arXiv:2602.01148v1 [cs.AI] 1 Feb 2026

Cap潜在思维链的能力与基本极限

Jiaxuan Zou [*1] **Yaozhong Xiong** [*2] **Yong Liu** [2]

摘要

潜在思维链（潜在思维链）模型通过连续表征实现高效推理，但其性能表现令人费解且不一致：在探索任务上表现出色（ProsQA：97.0%），在计算任务上却表现不佳（GSM8K：34.1%）。我们揭示，这种**权衡**受决策确定性支配。我们的贡献有三方面：（1）我们从理论上刻画了探索—执行权衡，证明高确定性有助于精准执行，却会抑制探索；低确定性则有利于搜索，但会导致错误累积。（2）我们提出**符号索引**，将其量化为决策承诺程度，并指出它是支配这一权衡的核心机制，同时确立了它与执行稳定性和探索能力之间的因果关系。（3）我们证明课程学习在理论上不可或缺，因为直接训练会因分布不匹配而必然失败。我们的框架将设计范式从二元架构选择转向自适应系统，使其能够根据任务需求动态调节决策确定性。

1. 引言

大语言模型（LLM）的出现（Guo 等人，2025；OpenAI，2025；DeepSeek-AI，2025）标志着人工智能领域的范式转变，从根本上改变了复杂推理任务的格局（Liu 等人，2025）。迄今为止，激发模型推理能力的主流方法一直是显式思维链（CoT）（Wei 等人，2022；Sun 等人，2023）。通过迫使模型将中间步骤表述为离散词元，思维链构建出结构化且可供人类理解的逻辑流程，从而显著提升性能。然而，这种显式表述代价高昂：由于序列长度过长以及自回归瓶颈带来的计算负担，这种方法存在固有的低效问题（Hong 等人，2025；Yue 等人，2025）。为了解决这些可扩展性限制，近期研究已大力转向隐式推理，通常称为潜在思维链（Ye 等人，2025）。在这一范式中，模型在连续内部状态空间中运行，通过“静默”的向量变换处理信息，而不生成中间词元。借助词元级操控（Tack 等人，2025；Sun 等人，2025）和轨迹优化（Cheng 和 Van Durme，2024；Hao 等人，2024a）等机制，这些模型有望降低计算复杂度，并支持更多样化、高维度的推理路径，使其不受自然语言词汇表的限制（Hao 等人，2024a；Zhang 等人，2025a）。

尽管出现了这些架构创新，实证性能却呈现出一种鲜明且令人费解的二分现象。如表1所示，显式思维链与潜在思维链表现出互补且互斥的失效模式。显式思维链擅长需要精确符号操作和严谨逻辑执行的任务，例如 GSM8K 中的算术推理（准确率为 42.9%）。然而，对于需要灵活探索或策略规划的任务，例如 ProsQA（77.5%），它则明显力有不逮，往往过早地固守一条僵化路径。相反，潜在思维链在探索性任务上表现更优（97.0%），借助其连续表示遍历更广阔的搜索空间。然而，这种灵活性也带来了沉重代价：在精度要求极高、精确状态维护至关重要的计算任务中，潜在思维链会遭遇灾难性失败。此外，潜在思维链还表现出极强的训练脆弱性；正如我们的消融实验结果所示，移除课程学习会导致性能骤降（例如，在 ProntoQA 上从 99.8% 降至 52.4%），这表明这些模型学习推理的方式存在根本性的不稳定。

这些实证观察凸显了我们在理论理解上的一个关键间隔。目前，该领域缺乏统一的框架来解释为何探索性能的提升必然要以牺牲执行精度为代价，以及为何隐式模型本质上难以训练。现有文献大多将这些架构视为二

[1]西安交通大学数学与统计学院，中国陕西省西安市 710049 [2]中国人民大学高瓴人工智能学院人工智能系，中国北京 100872。通讯请联系：Yong Liu < liuyonggsai@ruc.edu.cn>。

预印本。*2026*年*2*月*3*日。

表1. 离散思维链（CoT）与潜在思维（Coconut）推理模型的性能揭示了一种令人费解的权衡。思维链擅长精确计算（GSM8K），但在灵活推理（ProsQA）方面表现不佳。相反，潜在思维链在灵活性方面表现出色，却无法胜任计算，且对训练课程高度敏感。*结果来自（Hao 等，2024b）。

方法	GSM8K		ProntoQA		ProsQA	
	准确率（%）	词元数	准确率（%）	词元数	准确率（%）	词元数
CoT	42.9 ± 0.2	25.0	98.8 ± 0.8	92.5	77.5 ± 1.9	49.4
无思维链	16.5 ± 0.5	2.2	93.8 ± 0.7	3.0	76.7 ± 1.0	8.2
COCONUT	34.1 ± 1.5	8.2	99.8 ± 0.2	9.0	97.0 ± 0.3	14.2
- 无课程学习	14.4 ± 0.8	8.2	52.4 ± 0.4	9.0	76.1 ± 0.2	14.2

元选择——离散还是连续——却未能刻画支配它们各自优缺点的底层决策理论机制。因此，当前的架构设计依赖启发式方法，而非对模型内部推理过程进行有原则的调控，导致系统要么精确但僵化，要么灵活却容易产生幻觉。

在这项工作中，我们提出，这种权衡并非架构的任意产物，而是决策确定性的根本结果。我们将**符号指数**——一种量化模型对特定推理路径承诺程度的指标——作为统一这些现象的核心机制。我们的关键洞见是，显式思维链处于高确定性区间。该区间通过纠错离散化过程（选择特定词元）保证计算保真度，但由于过早承诺于单一轨迹，也限制了探索。相比之下，潜在思维链处于低确定性区间，维持着多条推理路径的叠加态。尽管这种叠加态有助于稳健地探索复杂的解空间，却会遭受噪声累积，最终破坏符号精度，因为没有离散量化步骤来重置内部状态。

我们的贡献有三点。首先，我们从理论上刻画了根本性的探索—执行权衡。我们证明，思维链的高决策确定性会导致探索消失，而潜在思维链的低确定性会放大亚决策噪声，这是算术任务失败的直接原因。其次，我们将符号指数形式化为一种调节指标，为能够在探索与执行之间动态转换的新一代架构提供具体的设计原则。第三，我们通过证明课程学习在理论上是潜在思维链的必要条件，解决了训练稳定性之谜。我们证明，如果没有课程，训练必然会失败，因为模型自行生成的潜在状态与有效推理轨迹之间存在分布不匹配——而课程能够弥合这一间隔，这一点已得到严格证明。

2. 相关工作

与受限于单一路径的显式思维链不同，潜在思维链利用连续潜在空间实现探索性多样化。该领域的研究使模型能够并行探索多条推理轨迹，从而增强其在复杂问题上的鲁棒性。相关技术包括对潜在轨迹进行采样（Chen et al., 2024）、使用概率加权概念（Zhang et al., 2025b），以及注入连续词元以丰富搜索空间（Xu et al., 2025b;a; Gozeten et al., 2025）。这些实证成果展现了探索能力的价值。我们通过证明探索能力源于保持较低的决策确定性——即较低的符号指数——并确立相对于理想均匀探索先验的 KL 散度形式化界，来形式化这一直觉（定理 4.5）。

另一条不同的研究路线聚焦于轨迹级潜在优化，将完整的显式推理链压缩为连续表示。早期研究旨在通过将潜在状态锚定到显式步骤来保持语义保真度（Cheng 和 Van Durme，2024；Liu 等人，2024；Shen 等人，2025b）。较新的研究则通过动态压缩或自适应控制提升效率（Zhang 等人，2025a；Ma 等人，2025；Tan 等人，2025；Wang 等人，2025a）。一种尤为重要的策略是渐进式内化，即通过课程学习（Deng 等人，2024；Hao 等人，2024a；Shen 等人，2025a）或内部迭代（Zeng 等人，2025；Ruan 等人，2025），逐步以潜在思维取代显式步骤。这些方法，尤其是采用课程学习的方法，为我们的分析提供了实证动机。我们的工作补充了这些实证发现，并证明渐进式训练不仅有益，而且在理论上也是克服根本性分布不匹配（定理 5.1）并确保收敛（定理 5.2，第 5 节）所必需的。

作为对这些架构进展的补充，另一条研究路线致力于对内部状态进行细粒度控制与分析。信号引导控制方法会插入专用的非文本生成词元，以引导内部推理（Herel & Mikolov, 2024; Goyal etal., 2024;

Zelikman 等，2024; Pfau 等，2024; Wang 等，2024）。与此同时，内部状态分析运用探测或蒸馏等技术，寻找隐式推理的机制性证据，例如发现注意力模式中编码的推理树（Deng 等，2023; Hou 等，2023; Wang 等，2025b; Yu 等，2024）。这两种方法都凸显了内部状态的核心地位。我们通过证明内部确定性的方差是探索—执行权衡的根本原因（定理 4.11 和 4.12，第 4.4 节），统一了这些机制性洞见。我们的符号指数为这些原本难以洞察的内部状态提供了可计算、可量化的指标，为未来的架构创新提供了有原则的评估标准。

3. 预备知识

我们考虑推理任务中的监督学习，其中输入是随机变量 X，从中抽取实例 x。每个输入 x 与答案 y 通过思维链 $S = (s_1, \ldots\ldots, s_M)$ 相连接。我们比较对这一过程进行建模的两种范式：思维链（CoT）和潜在思维链，后者通过 Coconut 课程进行训练。

思维链（CoT） 思维链模型以自回归方式生成离散推理步骤：

$$p_\theta(S|x) = \prod_{k=1}^{M} p_\theta(s_k|x, S^{(1\ldots k-1)}).$$

在推理时，词元选择会确定一条单一路径——这使精确计算成为可能，但也存在较早步骤出错的风险（Mohtashami 等，2025；Yu 等，2025；Xu 等，2024；Emmons 等，2025）。

潜在思维链（Latent CoT） 潜在思维链在连续空间中迭代，生成潜在状态 $H = (h_1, \ldots\ldots, h_M)$，$h_k \in \mathbb{R}^d$：

$$h_k = f_\theta(h_{k-1}).$$

每个 h_k 都编码了多条潜在推理路径，支持探索，但也会随着步骤增加而累积噪声（Kong 等，2025；Zhu 等，2025；Su 等，2025；Orlicki，2025）。

Coconut 训练（Hao 等，2024b）由于无法直接对 H 进行监督，Coconut 采用了课程学习。在阶段 k，它学习将前缀 $S^{(1\ldots k)}$ 压缩为潜在状态 h_k，以预测后缀 $S^{(k+1\ldots M)}$：

$$\mathcal{L}^{(k)}_{\text{Coconut}}(\theta) = \mathbb{E}_{p(x,S)}[-\log p_\theta(S^{(k+1\ldots M)}|h_k, x)].$$

课程从 $k = 0$（纯思维链）逐步推进到 $k = M$（完整潜在推理），逐渐内化符号步骤，以衔接离散与连续思维链。

4. 理论分析

本节深入分析思维链与潜在思维链之间的根本差异及权衡。我们首先从信息瓶颈理论的视角，揭示 Coconut 训练目标的数学基础。随后，我们直接从两个关键维度分析模型性能：**规划与探索能力**和**计算执行精度**。最后，我们指出，由符号指数量化的**决策确定性**是调节这一权衡的核心机制，并从理论上确立课程学习对于有效训练潜在思维链模型的必要性。

4.1. 与信息瓶颈的对偶关系

由于 Coconut 训练模型采用分阶段的隐式压缩目标，直接分析其行为颇具挑战。为此，我们证明，Coconut 课程在数学上等价于求解一个广为研究的信息论目标——条件信息瓶颈（Tishby & Zaslavsky, 2015）。这种等价关系使我们能够借助信息论中已有的工具，严格刻画模型的特性。

这种对偶性可以直观地理解：阶段 k 的 Coconut 目标函数提出了一个信息压缩问题。模型必须将过去的思维链 $S^{(1\ldots k)}$ 压缩为单个潜向量 h_k，其唯一目的是最大限度地提升该潜向量对未来序列 $S^{(k+1\ldots M)}$ 的预测效用。这正是信息瓶颈（IB）原理所要解决的问题：寻找一种对源（过去）进行压缩的“瓶颈”表示，同时最大限度地保留有关目标（未来）的信息。因此，Coconut 训练过程隐式地学会了一个最优的信息压缩器。定理 4.1 将这一直觉形式化。

定理 4.1（Coconut-CIB 对偶性）。在任何有限模型容量的约束下，阶段 k 的 *Coconut* 课程（预备知识 *3*）的优化目标都可以严格地重新表述为一个受约束的优化问题。其拉格朗日对偶恰好就是条件信息瓶颈（*CIB*）问题。具体而言：

原始问题： *Coconut* 的优化过程等价于求解：

$$\min_{p(h_k|X)} H(S^{(k+1\ldots M)} \mid h_k, X) \tag{1}$$

$$\textit{s.t.} \quad I(h_k; S^{(1\ldots k)} \mid X) \le R \tag{2}$$

$$\textbf{\textit{Dual:}} \min_{p(h_k|X)} \Big[I(h_k; S^{(1\ldots k)} \mid X) - \beta(k) I(h_k; S^{(k+1\ldots M)} \mid X) \Big] \tag{3}$$

其中 $\beta(k) > 0$ 理想情况下满足 $\beta(k) \sim \frac{k}{M-k}$。

对 $\beta(k)$的解读。 权衡参数 $\beta(k) \sim \frac{k}{M-k}$ 反映了课程在不同阶段侧重点的变化。随着 k 增加，模型会压缩更多过去的信息，同时保持对较短未来序列的预测能力，从而动态平衡压缩效率与预测效用。这种动态加权确保了每个阶段都达到最佳平衡；我们在 §5 中对训练的必要性进行形式化说明。

主要挑战在于将 *Coconut* 的实际训练损失与形式化的信息论目标联系起来。证明通过将优化表述为一个受约束问题并应用拉格朗日对偶，揭示了 *Coconut* 隐式地求解条件信息瓶颈。证明见附录 *A.2*。

