} \|w^*\|^2 \|S\|_{\mathrm{op}} \\ &= k^2 \rho(M)^{2k-2} \|w^*\|_{\ell_2}^2 \|S\|_{\mathrm{op}} ,\end{aligned}$$

其中第二步根据算子范数的次乘性得出。

参考文献

S. N. Akter、S. Prabhumoye、E. Nyberg、M. Patwary、M. Shoeybi、Y. Choi 和 B. Catanzaro。前置推理：预训练与后训练数据之间的协同作用。*arXiv* 预印本 *arXiv:2510.03264*，2025年。G. Aminian、A. R. Asadi、I. Shenfeld 和 Y. Mroueh。采用多个参考模型的 KL正则化RLHF：精确解与样本复杂度。载于第*39*届神经信息处理系统年会，2025年。S. Garg、D. Tsipras、P. S. Liang 和 G. Valiant。Transformer 能从上下文中学到什么？简单函数类案例研究。神经信息处理系统进展，35:30583–30598，2022年。E. Guha、R. Marten、S. Keh、N. Raoof、G. Smyrnis、H. Bansal、M. Nezhurina、J. Mercat、T. Vu、Z. Sprague 等。OpenThoughts：推理模型的数据配方。*arXiv* 预印本 *arXiv:2506.04178*，2025年。D. Guo、D. Yang、H. Zhang、J. Song、R. Zhang、R. Xu、Q. Zhu、S. Ma、P. Wang、X. Bi 等。DeepSeek-R1：通过强化学习激发大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*，2025年。J. Huang、Z. Wang 和 J. D. Lee。Transformer 学会通过思维链实现多步梯度下降。*arXiv* 预印本 *arXiv:2502.21212*，2025a年。X. Huang、J. Wu、H. Liu、X. Tang 和 Y. Zhou。m1：释放大语言模型测试时缩放在医学推理中的潜力。*arXiv* 预印本 *arXiv:2504.00869*，2025b年。A. Javanmard、B. Mirzasoleiman 和 V. Mirrokni。理解训练数据在测试时缩放中的作用。*arXiv* 预印本 *arXiv:2510.03605*，2025年。J. Li、A. Fang、G. Smyrnis、M. Ivgi、M. Jordan、S. Y. Gadre、H. Bansal、E. Guha、S. S. Keh、K. Arora 等。DataComp-LM：寻找新一代语言模型训练集。神经信息处理系统进展，37:14200–14282，2024年。Meta。《Llama 4 家族：原生多模态人工智能创新新时代的开端》。*https://ai.meta.com/blog/llama-4-multimodal-intelligence/*，2025年4月。N. Muennighoff、Z. Yang、W. Shi、X. L. Li、L. Fei-Fei、H. Hajishirzi、L. Zettlemoyer、P. Liang、E. Candès 和 T. B. Hashimoto。s1：简单的测试时缩放。载于*2025*年自然语言处理实证方法会议论文集，第20286–20332页，2025年。D. Nguyen、W. Yang、R. Anand、Y. Yang 和 B. Mirzasoleiman。用于数据混合中语言模型内存高效训练的小批量核心集。*arXiv* 预印本 *arXiv:2407.19580*，2024年。

OpenAI。Lear 习用大语言模型进行推理。*https://openai.com/index/learning-to-reason-with-llms/*，2024年 .

A. Radford、J. Wu、R. Child、D. Luan、D. Amodei、I. Sutskever 等。语言模型是无监督多任务学习器。*OpenAI 博客*，1(8):9，2019年。J. W. Silverstein。高维随机矩阵特征值经验分布的强收敛性。《多元分析杂志》，55(2):331–339，1995年。A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin。注意力就是你所需要的一切。神经信息处理系统进展，30，2017年。S. M. Xie、H. Pham、X. Dong、N. Du、H. Liu、Y. Lu、P. S. Liang、Q. V. Le、T. Ma 和 A. W. Yu。Doremi：优化数据混合，加速语言模型预训练。神经信息处理系统进展，36:69798–69818，2023年。W. Xiong、H. Dong、C. Ye、Z. Wang、H. Zhong、H. Ji、N. Jiang 和 T. Zhang。从人类反馈中迭代式学习偏好：弥合 KL 约束下基于人类反馈的强化学习的理论与实践。在第四十一届国际机器学习会议上。Y. Yang、S. Mishra、J. Chiang 和 B. Mirzasoleiman。Smalltolarge（s2l）：通过总结小型模型的训练轨迹，为微调大语言模型实现可扩展的数据选择。神经信息处理系统进展，37:83465–83496，2024年。Y. Yue、Z. Chen、R. Lu、A. Zhao、Z. Wang、S. Song 和 G. Huang。强化学习真的能在基础模型之外提升大语言模型的推理能力吗？*arXiv* 预印本 *arXiv:2504.13837*，2025年。W. Zeng、Y. Huang、Q. Liu、W. Liu、K. He、Z. Ma 和 J. He。SimpleRL-Zoo：在真实环境中研究并驯服开放基础模型的零强化学习。*arXiv* 预印本 *arXiv:2503.18892*，2025年。H. Zhao、C. Ye、W. Xiong、Q. Gu 和 T. Zhang。在线 KL 正则化强化学习的对数遗憾。$CoRR$，2025年。