4.2. 潜在思维链擅长探索

我们将推理建模为在决策有向无环图上的遍历$Q := (G, v_{\text{start}}, V_{\text{target}})$。在任意节点 v，令 $N_{\text{valid}}(v)$ 表示有效后续步骤的集合。理想的探索者采用均匀先验 $q_{\text{PR}}(u \mid v) = \frac{1}{|N_{\text{valid}}(v)|}\mathbb{I}[u \in N_{\text{valid}}(v)]$。我们通过 $D_{\text{KL}}(q_{\text{PR}} \| p)$ 衡量其与理想状态的偏差，其中 p 是模型的后续步骤分布：KL 散度较低意味着探索范围广；KL 散度较高则表明模型过度自信并过早承诺。

我们的分析首先聚焦于一条**单一、连贯的推理路径**的内在属性，它构成了思维链范式的基本构建单元。理解此类轨迹的生成过程，对于剖析执行的核心机制至关重要。为将其形式化，我们刻画思维链在成功推理过程中所表现出的高确定性行为。这并非任意的建模选择，而是思维链通过对确定性推理链进行教师强制训练所带来的固有结果，这种训练会自然地诱导出尖锐峰值分布。

假设 4.2 （思维链的κ-集中分布）**.** 在具有 B 个选项的决策点上，我们将单个思维链步骤 p_{CoT}的生成分布建模为抽自浓度参数较大的狄利克雷先验，其浓度参数为$\kappa = \sum_i \alpha_i$。较大的 κ 会产生尖锐峰值分布，反映出确定性、高保真执行所要求的高决策确定性。

在该假设下，思维链不可避免地会在每一步坍缩其分布：

定理 4.3 (思维链的探索不足)。 当 $\kappa \to \infty$ 时，熵 $H(p_{CoT}) \to 0$，并且

$$D_{KL}(q_{PR} \| p_{CoT}) = \frac{B-1}{B}\log\kappa - \log B - \frac{(B-1)\log c}{B} + \mathcal{O}\left(\frac{1}{\kappa}\right),$$

对于一个常数 $c > 0$，这意味着 $D_{KL} \to \infty$ 作为确定性$\kappa \to \infty$。

附录 *A.3* 中的证明。

备注 4.4（基于采样的方法）。我们的分析针对单条推理链的生成。虽然自一致性（Wang 等人，2023）等集成方法通过采样多条推理链来引入探索，但每条单独的 推理链在计算上的完整性仍取决于那些高确定性的承诺，而这正是思维链以执行为核心的特性。因此，我们的核心结果仍具有根本意义。

这量化了思维链在 ProsQA 等探索性任务上的糟糕性能（表1中为 77.5%，而潜在思维链为 97.0%）：这种散度 $D_{\text{KL}} \to \infty$ 将思维链的分布与均匀探索先验之间的距离形式化为可任意增大。

相比之下，潜在思维链的训练目标与条件信息瓶颈（定理 4.1）互为对偶，会施加隐式正则化以抑制过度自信，从而确保持续探索：

定理 4.5（潜在思维链的探索能力保证）。存在常数 $\delta \in (0, 1)$ 和有限的 c，使得

$$D_{KL}(q_{PR} \| p_{Coconut}) \leq -\frac{1}{2}\log\delta - c.$$

证明见附录*A.4*。 这一关于 D_{KL}的有限上界确保潜在思维链在各选项上保持非退化分布，与思维链的无界发散形成鲜明对比（定理 4.3）。这保证了稳健的探索，并与思维链在执行方面的优势相辅相成，我们将在下文分析这一点。

4.3. 潜在思维链在符号计算中的脆弱性

尽管潜在思维链的连续状态空间支持稳健的探索（第 4.2 节），但对于需要高保真符号推理的任务（如 GSM8K）而言，它却是一种不利因素。与思维链的离散符号操作不同，潜在思维链的连续状态表示本质上易受噪声累积影响，从而削弱精确的逐步计算。

为了将其形式化，我们分析微小内部误差的影响，这类误差称为**亚决策扰动**——即会破坏模型内部状态、但不足以改变即时输出的噪声。

定义 4.6 （亚决策级扰动）**：** 设 $l_k = l_k^* + \epsilon_k$ 为步骤 k处的logit向量。若扰动 ϵ_k 满足以下条件，则称其为亚决策级扰动：

$$\text{argmax}(l_k^* + \epsilon_k) = \text{argmax}(l_k^*).$$

这一概念凸显了一个关键的架构差异。思维链模型由于其离散生成过程，天然能够抵抗此类扰动。

定理 4.7(思维链的符号完整性)。设 $S^* = \hat{}(s_1^*, \ldots, s_M^*)$ 为无噪声的思维链轨迹，S 为受到亚决策扰动时的轨迹。那么

$$\mathbb{P}[\hat{s}_M \neq s_M^*] = 0.$$

附录 *A.5* 中的证明。

定理 4.7 形式化地说明了思维链对亚决策扰动的韧性。其底层机制是一种**离散化重置** 过程：在每一步，*argmax* 操作都会充当纠错滤波器，将连续的隐藏状态投影到离散词元，从而消除任何亚决策噪声。这种重置确保后续推理从干净的符号表示开始，防止误差沿推理链传播。

相比之下，潜在思维链缺少这种离散化重置机制。其连续状态 h_k 未经量化便直接传递到下一步，并携带任何扰动。对于转移函数为 f_θ（利普希茨常数为 L_F）且噪声 $\epsilon_h^{(k)} \sim \mathcal{N}$ 为独立同分布的模型（0，$\sigma_h^2 I_d$），这种误差传播可以量化为：

定理 4.8（潜在计算中的误差累积）。设 $E_M = h_M - h_M^*$。则对于 $M \geq 1$,

$$\mathbb{E}[\|E_M\|^2] = \frac{1 - L_F^{2M}}{1 - L_F^2} \cdot d\sigma_h^2 > 0.$$

证明见附录 *A.6*。 $\frac{1-L_F^{2M}}{1-L_F^2}$ 一项表明，期望平方误差会随着推理步骤数量 M 的增加而增大。如果模型的转移函数具有扩张性（$L_F > 1$），误差会呈指数级累积；即使对于稳定函数（$L_F \leq 1$），误差仍会不断累积。这一数学结果直接解释了潜在思维链在 GSM8K 等对精度敏感的任务上的性能下降。在较长的推理链中，即使是微小扰动的累积也会破坏最终状态，从而损害符号计算所要求的完整性，并补充了第 refexploration abi 节所分析的二分法。

4.4. 通过符号指数统一权衡

第 4.2 节和第 4.3 节的分析揭示了一种二分现象：思维链擅长执行，却缺乏探索；潜在思维链则恰恰相反。在这里，我们通过一个单一且可量化的原则将这两种行为统一起来：每个推理步骤的决策确定性程度。为对此进行形式化，我们引入**符号指数** ($\mathcal{I}_S$)，用于衡量模型对其首选选项的确信程度。

定义 4.9（符号指数 $\mathcal{I}_S$）。在词汇表中有效词元的集合为 $\mathcal{V}$ 的决策点，设 $p(u \mid h, x)$ 为模型的输出分布。符号指数是概率最高的词元的概率：

$$\mathcal{I}_S = \max_{u \in \mathcal{V}} p(u \mid h, x) \in [1/|\mathcal{V}|, 1].$$

高 $\mathcal{I}_S$ （≈ 1）表示高度确定性，这是思维链离散承诺的特征。低 $\mathcal{I}_S$ 表示对多个选项进行分布式考量，这是潜在思维链的特征。

首先，我们确立这种确定性与抵御计算噪声的鲁棒性之间的直接联系。这种鲁棒性的机制在于前两个候选项之间的分离程度，我们将其称为 Logit 决策间隔。

定义 4.10（Logit 决策间隔 Δ_l）。设 l_{i^*} 和 l_{j^*} 分别为最可能和次最可能词元的 logit。该间隔是二者之差：$\Delta_l = l_{i^*} - l_{j^*}$。

下面的定理证明，高确定性保证较大的决策边际，使模型能够抵御扰动。

定理 4.11（符号稳定性定理）。*Logit* 决策间隔 Δ_l 的下界由符号指数 $\mathcal{I}_S$ 给出，如下所示：

$$\Delta_l \geq \log\left(\frac{\mathcal{I}_S}{1 - \mathcal{I}_S}\right).$$

附录 *A.8* 中的证明。该定理确立了一个直接关系：更高的决策确定性保证更大的保护性间隔。例如，具有$\mathcal{I}_S = 0.99$ 的高确定性思维链（CoT）具有 $\Delta_l \geq \log(99) \approx 4.6$的间隔，使其决策能够抵御显著噪声。相比之下，具有 $\mathcal{I}_S = 0.6$ 的低确定性潜在思维链仅有$\Delta_l \geq \log(1.5) \approx 0.4$的间隔，因此容易受到微小扰动的影响。这解释了思维链的噪声免疫性和符号完整性（定理 4.7）。

然而，这种稳定性直接带来了可量化的探索代价。高确定性会迫使模型的分布偏离无偏探索所需的均匀理想分布。下一个定理将这一权衡形式化。

定理 4.12 (探索—执行权衡定理)。对于一个有 B 个有效选项的决策，与理想的均匀探索先验 q_{PR} 之间的 *KL* 散度下界为：

$$D_{KL}(q_{PR} \| p) \geq \log B + \mathcal{I}_S \log(\mathcal{I}_S) + (1 - \mathcal{I}_S) \log\left(\frac{1 - \mathcal{I}_S}{B - 1}\right).$$

当 $\mathcal{I}_S = 1/B$ 时，该界达到最小值，并随着 $\mathcal{I}_S \to 1$ 增长而增大。

附录 *A.9* 中的证明。 该不等式表明，通过增加 $\mathcal{I}_S$ 必然会获得执行稳定性，同时也会因增加与均匀策略的发散度而削弱探索能力。它量化了专注于单一路径与保留开放选项之间固有的张力，解释了思维链的探索不足（定理 4.3）。

综合来看，这些结果确立了 $\mathcal{I}_S$ 作为核心调节因素，统一了两种范式：

1. **思维链：** 高$\mathcal{I}_S$ 区间可确保较大的决策间隔，从而实现鲁棒执行（定理 4.11、4.7），但会削弱探索（定理 4.12）。
2. **潜在思维链：** 低$\mathcal{I}_S$ 区间能够实现广泛探索（定理 4.5），但意味着较小的决策间隔，从而易受亚决策噪声影响，且该噪声会随步骤累积（定理 4.8），导致在对精度敏感的任务上失败。

该框架将关注点从二元的架构选择转向**管理决策确定性**，并提出能够根据任务需求动态调节 $\mathcal{I}_S$ 的自适应系统。

5. 课程学习为何有效

我们已经确定， $\mathcal{I}_S$ 决定了训练模型中的探索—执行权衡（§4.4）。现在，我们要面对一个根本性问题：如何训练模型，使其能够在稳健探索所需的低-$\mathcal{I}_S$ 状态下运行？这揭示了潜在思维链训练中固有的悖论。不同于显式思维链能够从每个推理步骤的令牌级监督中受益，潜在思维链必须学会生成有意义的潜在表征 h_k，却没有明确的真实标签。

5.1. 潜在思维链的训练悖论

未经训练的模型没有动力去执行编码在 h_k中的多步推理；相反，它往往会学习“捷径”表征，只将表层输入模式映射为输出。训练需要可靠的梯度信号，而要尽量降低损失方差，就需要高确定性（高-$\mathcal{I}_S$）的预测。然而，最终目标是学会能够实现探索的低-$\mathcal{I}_S$ 表征。因此，直接训练会将模型困在高-$\mathcal{I}_S$ 状态，阻碍真正的潜在推理涌现。

为严格分析这一难题，我们在模仿学习（IL）框架下对问题进行建模。设 $P_{\theta^*}(h|x)$表示未观测到的专家策略，它会生成经过推理编码的最优潜在状态。我们定义二元估值函数 $V(h) \in \{0,1\}$，其中 $V(h) = 1$ 意味着该潜在状态会导向正确答案。首要目标是最大化**任务成功率** $\mathcal{R}(\theta)$，其定义为

该函数在诱导策略下的期望值：

$$\mathcal{R}(\theta) = \mathbb{E}_{x\sim\mathcal{D},h\sim P_\theta(\cdot|x)}[V(h)].$$

没有课程时，模型必须从自身生成的潜在状态中学习。这些自生成状态来自 P_{biased} 一种偏置分布，其中包含表面化表征；这种分布与真实的专家分布 P_{θ^*}存在根本差异。由于估值函数 $V(h)$ 是非均匀的（偏向推理状态），从不匹配的分布中学习必然会导致性能欠佳。

定理 5.1 （无课程训练的可证明失败）。 设 D_{nc} 为从偏置分布 P_{biased} 中抽取的潜在状态数据集。若该分布相对于专家本质上是次优的（即 $\mathcal{R}(P_{biased}) \leq \hat{}\mathcal{R}(\theta^*) - \Delta$，对于某个间隔 $\Delta > 0$），则在 D_{nc} 上训练的模型 θ_{MLE} 严格远离最优性：

$$\mathcal{R}(\hat{\theta}_{MLE}) \leq \mathcal{R}(\theta^*) - C(\Delta),$$

其中 $C(\Delta) > 0$ 是由偏差决定的常数。这意味着无论数据集规模如何，模型的成功率都将被永久限制在专家性能之下。

证明见附录*A.10*。 该定理表明，在“捷径”潜在状态上训练会使模型锁定于次优策略。与随数据消失的方差误差（如定理 5.2 所示）不同，这种偏差误差是不可约的。

5.2. 将课程学习作为受控的 $\mathcal{I}_S$ 转变

从符号指数的角度来看，这可以理解为一种受控转变。在早期阶段（纯思维链）中，较高的 $\mathcal{I}_S$ 通过纠错离散化确保了准确的词元级梯度。在混合阶段，前缀在潜在空间中运行（较低的 $\mathcal{I}_S$），但后缀仍保留显式的词元生成。这会将潜在表征锚定到有效的推理轨迹上。最终，模型达到了探索所需的较低 $\mathcal{I}_S$，并通过渐进式内化使其保持稳定。

在我们的模仿学习框架中，该过程实际上提供了以正确推理为基础的潜在思维监督样本，等价于直接从$P_{\theta^*}(h|x)$中抽取样本。关键在于，依分布收敛保证了性能收敛。由于估值函数$V(h)$在 $[0,1]$, 中有界，因此成功率之差$\hat{}|\mathcal{R}(\theta) - \mathcal{R}(\theta^*)|$ 受策略之间总变差距离的上界约束。因此，通过课程学习最小化分布差异，可以直接优化任务成功率。

定理 5.2 （课程训练成功的可证明性）。 在标准统计学习条件下，对通过课程生成的大小为 n 的数据集 D_c 进行最大似然估计

ˆ得到的模型 θ，其成功率趋近于专家的成功率：

$$\mathcal{R}(\hat{\theta}) \geq \mathcal{R}(\theta^*) - \mathcal{O}\left(\sqrt{\frac{d\log n + \log(1/\delta)}{n}}\right),$$

以高概率 $1-\delta$。随着数据集规模 $n \to \infty$，性能差距将消失。

证明见附录 *A.11*。该定理证实，通过将显式思维链用作生成有效训练数据的脚手架，课程能够确保学生模型可证明地收敛到最优专家策略。总之，课程学习在理论上对于解决分布不匹配问题（定理 5.1）并确保收敛（定理 5.2）是必要的。

6. 实验

我们开展实证评估，以验证第 4.4 节建立的理论框架。实验验证了三个假设：（1）潜在思维链和显式思维链运行于不同的决策确定性区间（验证定理 4.3 和 4.5）；（2）这种确定性差异决定了对计算噪声的鲁棒性（验证定理 4.11 和 4.12）；（3）潜在思维链表现出与连续状态动力学一致的误差累积（验证定理 4.8）。

实验设置 我们采用 GPT-2（124M）架构。我们比较标准的显式思维链模型与通过 Coconut 课程训练的潜在思维链模型（$M = 6$ 个潜在步骤）。我们在**GSM8K**（Cobbe 等，2021）上评估精确符号计算能力，并在**ProsQA**（Hao 等，2024b）上评估探索性推理能力。实现细节见附录 A.1。

6.1. 决策确定性分析（$\mathcal{I}_S$）

为验证探索—执行权衡（定理 4.12），我们测量**符号指数**（$\mathcal{I}_S$），其定义为给定步骤中词元概率的最大值。

图 1 和图 2 展示了 $\mathcal{I}_S$ 值的分布：

- **通过较低的 $\mathcal{I}_S$（ProsQA）：** 在 ProsQA（图 2）上，潜在思维链模型保持 $\mathcal{I}_S < 0.6$。这从经验上证实了定理 4.5。持续较低的 $\mathcal{I}_S$ 表明，模型在潜在空间中保留了多条潜在推理轨迹，从而避免过早收敛到次优解。这一机制有助于模型在探索性任务中取得较高的准确率（97.0%）。

- **缺乏离散化（GSM8K）：** 在 GSM8K（图 1）上，模型在没有高概率集中度（• ）的情况下运行。

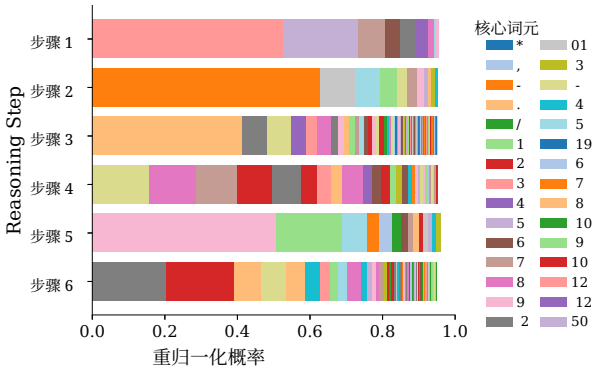


图 *1*。**GSM8K 上的符号指数**。潜在思维链（图示）保持较低的符号指数（$\mathcal{I}_S \in [0.2, 0.5]$），表明其概率分布较为分散。它缺乏概率集中在显式思维链中观察到的.....（$\mathcal{I}_S \approx 1.0$）。

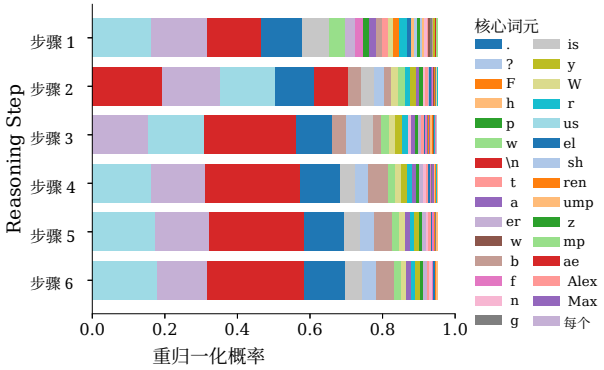


图*2.* **ProsQA 上的符号指数**。潜在思维链在各个推理步骤中呈现出稳定且较低的 $\mathcal{I}_S$ 分布。这验证了定理 4.5，表明模型会将概率质量分配到多条潜在路径上，而非收敛到单个词元。

根据定理 4.11，较低的 $\mathcal{I}_S$ 对应最小的 Logit 决策间隔 Δ_l。与利用“argmax”运算在每个词元处重置状态方差的显式思维链不同，潜在思维链会连续传播方差。由于缺乏中间误差校正，其计算准确率较低（34.1%）。

6.2. 噪声鲁棒性与稳定性分析

为验证误差传播（定理 4.8）和稳定性界（定理 4.11），我们向隐藏状态（潜在思维链）或预输出状态（思维链）中注入高斯噪声$\mathcal{N}(0, \sigma^2 I_d)$，并测量精确匹配准确率的下降。

图 3 中的结果展示了两种截然不同的失效模式：

- **显式思维链（阶跃函数响应）：** 在低噪声水平下，显式思维链的准确率曲线保持平坦

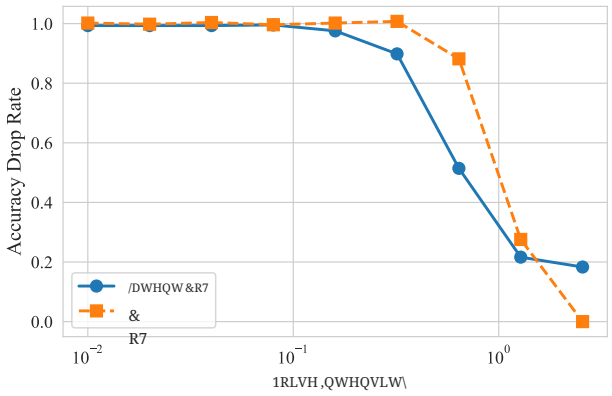


图 3. **噪声下的性能退化。** 标准思维链（橙色）呈现阈值效应：性能保持不变，直到 σ 超过决策边际。潜在思维链（蓝色）则从 $\sigma \approx 0$开始单调衰减。这与附录 A 中的推导 $A($ σ $=$ Φ 7 $)$ $($ $\frac{\Delta_l}{\sqrt{C}\sigma}$ $)$ 一致。

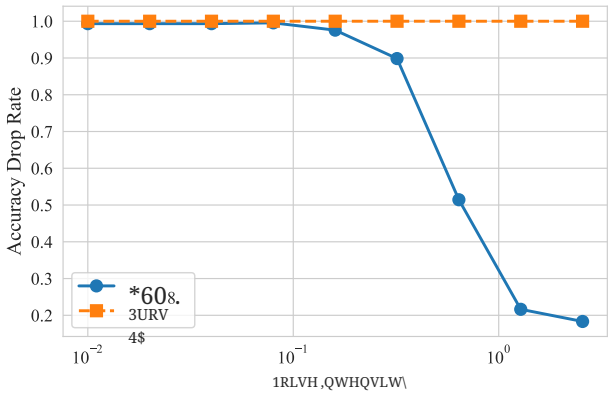


图 4. **任务依赖的敏感性。** 在相同噪声水平下，潜在思维链在 ProsQA（橙色）上的性能高于在 GSM8K（蓝色）上的表现。这表明，与精确的计算任务相比，探索性任务对状态偏差的容忍度更高。

，随后出现急剧下降。这证实了非零 Logit 决策间隔的存在 Δ_l。对于噪声向量 $\|\epsilon\| < \Delta_l$，离散投影（词元选择）能够有效滤除扰动（定理 4.7）。

- **潜在思维链（连续衰减）：** 潜在思维链会立即出现单调退化。这证实了 $\Delta_l \approx 0$ 存在于连续潜在表示中。由于没有离散边际，扰动会直接改变状态轨迹。GSM8K 上的衰减速率支持定理 4.8，表明转移函数 f_θ 会在推理步骤$M = 6$中累积误差。

图 4 显示，鲁棒性取决于任务领域。GSM8K 的性能迅速下降，而 ProsQA 在噪声下仍相对稳定。这表明，探索性任务的解空间是连续的（ h_M 中的微小偏差仍会映射到有效的语义输出），而计算任务则要求 h_M 保持在狭窄区域内，才能生成正确的数值词元。

6.3. 消融实验：课程学习

我们分析了不使用课程学习时潜在思维链的训练稳定性（表1：在 ProntoQA 上的准确率为 52.4%）。对训练过程的监测显示，不使用课程学习时，模型在训练早期便收敛到高-$\mathcal{I}_S$ 状态。这造成了分布不匹配：模型依据输入输出统计特征学习确定性映射，而非专家潜在策略 P_{θ^*} （定理 5.1）。因此，课程学习对于确保分布对齐是必要的，而这种对齐是低-$\mathcal{I}_S$推理能力涌现的前提。

7. 结论

在这项工作中，我们提出了符号指数，将其作为统一的理论框架，用以解释语言模型推理中探索与执行之间的根本权衡。我们的分析指出，决策确定性是其主导机制：显式思维链的高确定性通过纠错式离散化确保计算保真度，但过早承诺会限制探索；而潜在思维链的低确定性有助于广泛搜索，却因误差累积而损害符号完整性。我们还证明，课程学习对于潜在思维链在理论上不可或缺：它通过弥合分布不匹配来确保收敛，否则模型就会受困于次优策略。

尽管符号指数目前充当诊断指标，但它揭示了当前范式设计中的一个关键局限：决策确定性的僵化。我们的框架表明，推理架构的进步需要超越离散模态与连续模态之间的二元选择。相反，未来研究应优先考虑能够动态调节其决策确定性的自适应系统——在用于精确执行的高确定性承诺与用于规划的低确定性多样化之间进行调节。通过将决策确定性确立为核心设计原则，本工作为开发能够根据任务需求在严谨计算与灵活探索之间流畅转换的智能体奠定了基础。

影响声明

本文介绍了一项旨在推动机器学习领域发展的工作。这项工作可能带来许多社会影响，但我们认为没有哪一项需要在此

特别强调。

参考文献

Abbasi-yadkori, Y., P´al, D., and Szepesv´ari, C. 线性随机老虎机的改进算法。载于 Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K.（编），神经信息处理系统进展，第 24 卷。Curran Associates, Inc.，2011 年。

Chen, H., Feng, Y., Liu, Z., Yao, W., Prabhakar, A., Hei- necke, S., Ho, R., Mui, P., Savarese, S., Xiong, C., 等。语言模型是隐藏的推理者：通过自我奖励解锁潜在推理能力。*arXiv* 预印本 *arXiv:2411.04282*，2024 年。

Cheng, J. 和 Van Durme, B. 压缩思维链：通过稠密表示实现高效推理。*arXiv* 预印本 *arXiv:2412.13171*，2024 年。

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. 和 Schulman, J. 训练验证器来解决数学文字题，2021。网址：https://arxiv. org/abs/2110.14168。

DeepSeek-AI。Deepseek-v3.1 模型卡。Hugging Face 模型卡，2025。混合思考/非思考推理、改进的工具使用和智能体任务、长上下文扩展（128 K）、FP8 格式。

Deng, Y., Prasad, K., Fernandez, R., Smolensky, P.,Chaud- hary, V. 和 Shieber, S. 通过知识蒸馏实现隐式思维链推理。*arXiv* 预印本 *arXiv:2311.01460*，2023。

Deng, Y., Choi, Y., 和 Shieber, S. 从显式思维链到隐式思维链：逐步学习内化思维链。*arXiv* 预印本 *arXiv:2405.14838*，2024。

Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., Shafkat, I., 和 Shah, R. 思维链何时不可或缺：语言模型难以躲避监控器。*arXiv* 预印本 *arXiv:2507.05246*，2025。

Goyal, S., Ji, Z., Rawat, A. S., Menon, A. K., Kumar, S., 和 Nagarajan, V. 开口之前先思考：使用暂停词元训练语言模型。发表于第十二届国际学习表征会议，2024。

Gozeten, H. A., Ildiz, M. E., Zhang, X., Harutyunyan, H., Rawat, A. S., 和 Oymak, S. 连续思维链支持并行探索与推理。*arXiv* 预印本 *arXiv:2505.23648*，2025。

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., 等。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*，2025。

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., 和 Tian, Y. 训练大语言模型在连续潜在空间中进行推理。*arXiv* 预印本 *arXiv:2412.06769*，2024a。Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., 和 Tian, Y. 训练大语言模型在连续潜在空间中进行推理，2024b。Herel, D. 和 Mikolov, T. 用于语言建模的思考令牌。*arXiv* 预印本 *arXiv:2405.08644*，2024。

Hong, J., Zhen, T., Chen, K., Liu, J., Zhu, W., Huo, J., Gao, Y., Wang, D., Wan, H., Yang, X., 等。重新思考过度思考：在思维链推理中惩罚内部和外部冗余。*arXiv* 预印本 *arXiv:2508.02178*，2025。

Hou, Y., Li, J., Fei, Y., Stolfo, A., Zhou, W., Zeng, G., Bosselut, A., and Sachan, M. 迈向对语言模型多步推理能力的机制性解释。*arXiv* 预印本 *arXiv:2310.14491*，2023。

Kong, D., Zhao, M., Xu, D., Pang, B., Wang, S., Honig, E., Si, Z., Li, C., Xie, J., Xie, S., 等。通过潜在思维向量的后验推断构建可扩展语言模型。*arXiv* 预印本 *arXiv:2502.01567*，2025。

Liu, J., Qiu, Z., Li, Z., Dai, Q., Zhu, J., Hu, M., Yang,M., 和 King, I. 个性化大语言模型综述：进展与未来方向。*arXiv* 预印本 *arXiv:2502.11528*，2025。

Liu, T., Chen, Z., Liu, Z., Tian, M., 和 Luo, W. 通过隐藏思维链解码加速并提升大语言模型推理。*arXiv* 预印本 *arXiv:2409.08561*，2024。

Ma, X., Wan, G., Yu, R., Fang, G., 和 Wang, X. CoT-valve：长度可压缩思维链调优。*arXiv* 预印本 *arXiv:2502.09601*，2025。

Mohtashami, A., Pagliardini, M., 和 Jaggi, M. CoT-Former：一种思维链驱动的架构，可在推理时实现预算自适应计算成本。载于第十三届国际学习表征会议，2025。

OpenAI。推出 GPT-5。https://openai.com/index/introducing-gpt-5/，2025。访问日期：2025-08-29。

Orlicki, J. I. 超越文字：一种用于大语言模型内部推理的潜在记忆方法。*arXiv* 预印本 *arXiv:2502.21030*，2025。

Pfau, J., Merrill, W. 和 Bowman, S. R. 让我们逐点思考：Transformer语言模型中的隐藏计算。发表于第一届语言建模会议，2024。

Ruan, Y., Band, N., Maddison, C. J. 和 Hashimoto, T. 从潜在思维中学习推理。*arXiv* 预印本 *arXiv:2503.18866*，2025。

Shen, X., Wang, Y., Shi, X., Wang, Y., Zhao, P., andGu,J. 利用隐式思考进行高效推理。*arXiv* 预印本 *arXiv:2501.19201*，2025a。

Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., and He, Y. Codi：通过自蒸馏将思维链压缩至连续空间。*arXiv* 预印本 *arXiv:2502.21074*，2025b。Su, D., Zhu, H., Xu, Y., Jiao, J., Tian, Y., and Zheng, Q. Token assorted：混合潜在词元与文本词元以提升语言模型推理能力。*arXiv* 预印本 *arXiv:2502.03275*，2025。

Sun, J., Zheng, C., Xie, E., Liu, Z., Chu, R., Qiu, J., Xu, J., Ding, M., Li, H., Geng, M., 等。基础模型推理综述。*arXiv* 预印本 *arXiv:2312.11562*，2023。

Sun, Y., Chen, Y., Li, Y. 和 Ding, B. 使用潜在词元增强 Transformer 中的潜在计算。*arXiv* 预印本 *arXiv:2505.12629*，2025。Tack, J., Lanchantin, J., Yu, J., Cohen, A., Kulikov, I., Lan, J., Hao, S., Tian, Y., Weston, J. 和 Li, X. 使用连续概念进行大语言模型预训练。*arXiv* 预印本 *arXiv:2502.08524*，2025。Tan, W., Li, J., Ju, J., Luo, Z., Luan, J. 和 Song, R. 默默思考，快速思考：大语言模型推理链的动态潜在压缩。*arXiv* 预印本 *arXiv:2505.16552*，2025。Tishby, N. 和 Zaslavsky, N. 深度学习与信息瓶颈原理，2015。网址 https: //arxiv.org/abs/1503.02406。Wang, J., Wu, Z., Lai, F., Lian, S. 和 Zeng, Z. Synadapt：通过合成连续思维链在大语言模型中学习自适应推理。*arXiv* 预印本 *arXiv:2508.00574*，2025a。Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A. 和 Zhou, D. 自一致性提升语言模型中的思维链推理，2023。网址 https://arxiv.org/abs/2203. 11171。Wang, X., Caccia, L., Ostapenko, O., Yuan, X., Wang, W. Y. 和 Sordoni, A. 使用规划词元引导语言模型推理。载于第一届语言建模会议，2024。Wang, X., Wang, S., Zhu, Y. 和 Liu, B. System-1.5 推理：在语言与潜在空间中通过

动态捷径进行遍历。*arXiv* 预印本 *arXiv:2505.18962*，2025b。Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. 等。思维链提示可激发大语言模型的推理能力。神经信息处理系统进展，35:24824–24837，2022。Xu, G., Jin, P., Hao, L., Song, Y., Sun, L. 和 Yuan, L. Llava-o1：让视觉语言模型逐步进行推理。*arXiv* 预印本 *arXiv:2411.10440*，2024。Xu, Y., Guo, X., Zeng, Z. 和 Miao, C. Softcot++：使用软思维链推理实现测试时扩展。*arXiv* 预印本 *arXiv:2505.11484*，2025a。

Xu, Y., Guo, X., Zeng, Z., and Miao, C. Softcot：借助软思维链实现大语言模型高效推理。*arXiv* 预印本 *arXiv:2502.12134*，2025b。

Ye, J., Yao, Z., Huang, Z., Pan, L., Liu, J., Bai, Y., Xin, A., Weichuan, L., Che, X., Hou, L., 等。Transformer 如何学习隐式推理？*arXiv* 预印本 *arXiv:2505.23653*，2025。Yu, P., Xu, J., Weston, J. E., 和 Kulikov, I. 将系统2蒸馏到系统1中。载于 *NeurIPS'24* 大规模系统*2*推理第一届研讨会，2024。Yu, Q., He, Z., Li, S., Zhou, X., Zhang, J., Xu, J., 和 He, D. 通过循环对齐推理增强自回归思维链。*arXiv* 预印本 *arXiv:2502.08482*，2025。Yue, L., Du, Y., Wang, Y., Gao, W., Yao, F., Wang, L., Liu, Y., Xu, Z., Liu, Q., Di, S., 等。别想太多：R1风格大型推理模型高效化综述。*arXiv* 预印本 *arXiv:2508.02120*，2025。Zelikman, E., Harik, G. R., Shao, Y., Jayasiri, V., Haber, N., 和 Goodman, N. Quiet-STar：语言模型可以教会自己先思考再发言。载于第一届语言建模会议，2024。Zeng, B., Song, S., Huang, S., Wang, Y., Li, H., He, Z., Wang, X., Li, Z., 和 Lin, Z. 预训练语言模型，使其在连续空间中进行思考。*arXiv* 预印本 *arXiv:2505.20674*，2025。Zhang, J., Zhu, Y., Sun, M., Luo, Y., Qiao, S., Du, L., Zheng, D., Chen, H., 和 Zhang, N. Lightthinker：逐步思考压缩。*arXiv* 预印本 *arXiv:2502.15589*，2025a。Zhang, Z., He, X., Yan, W., Shen, A., Z hao, C., Wang, S., Shen, Y., 和 Wang, X. E. 软思考：释放大语言模型在连续概念空间中的推理潜力。*arXiv* 预印本 *arXiv:2505.15778*，2025b。

Zhu, H., Hao, S., Hu, Z., Jiao, J., Russell, S., 和
Tian, Y. 叠加推理：关于连续思维链的理论视角。
arXiv 预印本 *arXiv:2505.12514*，2025。

A. 附录

A.1. 实验细节

模型。 思维链和潜在思维链（Coconut）模型均基于最小的 GPT-2 变体，包含 124M 个参数。对于潜在思维链，我们在所有实验中使用 $M = 6$ 个潜在推理步骤。潜在思维向量 h_k 的维度与 GPT-2 的隐藏层大小（$d = 768$）相同。

训练。 我们使用 Adam优化器，学习率为 1e-4，权重衰减为 0.01。训练在 4 个图形处理器上采用混合精度进行（ProsQA 启用 bfloat16；由于稳定性问题，GSM8K 禁用 bfloat16）。对于 GSM8K，我们训练 25 个 epoch，批量大小为 32，梯度累积步数为 1；对于 ProsQA，我们训练 30 个 epoch，批量大小为 16，梯度累积步数为 2。Coconut 课程按阶段推进：对于 GSM8K，每个阶段训练 3 个 epoch，直到潜在阶段 3（即将前 3 个思维链步骤压缩为潜在思维）；对于 ProsQA，每个阶段训练 4 个 epoch，直到潜在阶段 6（完全内化）。基础模型检查点初始化自标准 GPT-2。

数据集。我们在两个推理基准测试上进行评估：

1. **GSM8K**：我们采用标准划分，其中留出包含 330 个样例的测试集。训练和验证分别使用官方训练集（7,473 个样例）和 10% 的验证子集。
2. **ProsQA**：按照 Hao等人（2024b）的做法，我们使用包含 300 个样例的测试集。训练集和验证集分别包含 2,000 个和 200 个样例。

所有报告的准确率均在这些测试集上计算。

符号索引可视化详情。图 1 和图 2 中的可视化基于各自测试集中的代表性样例。

• **图 1（GSM8K）：** 该图根据以下数学推理题生成：“Janet 的鸭子每天产 16 个蛋。她每天早晨吃 3 个，还每天用 4 个给朋友烤松饼。她每天把剩下的蛋拿到农贸市场出售，每个新鲜鸭蛋卖 2 美元。她每天在农贸市场能赚多少美元？”

• **图2（ProsQA）：** 该图根据以下逻辑推理问题生成：“每个 shumpus 都是 yumpus。每个 worpus 都是 yimpus。每个 shumpus 都是 gwompus。每个 tumpus 都是 boompus。每个 worpus 都是 shumpus。每个 storpus 都是 terpus。Max 是 yimpus。每个 shumpus 都是 rompus。每个 wumpus 都是 jelpus。每个 boompus 都是 terpus。Fae 是 tumpus。每个 tumpus 都是 worpus。每个 rompus 都是 gorpus。每个 timpus 都是 impus。每个 jompus 都是 gerpus。每个 boompus 都是 rompus。Fae 是 boompus。每个 boompus 都是 kerpus。每个 zumpus 都是 bompus。Max 是 rempus。每个 rompus 都是 kerpus。Max 是 impus。每个 rempus 都是 impus。每个 wumpus 都是 yumpus。每个 grimpus 都是 terpus。每个 tumpus 都是 jompus。每个 yumpus 都是 felpus。每个 jelpus 都是 felpus。每个 shumpus 都是 felpus。每个 rempus 都是 timpus。每个 storpus 都是 jompus。每个 rompus 都是 storpus。每个 tumpus 都是 wumpus。每个 wumpus 都是 jompus。每个 boompus 都是 worpus。Fae 是 storpus。每个 worpus 都是 jelpus。每个 grimpus 都是 felpus。每个 worpus 都是 yumpus。每个 rempus 都是 zumpus。每个 kerpus 都是 grimpus。Fae 是 gwompus 还是 bompus？”

噪声注入协议。 在推理过程中，我们按如下方式注入高斯噪声 $\mathcal{N}(0,\ \sigma^2 I_d)$ ：

1. 对于**思维链**：噪声 会被添加到语言模型输出头之前的最后一个隐藏状态，即 $h \leftarrow h + \sigma \cdot \epsilon$，其中 $\epsilon \sim \mathcal{N}(0,\ I_d)$。
2. 对于 **潜在思维链**：噪声会在每个推理步骤添加，即 $h_k \leftarrow f_\theta(h_{k-1}) + \sigma \cdot \epsilon_k$，且彼此独立的 $\epsilon_k \sim \mathcal{N}(0,\ I_d)$。

我们将准确率报告为测试集上的精确匹配率。 图 3 中的“准确率下降比率”定义为 Acc(σ)/Acc(0).

可复现性。 所有实验均使用随机种子 0。由于我们的评估协议具有确定性，并且在估计符号指数时使用了固定 $p = 0.95$的核采样，因此给定相同的模型检查点，结果完全可复现。

A.2. 定理 4.1 的证明

引理 A.1（损失与互信息）。假设解码器模型族 $\{p_\theta(S^{(k+1\ldots M)} \mid h_k, X)\}$ 设定正确，即存在一个参数 θ^*，使得 $p_{\theta^*}(S^{(k+1\ldots M)} \mid h_k, X) = p(S^{(k+1\ldots M)} \mid h_k, X)$ 几乎处处成立。进一步假设 $H(S^{(k+1\ldots M)} \mid X) < \infty$。那么，

$$\min_\theta \mathcal{L}^{(k)}_{\textit{Coconut}}(\theta) \iff \max_{p(h_k|X)} I(h_k; S^{(k+1\ldots M)} \mid X).$$

证明。展开损失函数：

$$\begin{aligned}
\mathcal{L}^{(k)}_{\text{Coconut}}(\theta) &= \mathbb{E}_{p(X,S,h_k)}[-\log p_\theta(S^{(k+1\ldots M)} \mid h_k, X)] \\
&= \mathbb{E}_{p(X,h_k)}\left[\mathbb{E}_{p(S^{(k+1\ldots M)}|X,h_k)}[-\log p_\theta(S^{(k+1\ldots M)} \mid h_k, X)]\right] \\
&= \mathbb{E}_{p(X,h_k)}\left[\mathbb{E}_{p(S^{(k+1\ldots M)}|X,h_k)}\left[-\log p(S^{(k+1\ldots M)} \mid h_k, X) + \log \frac{p(S^{(k+1\ldots M)} \mid h_k, X)}{p_\theta(S^{(k+1\ldots M)} \mid h_k, X)}\right]\right] \\
&= \mathbb{E}_{p(X,h_k)}\left[H(S^{(k+1\ldots M)} \mid h_k, X) + D_{\text{KL}}(p(\cdot \mid h_k, X) \| p_\theta(\cdot \mid h_k, X))\right].
\end{aligned}$$

根据设定正确假设，存在 w^* 使得 $D_{\text{KL}}(p(\cdot \mid h_k, X) \| p_\theta(\cdot \mid h_k, X)) \to 0$。因此，

$$\min_\theta \mathcal{L}^{(k)}_{\text{Coconut}}(\theta) = \min_{p(h_k|X)} H(S^{(k+1\ldots M)} \mid h_k, X).$$

根据条件互信息的定义，$I(A; B \mid C) = H(A \mid C) - H(A \mid B, C)$，我们有

$$H(S^{(k+1\ldots M)} \mid h_k, X) = H(S^{(k+1\ldots M)} \mid X) - I(h_k; S^{(k+1\ldots M)} \mid X).$$

由于 $H(S^{(k+1\ldots M)} \mid X)$是常数，因此

$$\min_{p(h_k|X)} H(S^{(k+1\ldots M)} \mid h_k, X) \iff \max_{p(h_k|X)} I(h_k; S^{(k+1\ldots M)} \mid X).$$

□

现在开始证明定理 4.1。

证明。根据引理 A.1，最小化 Coconut 损失等价于 $\min_{p(h_k|X)} H(S^{(k+1\ldots M)} \mid h_k, X)$。任何物理上可实现的编码器都受有限模型容量的限制，形式化表示为 $I(h_k; S^{(1\ldots k)} \mid X) \le R$。我们构造拉格朗日函数：

$$\mathcal{J}(p, \lambda) = H(S^{(k+1\ldots M)} \mid h_k, X) + \lambda \left(I(h_k; S^{(1\ldots k)} \mid X) - R\right), \quad \lambda \ge 0.$$

利用恒等式 $H(A \mid B, C) = H(A \mid C) - I(B; A \mid C)$，我们可改写为：

$$\mathcal{J} = \left[H(S^{(k+1\ldots M)} \mid X) - I(h_k; S^{(k+1\ldots M)} \mid X)\right] + \lambda I(h_k; S^{(1\ldots k)} \mid X) - \lambda R.$$

在对 $p(h_k \mid X)$进行最小化时，忽略常数项 $H(S^{(k+1\ldots M)} \mid X)$ 和 λR，得到：

$$\min_{p(h_k|X)} \left\{\lambda I(h_k; S^{(1\ldots k)} \mid X) - I(h_k; S^{(k+1\ldots M)} \mid X)\right\}.$$

定义 $\beta = 1/\lambda > 0$ 并除以 λ，即可得到标准的条件信息瓶颈目标函数：

$$\min_{p(h_k|X)} \left\{I(h_k; S^{(1\ldots k)} \mid X) - \beta I(h_k; S^{(k+1\ldots M)} \mid X)\right\}.$$

记 $I_{\text{past}} = I(h_k; S^{(1\ldots k)} \mid X)$ 和 $I_{\text{future}} = I(h_k; S^{(k+1\ldots M)} \mid X)$。任何帕累托最优编码器都位于信息平面的效率前沿上，其中保留信息与预测信息之间的权衡满足一阶条件：

$$\left.\frac{dI_{\text{future}}}{dI_{\text{past}}}\right|_{h_k^*} = \frac{1}{\beta(k)}.$$

我们将该前沿建模为:

$$I_{\text{future}}(I_{\text{past}}; k) = I_{\max}(k)\left(1 - e^{-\alpha(k)\cdot I_{\text{past}}}\right),$$

参数的设定依据:

$I_{\max}(k) \approx C_2(M-k)$，与剩余序列的长度成正比。

$\alpha(k) \approx 1/H(S^{(1\cdots k)} \mid X) \approx 1/(C_1 k + C_0)$，与过去词元的信息量成反比。求导得:

$$\frac{dI_{\text{future}}}{dI_{\text{past}}} = I_{\max}(k)\cdot\alpha(k)\cdot e^{-\alpha(k)\cdot I_{\text{past}}}.$$

At o在最优状态下，编码器的瓶颈达到饱和: $I_{\text{past}} \approx H(S^{(1\cdots k)} \mid X) \approx 1/\alpha(k)$。代入:

$$\left.\frac{dI_{\text{future}}}{dI_{\text{past}}}\right|_{h_k^*} \approx I_{\max}(k)\cdot\alpha(k)\cdot e^{-1} \approx \frac{C_2(M-k)}{e(C_1k + C_0)}.$$

应用 $\beta(k) = (dI_{\text{future}}/dI_{\text{past}})^{-1}$:

$$\beta(k) = e\cdot\frac{C_1k + C_0}{C_2(M-k)} = e\frac{C_1}{C_2}\cdot\frac{k}{M-k} + O\left(\frac{1}{M-k}\right).$$

因此，当 $k \to M$ 时，主要的渐近行为为 $\beta(k) \sim {}^{k}_{M-k}$，这表明最优权衡随已处理序列长度与剩余序列长度之比而变化。 □

A.3. 定理 4.3 的证明

引理 A.2 (熵的上界 $H(p_{\text{CoT}})$ 思维链输出分布)。 设 $\bar{p} = \mathbb{E}[p_{CoT}]$ 为从参数为 $\boldsymbol{\alpha}$、浓度为 $\kappa = \sum_{i=1}^{B}\alpha_i$的狄利克雷先验中抽取的、在 B 个选项上的下一步期望分布。在高浓度情形下——即当某个 j 满足 $\alpha_j \gg \alpha_{i\neq j}$ 时——熵 $H(\bar{p})$ 会随着 κ 增大而趋于零。具体而言，如果$\alpha_j = \kappa - (B-1)c$，且对于较小的 $c > 0$, $\alpha_{i\neq j} = c$，那么

$$H(\bar{p}) = O\left(\frac{\log\kappa}{\kappa}\right),$$

证明。 根据狄利克雷分布的性质，期望概率为 $\bar{p}_i = \alpha_i/\kappa$。我们考虑一种典型的单峰情形，其中绝大部分权重集中在单个选项 j上。我们令 $\alpha_j = \kappa - (B-1)c$，并对所有 $i \neq j$令 $\alpha_{i\neq j} = c$，其中 c 是一个较小的正数。

在这种设定下，期望概率为:

$$\bar{p}_j = \frac{\kappa - (B-1)c}{\kappa} = 1 - \frac{(B-1)c}{\kappa}$$
$$\bar{p}_{i\neq j} = \frac{c}{\kappa}$$

我们计算该分布的熵:

$$\begin{aligned} H(\bar{p}) &= -\bar{p}_j\log\bar{p}_j - \sum_{i\neq j}\bar{p}_i\log\bar{p}_i \\ &= -\left(1 - \frac{(B-1)c}{\kappa}\right)\log\left(1 - \frac{(B-1)c}{\kappa}\right) - (B-1)\frac{c}{\kappa}\log\left(\frac{c}{\kappa}\right) \end{aligned}$$

我们使用泰勒展开 $\log(1-x) \approx -x$ 来分析第一项。当 $\kappa \to \infty$ 时，有 $(B-1)_{\kappa}{}^{c} \to 0$, 因此:

$$-\left(1 - \frac{(B-1)c}{\kappa}\right)\log\left(1 - \frac{(B-1)c}{\kappa}\right) \approx -\left(1 - \frac{(B-1)c}{\kappa}\right)\left(-\frac{(B-1)c}{\kappa}\right) = O\left(\frac{1}{\kappa}\right).$$

接下来，我们分析第二项：

$$-(B-1)\frac{c}{\kappa}\log\left(\frac{c}{\kappa}\right) = -(B-1)\frac{c}{\kappa}(\log c - \log\kappa)$$
$$= (B-1)c\frac{\log\kappa}{\kappa} - (B-1)c\frac{\log c}{\kappa} = O\left(\frac{\log\kappa}{\kappa}\right).$$

综合这两项，随着 κ 变大，熵由 $O\left(\frac{\log\kappa}{\kappa}\right)$ 项主导。并且我们有 $\lim_{\kappa\to\infty}\frac{\log\kappa}{\kappa}=0$。

□

现在我们开始证明定理 4.3。

证明。 我们分析在集中狄利克雷先验下， $D_{\mathrm{KL}}(q_{\mathrm{PR}}\|p_{\mathrm{CoT}})$的几乎必然渐近行为，其中$\alpha_j = \kappa - (B-1)c$ 且 $\alpha_{i\neq j} = c$。

根据狄利克雷分布的大数定律，当 $\kappa\to\infty$时，随机向量 p_{CoT} 几乎必然收敛于其均值：

$$p_{\mathrm{CoT}} \xrightarrow{a.s.} \bar{p} = \left(1 - \frac{(B-1)c}{\kappa}, \frac{c}{\kappa}, \ldots, \frac{c}{\kappa}\right).$$

由于 KL 散度　由于散度是概率向量的连续函数（在远离边界处），我们有：

$$D_{\mathrm{KL}}(q_{\mathrm{PR}}\|p_{\mathrm{CoT}}) \xrightarrow{a.s.} D_{\mathrm{KL}}(q_{\mathrm{PR}}\|\bar{p}) \quad \text{as } \kappa\to\infty.$$

现在计算 $D_{\mathrm{KL}}(q_{\mathrm{PR}}\|\bar{p})$：

$$D_{\mathrm{KL}}(q_{\mathrm{PR}}\|\bar{p}) = \sum_{i=1}^{B}\frac{1}{B}\log\frac{1/B}{\bar{p}_i}$$
$$= -\log B - \frac{1}{B}\sum_{i=1}^{B}\log\bar{p}_i$$
$$= -\log B - \frac{1}{B}\left[\log\left(1 - \frac{(B-1)c}{\kappa}\right) + (B-1)\log\left(\frac{c}{\kappa}\right)\right].$$

利用泰勒展开 $\log(1-x) = -x + O(x^2)$：

$$D_{\mathrm{KL}}(q_{\mathrm{PR}}\|\bar{p}) = -\log B - \frac{1}{B}\left[-\frac{(B-1)c}{\kappa} + O\left(\frac{1}{\kappa^2}\right) + (B-1)\log c - (B-1)\log\kappa\right]$$
$$= \frac{B-1}{B}\log\kappa - \log B - \frac{(B-1)\log c}{B} + O\left(\frac{1}{\kappa}\right).$$

因此，当 $\kappa\to\infty$时，几乎必然有：

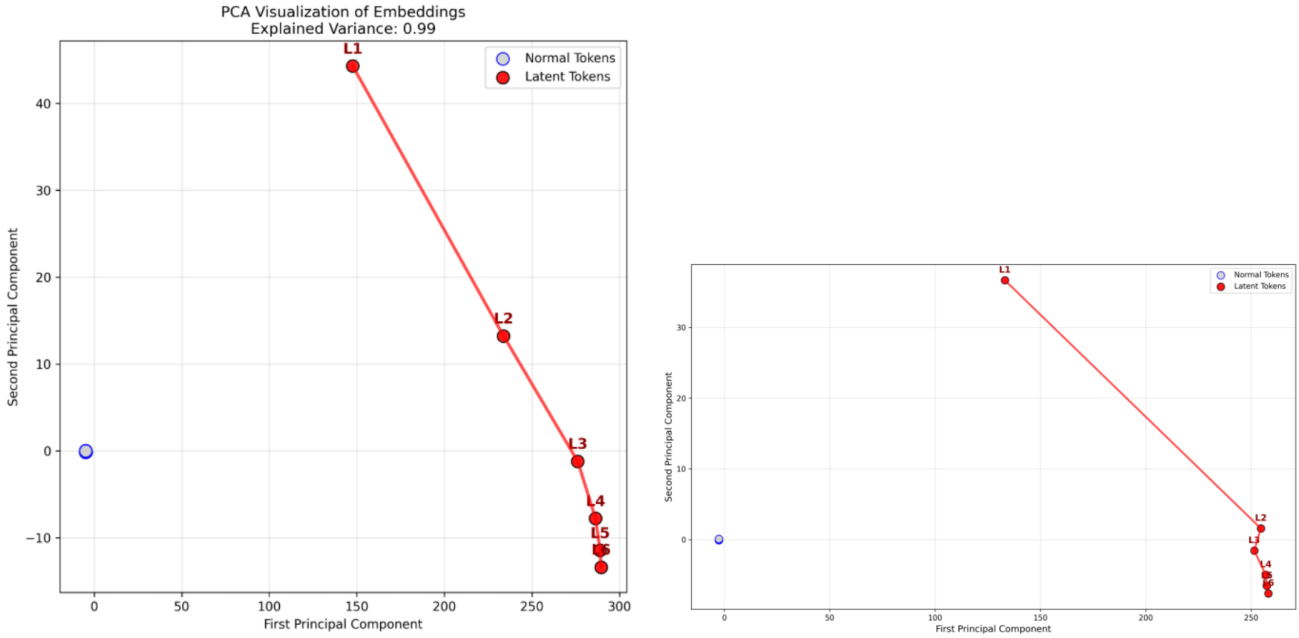
$$D_{\mathrm{KL}}(q_{\mathrm{PR}}\|p_{\mathrm{CoT}}) = \frac{B-1}{B}\log\kappa - \log B - \frac{(B-1)\log c}{B} + \mathcal{O}\left(\frac{1}{\kappa}\right).$$

由于引理 A.2 已证明 $H(p_{\mathrm{CoT}}) \to 0$，且 $D_{\mathrm{KL}}(q_{\mathrm{PR}}\|p_{\mathrm{CoT}})$ 随 κ对数增长，这证明了思维链的探索不足。

□

A.4. 定理 4.5 的证明

假设 A.3 （潜在状态的收敛性与紧致性）。 我们假设，在推理过程中，随着推理步骤 k 增加，潜在状态序列 $\{h_k\}$ 会收敛到一个不动点，或进入一个紧致吸引集。这并非仅仅出于理论上的便利；这是我们通过实证持续观察到的一种行为。如图 5 所示，潜在思维嵌入的主成分分析可视化呈现出清晰的收敛轨迹，其中连续状态 $\{L1,\ldots,L6\}$ 在有界区域内逐渐相互靠近。前两个主成分较高的解释方差（0.99）证实了这一二维投影能够忠实地表示原高维空间中的动力学。



(a) GSM8K 任务上的主成分分析可视化。

(b) ProsQA 任务上的主成分分析可视化。

图 *5.* 推理轨迹的潜在状态嵌入主成分分析可视化。在 (a) 和 (b) 中，潜在思维（L1-L6）均呈现出明显的收敛趋势，支持了我们关于潜在状态存在紧致吸引集的假设。这验证了收敛动力学是潜在推理过程的一般属性，而非特定于单一任务。

引理 A.4（Coconut 的非退化输出分布）。 对于通过 *Coconut* 目标函数训练的模型，其优化过程可以建模为求解一个具有有限权衡参数 $\beta(k) > 0$ 的条件信息瓶颈（*CIB*）问题（定理 *4.1*）。在此框架下，模型的输出概率分布是非退化的：存在一个常数 $\delta > 0$，使得对于任意有效的下一个词元 u 和任意可达的潜在状态 h，其概率都严格低于 *1*：

$$\sup_{h \in \mathcal{H}} \left(\max_{u \in N_{valid}(v)} p(u \mid h) \right) \leq 1 - \delta,$$

其中 $\mathcal{H}$ 是由假设 *A.3* 保证的紧致潜在状态空间。

证明。 反设 $\sup_{h \in \mathcal{H}} \max_u p(u \mid h) = 1$。由于 $\mathcal{H}$ 是紧致的，且解码器 $p(u \mid h)$ 连续，因此必然存在一个潜在状态 $h^* \in \mathcal{H}$ 和一个词元 u^*，使得 $p(u^* \mid h^*) = 1$。
在 h^*，未来不确定性已被完全消除： $H(S^{(k+1\ldots M)} \mid h^*, X) = 0$，这意味着预测信息 $I(h; S^{(k+1\ldots M)} \mid X)$ 达到了理论最大值。
根据定理 4.1，Coconut 解对应于 CIB问题的一个最优解。在信息平面（$I_{\text{past}}, I_{\text{future}}$）上，所有最优解都必须位于效率前沿上，且任意最优点 h_{opt} 都必须满足：

$$\left. \frac{dI_{\text{future}}}{dI_{\text{past}}} \right|_{h_{\text{opt}}} = \frac{1}{\beta(k)}.$$

在我们的假设点 h^*， I_{future} 达到了其最大值 $I_{\max}(k)$。根据定理 4.1 中采用的效率前沿模型：

$$I_{\text{future}}(I_{\text{past}}; k) = I_{\max}(k) \left(1 - e^{-\alpha(k) \cdot I_{\text{past}}}\right),$$

它关于 I_{past} 严格递增且为严格凹函数，其导数为：

$$\frac{dI_{\text{future}}}{dI_{\text{past}}} = I_{\max}(k) \cdot \alpha(k) \cdot e^{-\alpha(k) \cdot I_{\text{past}}}.$$

当 $I_{\text{future}} \to I_{\max}(k)$ 时，我们必须有 $I_{\text{past}} \to \infty$，且导数 $\to 0^+$。如果 I_{future} 在有限的 I_{past} 处达到 $I_{\max}(k)$（如 h^*处所示），则导数必须为零，才能符合渐近行为。因此，在 h^*处：

$$\left.\frac{dI_{\text{future}}}{dI_{\text{past}}}\right|_{h^*} = 0.$$

如果 h^* 是 CIB 最优点，则其必须满足：

$$\left.\frac{dI_{\text{future}}}{dI_{\text{past}}}\right|_{h^*} = \frac{1}{\beta(k)},$$

这意味着 $1/\beta(k) = 0$，即 $\beta(k) \to \infty$。然而，根据定理 4.1，在课程学习的任何非终止阶段（$k < M$），$\beta(k)$ 都是有限的正常数（其渐近行为为 $\beta(k) \sim k/(M-k) < \infty$），因此 $1/\beta(k) > 0$。这与斜率为零的要求相矛盾。

因此，假设 $p(u^* \mid h^*) = 1$ 会导致矛盾。由此，$\sup \max p(u \mid h)$ 必须严格小于 1。引理得证。

直觉： 本质上，Coconut 通过设计从根本上避免了确定性坍缩：其信息瓶颈目标施加了一种权衡，从而保留输出的多样性，即使在高级推理阶段也是如此。 □

现在开始证明定理 4.5

证明。 根据引理 A.4，存在 $\delta > 0$，使得对于任意有效的下一个词元 u 和可达状态 h，概率满足 $\max_u p(u \mid h) \le 1 - \delta$。我们的目标是推导 KL 散度 $D_{\text{KL}}(q_{\text{PR}} \| p_{\text{Coconut}})$ 的上界。
根据定义，

$$D_{\text{KL}}(q_{\text{PR}} \| p) = -\log B - \frac{1}{B} \sum_u \log p(u \mid v).$$

为了对该量求上界，我们要寻找在约束 $\sum p_i = 1$和 $\max\ p_i \le 1 - \delta$下使 D_{KL} 最大化的概率分布，这等价于最小化 $\sum \log\ p(u \mid v)$。
The $\sum \log\ p_i$ 的最小值由满足约束的最集中分布取得 t:

$$p(u^* \mid v) = 1 - \delta, \quad \text{and for all } u \neq u^*, \quad p(u \mid v) = \frac{\delta}{B-1}.$$

S将该最坏情况分布代入 KL 散度表达式，可得上界 nd:

$$\begin{aligned} D_{\text{KL}}(q_{\text{PR}} \| p_{\text{Coconut}}) &\le -\log B - \frac{1}{B}\left[\log(1-\delta) + (B-1)\log\left(\frac{\delta}{B-1}\right)\right] \\ &= -\frac{B-1}{B}\log\delta - \left[\log B - \frac{B-1}{B}\log(B-1) + \frac{\log(1-\delta)}{B}\right]. \end{aligned}$$

将括号中的项记为 $f(B) = \log\ B - \frac{B-1}{B}\log(B-1) + \log(1-\delta)$。由于 $f(B)$ 在 $[2, \infty)$ 上连续且有界（且 $\lim_{B\to\infty} f(B) = 0$），其下确界 $c = \inf_{B \ge 2} f(B)$ 是有限常数。因此，$-f(B)$ 有上界 $-c$。

同时，由于 $B \ge 2$ $\frac{B-1}{B} \ge \frac{1}{2}$ $\delta \in (0, 1)$ $\delta < 0$ $-\frac{B-1}{B}$ $\delta \le$，我们有 。此外，由于 ()，可知 $-\frac{1}{2}$ δ log，因此 log log 。

综合这些界，KL 散度具有一个与分支因子 B 无关的有限上界：

$$D_{\text{KL}}(q_{\text{PR}} \| p_{\text{Coconut}}) \le -\frac{B-1}{B}\log\delta - c \le -\frac{1}{2}\log\delta - c.$$

□

A.5. 定理 4.7 的证明

ˆ证明。 我们证明，在亚决策扰动下，ExplicitCoT 轨迹 S 与无噪声轨迹S^* 相同，其方法是对推理步骤 k使用归纳法。

令 f_θ(·) 表示模型的确定性前向传播。在步骤 k，模型将输入 x 和离散历史 $S^{(1\dots k-1)}$ 作为输入，生成 logit向量 $l_k^* = f_\theta(x,\ S^{(1\dots k-1)})$。无噪声词元选择由 $s_k^* = \text{argmax}(l_k^*)$ 给出。受扰动过程生成 logit $l_k = l_k^* + \epsilon_k$，其中 ϵ_k 满足亚决策级条件：$\text{argmax}(l_k^* + \epsilon_k)$ =$\text{argmax}(l_k^*)$。

在第一步，干净过程和受扰动过程接收完全相同的输入 x （历史为空）。干净的 logit 为 $l_1^* = f_\theta(x,\ \emptyset)$。受扰动的 logit 为 ˆ $l_1 = l_1^* + \epsilon_1$。根据亚决策级扰动的定义：

$$\hat{s}_1 = \text{argmax}(\hat{l}_1) = \text{argmax}(l_1^* + \epsilon_1) = \text{argmax}(l_1^*) = s_1^*.$$

因此，第一个词元相同，传递到步骤 2 的历史没有误差。

假设对于所有步骤 $t < k$，生成的词元都相同，即 $\hat{s}_t = s_t^*$。这意味着离散历史相同：ˆ$S^{(1\dots k-1)} = S^{*(1\dots k-1)}$。在步骤 k，由于模型 f_θ 是确定性的，且严格以离散历史为条件：

$$\text{Logits}_{\text{pre-noise}} = f_\theta(x, \hat{S}^{(1\dots k-1)}) = f_\theta(x, S^{*(1\dots k-1)}) = l_k^*.$$

ˆ 噪声过程添加扰动： $l_k = l_k^* + \epsilon_k$。再次应用子决策约束，离散投影会滤除噪声：

$$\hat{s}_k = \text{argmax}(l_k^* + \epsilon_k) = s_k^*.$$

因此，噪　　声项 ϵ_k 会被 argmax 算子丢弃，且不会传播到下一步的状态　　p.

根据归纳法，对于所有 $k = 1$，.....， M，均有$\hat{s}_k = s_k^*$。因此，轨迹分歧概率为零：

$$\mathbb{P}[\hat{S} \neq S^*] = 0.$$

这从数学上证实了第 4.3 节所述的离散化-重置 机制：argmax 算子实际上充当非线性滤波器，在每一步都将内部状态重置到干净的整数网格上，从而防止亚决策噪声不断累积。

□

A.6. 定理 4.8 的证明

证明。 根据递推关系

$$E_k = f_\theta(h_{k-1}) - f_\theta(h_{k-1}^*) + \epsilon_h^{(k)},$$

取平方范数并求期望，再利用噪声的独立性和零均值性质（这会使交叉项消失），可得：

$$\mathbb{E}[\|E_k\|^2] \le L_F^2\, \mathbb{E}[\|E_{k-1}\|^2] + d\sigma_h^2.$$

This 这是一个非齐次线性递推关系。利用初始条件 $\mathbb{E}[\|E_0\|^2] = 0$ 求解，得到　　：

$$\mathbb{E}[\|E_M\|^2] = \begin{cases} \dfrac{1 - L_F^{2M}}{1 - L_F^2} \cdot d\sigma_h^2, & L_F \neq 1, \\ M \cdot d\sigma_h^2, & L_F = 1. \end{cases}$$

在两种情况下，对于所有 $M \ge 1$， $\mathbb{E}[\|E_M\|^2] > 0$ 均成立，由此完成证明。 □

A.7. 归一化准确率函数的推导

在本节中，我们将严谨推导**归一化准确率**曲线 $A(\sigma)$，以解释图 3 中观察到的典型反 S 形（类似 sigmoid）退化。我们对从潜在空间中的注入噪声到输出空间中词元排序发生概率性失效的因果链进行建模。

我们首先对隐藏状态中的噪声累积进行建模。设 h_M^* 表示最终推理步骤 M中的干净潜在状态。根据定理 4.8（误差累积），在 $\epsilon \sim \mathcal{N}(0,\ \sigma^2 I_d)$处注入的高斯噪声

会在每个步骤中沿推理轨迹不断累积。假设转移动力学可用线性化近似表示，则$\tilde{}$最终受扰状态 h_M 可表示为：

$$\tilde{h}_M = h_M^* + \epsilon_{\text{acc}}, \quad \text{where} \quad \epsilon_{\text{acc}} \sim \mathcal{N}(0, \sigma_{\text{eff}}^2 I_d).$$

其中，ϵ_{acc} 表示 $\mathbb{R}^d$中的有效累积噪声向量。方差 $\sigma_{\text{eff}}^2 \approx \alpha \cdot \sigma^2$ 随注入方差 σ^2线性缩放，其中缩放因子 α 取决于路径长度 M 和网络的利普希茨常数。

由于离散词元决策并非在潜在空间中，而是在 logit 空间中做出，我们必须追踪这种扰动如何通过反嵌入层传播。令 $W_U \in \mathbb{R}^{|\mathcal{V}|\times d}$ 为语言模型头的权重矩阵。扰动后的 logit $\tilde{} \; l$ 通过投影最终潜在状态得到：

$$\tilde{l} = W_U \tilde{h}_M = W_U(h_M^* + \epsilon_{\text{acc}}) = l^* + \eta,$$

其中，l^* 表示无噪声的 logit，$\eta = W_U \epsilon_{\text{acc}}$ 表示投影到 logit 空间中的噪声。由于 ϵ_{acc} 服从高斯分布，且 W_U 是线性变换，η 在 $\mathbb{R}^{|\mathcal{V}|}$上服从多元高斯分布。

如果噪声不足以改变排名靠前的词元的排序，则预测成功。令 $i^* = \text{argmax}(l^*)$为正确词元的索引$\tilde{}$ $\tilde{}$，令 j^* 为最强竞争词元的索引。如果 $l_{i^*} > l_{j^*}$，模型就能保留正确预测。$\tilde{}$代入 $l = l^* + \eta$后，该条件变为：

$$l_{i^*}^* + \eta_{i^*} > l_{j^*}^* + \eta_{j^*} \implies \eta_{j^*} - \eta_{i^*} < l_{i^*}^* - l_{j^*}^*.$$

关键的是，右侧恰好就是定义 4.10 中定义的 Logit 决策间隔 Δ_l。因此，成功条件可简化为：

$$\xi < \Delta_l, \quad \text{where} \quad \xi = \eta_{j^*} - \eta_{i^*}.$$

为了量化该事件发生的概率，我们分析随机变量 ξ 的分布。我们将 ξ 表示为累积噪声向量的线性组合：$\xi = (w_{j^*} - w_{i^*})^\top \epsilon_{\text{acc}}$。由于 ϵ_{acc} 是各向同性高斯分布，$\xi \sim \mathcal{N}(0, \sigma_1^2)$。标量方差 σ_1^2 为：

$$\sigma_1^2 = \|w_{j^*} - w_{i^*}\|_2^2 \cdot \sigma_{\text{eff}}^2.$$

令 $C = \alpha \cdot \|w_{j^*} - w_{i^*}\|_2^2$，我们可推导出关系式 $\sigma_1^2 = C \cdot \sigma^2$。因此，$\sigma_1 = \sqrt{C \cdot \sigma}$。

归一化准确率 $A(\sigma)$ 定义为成功概率 $\mathbb{P}(\xi < \Delta_l)$。我们使用标准正态分布的累积分布函数 $\Phi(z)$ 对其进行计算。通过 $Z = \xi/\sigma_1 \sim \mathcal{N}(0, 1)$ 进行标准化后，得到：

$$\begin{aligned} A(\sigma) &= \mathbb{P}(\xi < \Delta_l) \\ &= \mathbb{P}\left(\frac{\xi}{\sigma_1} < \frac{\Delta_l}{\sigma_1}\right) \\ &= \Phi\left(\frac{\Delta_l}{\sqrt{C} \cdot \sigma}\right). \end{aligned} \tag{4}$$

这一推导为实验结果提供了严谨的物理解释。在 $\sigma \to 0$ 的极限下，论证$\frac{\Delta_l}{\sqrt{C}\sigma} \to \infty$，且 $A(\sigma) \to 1$。这解释了“亚决策平台期”：由于噪声严格处于安全裕度 Δ_l 之内，准确率保持稳定。随着 σ 增加，论证随之减弱，从而出现准确率曲线上典型的类 S 形衰减。这证实了鲁棒性在解析上由符号指数所导致的决策边际 Δ_l 与累积噪声方差之比决定。

A.8. 定理 4.11 的证明

证明。Logit 决策间隔 Δ_l 定义为概率最高与第二高的词元对应的 logit 之差，即 $\Delta_l = l_{i^*} - l_{j^*}$。我们的目标是基于符号指数 $\mathcal{I}_S$，为该量建立一个下界。

首先，我们使用 softmax 函数 $p_i = e^{l_i} / \sum_k e^{l_k}$ 将 logit 与概率联系起来。排名前两位的词元的概率之比为：

$$\frac{p_{i^*}}{p_{j^*}} = \frac{e^{l_{i^*}}}{e^{l_{j^*}}} = e^{l_{i^*} - l_{j^*}} = e^{\Delta_l}.$$

作者: taki

对两边取自然对数后，我们可以用概率表示 logit 边际 :

$$\Delta_l = \log(p_{i^*}) - \log(p_{j^*}).$$

根据符号指数的定义，最可能词元的概率为 $p_{i^*} = \mathcal{I}_S$。

接下来，我们为第二可能词元的概率 p_{j^*} 建立一个上界。所有可能词元的概率之和为 1。因此，除最可能词元之外的所有词元的概率之和为 $1 - p_{i^*} = 1 - \mathcal{I}_S$。由于 p_{j^*} 是这些剩余概率中最大的，因此它不会大于它们的总和。于是，我们得到以下不等式:

$$p_{j^*} \leq 1 - \mathcal{I}_S.$$

由于对数函数单调递增，这意味着 $\log(p_{j^*}) \leq \log(1 - \mathcal{I}_S)$。因此，其相反数 $-\log(p_{j^*})$ 满足反向不等式: $-\log(p_{j^*}) \geq -\log(1 - \mathcal{I}_S)$。

最后，我们将所得结果代回关于 Δ_l 的方程:

$$\Delta_l = \log(\mathcal{I}_S) - \log(p_{j^*}) \geq \log(\mathcal{I}_S) - \log(1 - \mathcal{I}_S).$$

利用对数的性质，我们得到所需的下界:

$$\Delta_l \geq \log\left(\frac{\mathcal{I}_S}{1 - \mathcal{I}_S}\right).$$

□

A.9. 定理 4.12 的证明

证明。 理想均匀先验 q_{PR}（其中 $q_i = 1/B$ 对所有 $i = 1$、.....、 B）与模型的输出分布 p之间的 KL 散度为:

$$D_{\text{KL}}(q_{\text{PR}} \| p) = \sum_{i=1}^{B} q_i \log\left(\frac{q_i}{p_i}\right) = \sum_{i=1}^{B} \frac{1}{B} \log\left(\frac{1/B}{p_i}\right) = -\log B - \frac{1}{B} \sum_{i=1}^{B} \log p_i.$$

为了求出 $D_{\text{KL}}(q_{\text{PR}} \| p)$ 的下界，我们必须在模型分布所施加的约束条件下，求出项 $\sum_{i=1}^{B} \log p_i$ 的上界。这些约束为:

1. $\sum_{i=1}^{B} p_i = 1$
2. $\max_i p_i = \mathcal{I}_S$

我们要求解以下优化问题:

$$\max_{p_1,\ldots,p_B} \sum_{i=1}^{B} \log p_i \quad \text{s.t.} \quad \sum p_i = 1 \text{ and } \max p_i = \mathcal{I}_S.$$

令 $p_{i^*} = \mathcal{I}_S$ 为最大概率。剩余 $B - 1$ 概率之和为 $\sum_{k \neq i^*} p_k = 1 - \mathcal{I}_S$。函数 $f(p_1, \ldots..., p_{B-1}) = \sum_{k \neq i^*} \log p_k$ 是凹函数。根据詹森不等式，当剩余概率尽可能均匀地分布时，也就是当所有 $k \neq i^*$ 都满足 $p_k = \frac{1 - \mathcal{I}_S}{B-1}$ 时，该和达到最大值。

——

在峰值为 $\mathcal{I}_S$ 的条件下，这种分布使概率尽可能“平坦”，从而使 $\sum \log p_i$ 取得可能的最大值。代入这个极值分布后，我们得到上界:

$$\max \sum_{i=1}^{B} \log p_i = \log(\mathcal{I}_S) + (B - 1) \log\left(\frac{1 - \mathcal{I}_S}{B - 1}\right).$$

现在，我们将这个上界代回 KL 散度的表达式中。由于我们要减去这一项，它的上界便给出了 KL 散度的下界:

$$D_{\text{KL}}(q_{\text{PR}} \| p) \geq -\log B - \frac{1}{B}\left[\log(\mathcal{I}_S) + (B - 1) \log\left(\frac{1 - \mathcal{I}_S}{B - 1}\right)\right].$$

证明至此完成。 □

A.10. 定理 5.1 的证明

证明。 我们通过构造一个反例来证明这一点。我们定义一个推理任务，其中有偏的训练分布的成功率严格低于专家。随后我们证明，最大似然估计量（MLE）会收敛到模仿这种次优行为的策略，从而导致永久性的性能差距。

1. 实例构造 考虑一个 $d = 3$维潜在空间，其中包含三种离散状态：

- $h_{\text{专家}}$：正确推理路径。特征 $\phi(h_{\text{专家}}) = [1, 0, 0]^\top$。
- $h_{\text{捷径}}$：表面捷径。特征 $\phi(h_{\text{捷径}}) = [0, 1, 1]^\top$。
- $h_{\text{不良}}$：无关噪声。特征 $\phi(h_{\text{不良}}) = [0, 1, 0]^\top$。

估值函数 $V(h)$（由 h 推导出的答案的正确性）定义如下：

$$V(h) = \begin{cases} 1 & \text{if } h = h_{\text{expert}} \\ 0 & \text{otherwise} \end{cases}$$

因此，参数化为 θ 的策略的**任务成功率** 为 $\mathcal{R}(\theta) = P_\theta(h_{\text{expert}})$。

2. 专家与偏置分布 设专家参数为 $\theta^* = [10, 0, 0]^\top$。专家状态的未归一化对数概率（得分）为 10，而其他状态为 0。专家的成功率为：

$$\mathcal{R}(\theta^*) = \frac{e^{10}}{e^{10} + e^0 + e^0} \approx 1.0.$$

现在，考虑有偏数据分布 P_{biased}。假设数据集完全由捷径推理构成，即$P_{\text{biased}}(h_{\text{shortcut}}) = 1$。该数据分布的成功率为：

$$\mathcal{R}(P_{\text{biased}}) = \mathbb{E}_{h \sim P_{\text{biased}}}[V(h)] = 1 \cdot V(h_{\text{shortcut}}) = 0.$$

这满足定理条件 $\mathcal{R}(P_{\text{biased}}) \leq \mathcal{R}(\theta^*) - \Delta$，间隔为 $\Delta \approx 1$。

ˆ3. 最大似然估计收敛与性能差距 我们在这个有偏数据集上训练模型 θ_{MLE}。最大似然估计会最大化观测数据的对数似然（h_{shortcut}）。为了最大化得分 $\langle \theta,\ \phi(h_{\text{shortcut}}) \rangle = \theta_2 + \theta_3$，同时遵循归一化约束，优化器会使 $\theta_2 + \theta_3 \to \infty$。

当 N 较大时，一个特征解是较大 K 下的 ˆ $\theta_{\text{MLE}} \approx [0, K, K]^\top$。在该学习到的策略下，概率质量完全转移到捷径：

$$P_{\hat{\theta}_{\text{MLE}}}(h_{\text{shortcut}}) \propto e^{2K}, \quad P_{\hat{\theta}_{\text{MLE}}}(h_{\text{expert}}) \propto e^0 = 1.$$

学习到的模型的成功率为：

$$\mathcal{R}(\hat{\theta}_{\text{MLE}}) = P_{\hat{\theta}_{\text{MLE}}}(h_{\text{expert}}) = \frac{1}{1 + e^{2K} + e^K} \xrightarrow{K \to \infty} 0.$$

4. 结论 比较最终成功率：

$$\mathcal{R}(\hat{\theta}_{\text{MLE}}) \approx 0 \quad \text{vs.} \quad \mathcal{R}(\theta^*) \approx 1.$$

因此，我们推导出了该界：

$$\mathcal{R}(\hat{\theta}_{\text{MLE}}) \leq \mathcal{R}(\theta^*) - C,$$

其中，$C \approx 1$ 是由初始差距 Δ导出的严格正常数。这证明，若不通过课程纠正分布偏差，模型的性能就会始终与最优性相距甚远。 □

A.11. 定理 5.2 的证明

模仿学习框架的合理性说明我们的理论分析，尤其是定理 5.2，在标准模仿学习（IL）框架下对训练过程进行建模。该框架假设可以访问从专家潜在策略$P_{\theta^*}(h|x)$中采样得到的输入—潜在状态对 $D_c = \{(x_i,\ h_i)\}_{i=1}^n$ 的独立同分布数据集。然而，Coconut 训练范式是通过专家轨迹 $D_S = \{(x_i,\ S_i^*)\}_{i=1}^n$来监督未来词元的预测。本节将正式论证：训练 Coconut 模型等价于在这一模仿学习设置下进行最大似然估计（MLE）。

核心论证如下：

1. **隐式专家潜在策略的定义：**根据定理 4.1（Coconut-CIB 对偶性），Coconut 训练的目标是学习一个编码器 E_θ，使得潜在状态 $h_k = E_\theta(S^{(1\ldots k)})$ 成为过去 $S^{(1\ldots k)}$ 关于预测未来 $S^{(k+1\ldots M)}$ 的**最小充分统计量**。令 E_{θ^*} 为解决该 CIB问题的最优编码器。我们可以将**隐式专家潜在策略** $P_{\theta^*}(h|x)$ 定义为：将最优编码器应用于专家轨迹前缀的分布后所诱导的分布：

$$h \sim P_{\theta^*}(h|x) \quad \text{where} \quad h = E_{\theta^*}(S^*_{\text{past}}) \text{ and } S^*_{\text{past}} \sim P_{\text{expert}}(S_{\text{past}}|x).$$

   该策略的支集 $P_{\theta^*}(h|x)$ 构成了理想潜在状态的空间。

2. **优化过程的等价性：**Coconut 目标函数旨在最小化期望负对数似然：

$$\min_\theta \mathcal{L}(\theta) = \mathbb{E}_{(x,S^*)\sim P_{\text{expert}}}[-\log p(S^*_{\text{future}}|E_\theta(S^*_{\text{past}}), x)].$$

   根据数据处理不等式，对于任意编码器 E_θ，我们有：

$$I(S^*_{\text{past}}; S^*_{\text{future}}|x) \geq I(E_\theta(S^*_{\text{past}}); S^*_{\text{future}}|x).$$

   最优编码器 E_{θ^*} 能够达到等号，因为它保留了所有预测信息。这意味着条件分布可以分解为 $p(S^*_{\text{future}}|S^*_{\text{past}}, x) = p(S^*_{\text{future}}|E_{\theta^*}(S^*_{\text{past}})\ x) = p(S^*_{\text{future}}|h^*, x)$。因此，损失 $\mathcal{L}(\theta)$ 实际上衡量的是由 E_θ 诱导的潜在分布与理想潜在分布 $P_{\theta^*}(h|x)$之间预测能力的差异。当且仅当诱导的潜在分布 $p_\theta(h|x)$ 收敛到$P_{\theta^*}(h|x)$时，损失才会最小化。

结论： Coconut 训练过程虽然在形式上以未来词元生成为监督目标，但其内在CIB目标函数会**促使**编码器生成的潜在状态分布 $p_\theta(h|x)$ 拟合隐式专家潜在策略$P_{\theta^*}(h|x)$。因此，在专家轨迹数据集 D_S 上最小化 Coconut 损失，在数学上等价于对由 $h_i^* = E_{\theta^*}(S^*_{i,\text{past}})$构造的隐式数据集 D_c 执行最大似然估计。这为将基于模仿学习的理论保证应用于 Coconut 模型提供了严谨的理论基础。

引理 A.5（向量鞅的自归一化界（（Abbasi-yadkori 等人，2011））） 。 设 $\{\mathcal{F}_i\}_{i=0}^n$ 为一个滤过。设 $\{g_i\}_{i=1}^n$ 为 $\mathbb{R}^d$ 中的随机向量序列，满足 g_i 是 $\mathcal{F}_i$-可测的，且 $\mathbb{E}[g_i \mid \mathcal{F}_{i-1}] = 0$。假设存在常数 $\sigma > 0$，使得对于任意单位向量 $v \in \mathbb{R}^d$，实值随机变量 $\langle v, g_i\rangle$ 在给定条件下是 σ-次高斯的，即，

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda\langle v, g_i\rangle) \mid \mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

设 V 为一个 $d \times d$ 正定矩阵。定义

$$V_n = V + \sum_{i=1}^n g_i g_i^\top \quad \textit{and} \quad G_n = \sum_{i=1}^n g_i.$$

那么，对于任意 $\delta \in (0,1)$，以下不等式以至少 $1-\delta$ 的概率成立：

$$\|G_n\|^2_{V_n^{-1}} \leq 2\sigma^2 \log\left(\frac{\det(V_n)^{1/2}\det(V)^{-1/2}}{\delta}\right).$$

引理 A.6 (参数估计的置信集构造)。 设 θ_{MLE} 是在课程数据集 $D_c = \{(x_i, h_i)\}_{i=1}^n$ 上通过最大化对数似然函数得到的参数估计值，其中数据是从专家分布P_{θ^*}中独立同分布采样的：

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \log P_\theta(h_i \mid x_i)$$

假设对数线性模型满足以下正则条件：

1. （有界特征）：对于所有 $(x,\ h)$，$\|\phi(x,\ h)\|_2 \le L$。

2. （强凸性）：期望负对数似然函数 $L(\theta) = \mathbb{E}_{x,\ h\sim P_{\theta^*}}[-\text{对数}\ P_\theta(h|x)]$ 关于参数空间上的参数 θ 是 $\lambda_{\min}$-强凸的。其海森矩阵（费 *Fisher* 信息矩阵）$\nabla^2 L(\theta) = I(\theta)$ 满足$I(\theta) \succeq \lambda_{\min} I$。

3. （次高斯得分）：真实参数处的得分函数 θ^*, ∇_θ 对数 $P_{\theta^*}(h \mid x)$是一个均值为零的 σ^2-次高斯随机向量。

4. （有界参数）：$\|\theta^*\|_2 \le B$，且优化是在紧集 $\mathcal{B} = \{\theta \in \mathbb{R}^d :\ \|\theta\|_2 \le B\}$内进行的。

则对任意 $\delta \in (0,\ 1)$，存在常数 $C > 0$（仅依赖于 L、σ、$\lambda_{\min}$、B），使得以至少$1-\delta$的概率，真实专家参数 θ^* 满足：

$$\left\|\hat{\theta}_{MLE} - \theta^*\right\|_{\hat{H}_n}^2 \le \beta_n^2(\delta)$$

其中：

- $\hat{H}_n = \frac{1}{n}\sum_{i=1}^n \nabla^2(-\log P_{\hat{\theta}_{\text{最大似然估计}}}(h_i \mid x_i))$ 是在$\hat{\theta}_{\text{最大似然估计}}$处计算的经验海森矩阵，
- $\beta_n^2(\delta) = C \cdot \frac{d_{\log(n)+\log(1/\delta)}}{n}$ 是置信集的半径平方。

换言之，置信集 $\Theta(\hat{\theta}_{MLE}) = \left\{\theta \in \mathbb{R}^d :\ \left\|\theta - \hat{\theta}_{MLE}\right\|^2_{\hat{H}_n} \le \beta_n^2(\delta)\right\}$ 以高概率覆盖真实参数 θ^*。

证明。 本证明结合一阶最优性条件、集中不等式和自归一化分析技术，构造一个非渐近的高概率界。

1. 记号与基本关系 令 $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i(\theta)$ 表示经验负对数似然，其中 $\ell_i(\theta) = -\log P_\theta(h_i|x_i)$。根据定义，$\theta_{\text{MLE}}$ 是参数空间 $\mathcal{B}$ 内 $L_n(\theta)$ 的极小化解。由于 θ^* 是期望风险 $L(\theta)$ 的极小化解（因为 KL 散度 $D_{KL}(P_{\theta^*}\|P_\theta) = L(\theta) - L(\theta^*)$ 非负），因此我们有 $\nabla L(\theta^*) = 0$。

2. 利用最优性与凸性 根据 θ_{MLE} 的最优性，我们有 $L_n(\theta_{\text{MLE}}) \le L_n(\theta^*)$。由于 $L_n(\theta)$ 是凸函数，我们可以推导出一个将参数误差与经验过程联系起来的基本不等式：

$$\begin{aligned} L(\hat{\theta}_{\text{MLE}}) - L(\theta^*) &\le L(\hat{\theta}_{\text{MLE}}) - L(\theta^*) - (L_n(\hat{\theta}_{\text{MLE}}) - L_n(\theta^*)) \\ &= (L - L_n)(\hat{\theta}_{\text{MLE}}) - (L - L_n)(\theta^*) \\ &= \frac{1}{n}\sum_{i=1}^n \Big(\mathbb{E}[\ell_i(\hat{\theta}_{\text{MLE}})] - \ell_i(\hat{\theta}_{\text{MLE}})\Big) - (\mathbb{E}[\ell_i(\theta^*)] - \ell_i(\theta^*)) \end{aligned}$$

根据假设2，$L(\theta)$ 是 $\lambda_{\min}$-强凸的，因此 $L(\hat{\theta}_{\text{MLE}}) - L(\theta^*) \ge \frac{\lambda_{\min}}{2}\|\hat{\theta}_{\text{MLE}} - \theta^*\|_2^2$。结合上述不等式可得：

$$\frac{\lambda_{\min}}{2}\|\hat{\theta}_{\text{MLE}} - \theta^*\|_2^2 \le \sup_{\theta\in\mathcal{B}} |(L - L_n)(\theta) - (L - L_n)(\theta^*)|$$

对右端的经验过程给出界，可以得到一个初步的 L_2-范数收敛界 $\mathcal{O}(\sqrt{d/n})$。然而，为了得到更尖锐的加权范数界，我们转向自归一化分析。

3. 应用自归一化界 我们直接分析由一阶条件导出的误差关系。令 $\hat{\Delta} = \theta_{\text{MLE}} - \theta^*$。对 $\nabla L_n(\theta_{\text{MLE}}) = 0$ 在 θ^* 附近作一阶泰勒展开，得到：

$$0 = \nabla L_n(\hat{\theta}_{\text{MLE}}) = \nabla L_n(\theta^*) + \int_0^1 \nabla^2 L_n(\theta^* + t\Delta)dt \cdot \Delta$$

重新整理后，我们得到：

$$\bar{H}_n \Delta = -\nabla L_n(\theta^*) \tag{5}$$

其中，$\bar{H}_n = \int_0^1 \nabla^2 L_n(\theta^* + t\Delta)dt$ 是连接 θ^* 与$\hat{}$ θ_{MLE}的线段上的平均海森矩阵。

分析的核心在于界定经验得分 $\nabla L_n(\theta^*)$的范数。令 $g_i = \nabla_\theta \ell_i(\theta^*) = \nabla_\theta(-\log P_{\theta^*}(h_i|x_i))$。由于 $\mathbb{E}[g_i|\mathcal{F}_{i-1}] = \nabla L(\theta^*) = 0$，序列 $\{g_i\}_{i=1}^n$ 是一个向量值鞅差序列。在我们的正则条件下（具体而言，假设3），$\{g_i\}$ 也满足条件次高斯性质。因此，我们可以应用引理 A.5 中的自归一化界。

将变量对应到引理 A.5：对于某个较小的正则化常数 $\lambda > 0$，我们设定 $V = \lambda I$，这可确保初始正定性。于是得到 $G_n = \sum_{i=1}^n g_i$ 和 $V_n = \lambda I + \sum_{i=1}^n g_i g_i^\top$。应用该引理可知，以至少 $1-\delta$ 的概率：

$$\left\| \sum_{i=1}^n g_i \right\|^2_{(\lambda I + \sum_{i=1}^n g_i g_i^\top)^{-1}} \leq 2\sigma^2 \log\left(\frac{\det(\lambda I + \sum g_i g_i^\top)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right).$$

在有界特征假设下，我们可以对行列式项进行界定，并简化该表达式。利用以下事实：$\sum g_i g_i^\top \approx n \cdot \mathbb{E}[gg^\top] = n \cdot I(\theta^*)$，以及行列式的性质，可以推导出：

$$\|\nabla L_n(\theta^*)\|^2_{(I(\theta^*))^{-1}} = \frac{1}{n^2} \left\| \sum_{i=1}^n g_i \right\|^2_{(I(\theta^*))^{-1}} \lesssim \frac{d\log(n) + \log(1/\delta)}{n} \tag{6}$$

其中 $\lesssim$ 隐去了对数因子以及依赖于模型参数的常数。

4. 连接误差界与梯度界 回到公式（5），我们有 $\Delta = -\bar{H}_n^{-1} \nabla L_n(\theta^*)$。我们希望对 $\|\Delta\|^2_{\hat{H}_n}$ 进行界定。通过矩阵集中不等式以及假设1和假设4，可以证明，经验海森矩阵 $\nabla^2 L_n(\theta)^-$ $\hat{}$ 在参数空间 $\mathcal{B}$ 上一致地集中于其期望 $I(\theta)$ 附近。这意味着，以高概率，H_n 和 H_n 都接近 $I(\theta^*)^-$。形式化地说，存在 $\epsilon_n \to 0$，使得以高概率有 $\|H_n - I(\theta^*)\|_{\text{op}} \leq \epsilon_n$ 和 $\hat{}\|H_n - I(\theta^*)\|_{\text{op}} \leq \epsilon_n$。因此，我们可以近似得到：

$$\begin{aligned}
\|\Delta\|^2_{\hat{H}_n} &= \Delta^\top \hat{H}_n \Delta \\
&= (\nabla L_n(\theta^*))^\top \bar{H}_n^{-1} \hat{H}_n \bar{H}_n^{-1} \nabla L_n(\theta^*) \\
&\approx (\nabla L_n(\theta^*))^\top (I(\theta^*))^{-1} \nabla L_n(\theta^*) \\
&= \|\nabla L_n(\theta^*)\|^2_{(I(\theta^*))^{-1}}
\end{aligned}$$

通过更严格地处理近似 $\bar{H}_n^{-1\hat{}} H_n \bar{H}_n^{-1} \approx (I(\theta^*))^{-1}$ 并控制误差项，再结合公式（6），最终可以证明：

$$\left\| \hat{\theta}_{\text{MLE}} - \theta^* \right\|^2_{\hat{H}_n} \leq C \cdot \frac{d\log(n) + \log(1/\delta)}{n}$$

其中常数 C吸收了所有依赖于 $L, \sigma, \lambda_{\min}, B$ 的项。至此完成该引理的证明。 □

现在我们可以证明本节的主定理。

证明。 证明分三步进行：首先，利用引理 A.6，界定模型分布 $P_{\hat{\theta}}$ 与专家分布 P_{θ^*}之间的 KL 散度；其次，使用 Pinsker 不等式将 KL 散度的界转换为全变差（TV）距离的界；最后，利用 TV 距离的性质，界定两种策略之间任务成功率的差异。

1. 界定 KL散度 我们的目标是界定 KL 散度 $D_{KL}(P_{\theta^*}\|P_{\hat{\theta}})$。对于我们所考虑的对数线性模型族，KL 散度的形式如下：

$$D_{KL}(P_{\theta^*}(\cdot|x)\|P_{\hat{\theta}}(\cdot|x)) = \mathbb{E}_{h\sim P_{\theta^*}(\cdot|x)}[\log P_{\theta^*}(h|x) - \log P_{\hat{\theta}}(h|x)]$$

在 θ^* 附近对 $\log P_{\hat{\theta}}(h|x)$ 进行二阶泰勒展开，我们得到：

$$\log P_{\hat{\theta}}(h|x) \approx \log P_{\theta^*}(h|x) + \langle \nabla_\theta \log P_{\theta^*}(h|x), \Delta\rangle + \frac{1}{2}\Delta^\top \nabla^2_\theta \log P_{\theta^*}(h|x)\Delta$$

其中 $\Delta = \theta - \theta^*$。将其代入 KL 散度表达式并取期望，我们注意到 $\mathbb{E}_{h\sim P_{\theta^*}}[\nabla_\theta \log P_{\theta^*}] = 0$，由此得到用费 Fisher 信息矩阵表示的 KL 散度二次近似：

$$\mathbb{E}_{x\sim\rho}[D_{KL}(P_{\theta^*}\|P_{\hat{\theta}})] \approx \frac{1}{2}\Delta^\top I(\theta^*)\Delta = \frac{1}{2}\|\hat{\theta} - \theta^*\|^2_{I(\theta^*)}$$

更严谨地说，可以证明该近似在高阶项可控的情况下成立。由于引理 A.6 表明，$\|\theta - \theta^*\|_{H_n}$ 以高概率较小，且 H_n 集中在 $I(\theta^*)$ 附近，因此我们可以将引理 A.6 的结果转化为 KL 散度的一个界。其概率至少为 $1-\delta$：

$$\mathbb{E}_{x\sim\rho}[D_{KL}(P_{\theta^*}\|P_{\hat{\theta}})] \le C' \cdot \|\hat{\theta} - \theta^*\|^2_{\hat{H}_n} \le C \cdot \frac{d\log n + \log(1/\delta)}{n} \tag{7}$$

其中 C' 和 C 是正的常数。

2. 从 KL 散度到全变差距离全变差（TV）距离定义为 $TV(P,Q) = \frac{1}{2}\int |p(x) - q(x)|dx$。Pinsker 不等式确立了 KL 散度与 TV 距离之间的关系：

$$TV(P,Q) \le \sqrt{\frac{1}{2}D_{KL}(P\|Q)}$$

将该不等式应用于我们的模型和专家分布，并结合公式（7），我们得到 TV 距离的高概率界：

$$\begin{aligned}
\mathbb{E}_{x\sim\rho}[TV(P_{\theta^*}\|P_{\hat{\theta}})] &\le \mathbb{E}_{x\sim\rho}\left[\sqrt{\frac{1}{2}D_{KL}(P_{\theta^*}\|P_{\hat{\theta}})}\right] \\
&\le \sqrt{\frac{1}{2}\mathbb{E}_{x\sim\rho}[D_{KL}(P_{\theta^*}\|P_{\hat{\theta}})]} \quad \text{(by Jensen's inequality)} \\
&\le \sqrt{\frac{C}{2}\cdot\frac{d\log n + \log(1/\delta)}{n}} \\
&= \mathcal{O}\left(\sqrt{\frac{d\log n + \log(1/\delta)}{n}}\right)
\end{aligned} \tag{8}$$

3. 从全变差距离到任务成功率 TV 距离的一个关键性质是，它能够界定任意有界函数期望值之差。我们的任务成功函数 $V(h) \in \{0,1\}$ 是一个有界函数（其界为 1）。因此，

$$\begin{aligned}
|\text{SuccessRate}(\hat{\theta}) - \text{SuccessRate}(\theta^*)| &= \left|\mathbb{E}_{x\sim\rho,h\sim P_{\hat{\theta}}}[V(h)] - \mathbb{E}_{x\sim\rho,h\sim P_{\theta^*}}[V(h)]\right| \\
&= \left|\mathbb{E}_{x\sim\rho}\left[\mathbb{E}_{h\sim P_{\hat{\theta}}}[V(h)] - \mathbb{E}_{h\sim P_{\theta^*}}[V(h)]\right]\right| \\
&\le \mathbb{E}_{x\sim\rho}\left|\mathbb{E}_{h\sim P_{\hat{\theta}}}[V(h)] - \mathbb{E}_{h\sim P_{\theta^*}}[V(h)]\right| \\
&\le \mathbb{E}_{x\sim\rho}[\sup_{H'\subseteq H}|P_{\hat{\theta}}(H') - P_{\theta^*}(H')|] \\
&= \mathbb{E}_{x\sim\rho}[TV(P_{\theta^*}\|P_{\hat{\theta}})]
\end{aligned} \tag{9}$$

代入 b将式 (8) 中的界代入式 (9)，最终得到成功率之差的界：

$$|\text{SuccessRate}(\hat{\theta}) - \text{SuccessRate}(\theta^*)| \leq \mathcal{O}\left(\sqrt{\frac{d\log n + \log(1/\delta)}{n}}\right)$$

由于我们关注的是成功率的下界，这等价于：

$$\text{SuccessRate}(\hat{\theta}) \geq \text{SuccessRate}(\theta^*) - \mathcal{O}\left(\sqrt{\frac{d\log n + \log(1/\delta)}{n}}\right)$$

这一结果表明，在课程训练下，学习到的策略的任务成功率以 $\mathcal{O}(\sqrt{(d\log\ n)/n})$ 的速率高概率收敛到专家策略的任务成功率。证明至此完成。 □