可证明且实用的自我改进上下文策略优化

Tianrun Yu[1*], **Yuxiao Yang**[2*], **Zhaoyang Wang**[2], **Kaixiang Zhao**[1], **Porter Jenkins**[1], **Xuchao Zhang**[3], **Chetan Bansal**[3], **Huaxiu Yao**[2], **Weitong Zhang**[2]杨百翰大学, [2]北卡罗来纳大学教堂山分校, [3]微软
*贡献相同。tianruny@byu.edu, {yxyang, weitongz}@unc.edu

arXiv:2603.01335v1 [cs.LG] 2 Mar 2026

摘要

我们研究测试时扩展，即模型在推理过程中通过多轮自我反思来改进答案。我们提出上下文策略优化（ICPO），其中智能体在上下文中利用自评或外部观测到的奖励优化其响应，而无需修改参数。为解释这一 ICPO 过程，我们从理论上证明，在新颖的费舍尔加权对数几率匹配目标下进行充分预训练后，单层线性自注意力模型能够可证明地模仿用于线性老虎机的策略优化算法。基于这一理论，我们提出最小熵 ICPO（ME-ICPO），这是一种实用算法，可在推理时迭代利用模型的响应及其自评奖励，在上下文中不断完善响应。ME-ICPO 通过选择熵最低的响应及其奖励，利用多数投票确保自评奖励的稳健性。在标准数学推理任务中，与其他推理时算法相比，ME-ICPO 在保持推理成本可承受的同时，取得了有竞争力的顶尖表现。总体而言，ICPO 为大语言模型中的自我反思提供了有原则的理解，并为数学推理的测试时扩展带来实际益处。

1 引言

近年来，大语言模型（LLM）在数学推理（Yang 等，2024a；Wei 等，2022）、问题求解（Rein 等，2024；Zhou 等，2024）和工具使用（Yao 等，2023b）方面的能力不断增强。在这些新兴能力中，测试时扩展发挥着重要作用：大语言模型能够通过多轮自我反思逐步改进回答，而无需进行参数更新。测试时扩展已展现出强大的能力，使大语言模型能够执行后训练搜索（Yao 等，2023a；Besta 等，2024）、自我反思与自我奖励（Madaan 等，2023；Shinn 等，2023；Lightman 等，2023），以及思维链（CoT，Wei 等，2022）。这一过程的关键在于模型能否消化上下文中的信息，从而改进回答。这些上下文信息可以是先前的回答及用户提供的微调指令，也可以是附有自评奖励的思维链过程。然而，尽管这一现象已得到反复的经验验证，文献对这种上下文内自我改进背后的机制仍缺乏深入探究。现有研究（Park 等，2024）通常将大语言模型能够在内部进行后验采样或策略优化视为既定能力，却没有解释这种能力为何会在预训练过程中涌现。

另一方面，近期工作试图理解监督学习中的上下文学习（例如，线性回归 Zhang 等人，2024b; Garg 等人，2022）以及强化学习（例如，时序差分学习 Wang 等人，2024），这些研究表明，一些精心设计的 Transformer 在充分预训练后能够学习这些算法。然而，这些工作大多考虑让大语言模型根据输入预测输出，而文献中对于 Transformer 如何学习优化其行为 $\mathbf{x}$（即通过优化其策略以最大化响应 y）仍缺乏理解。此外，当前对上下文学习的理论理解与上下文测试时扩展的经验实现之间存在巨大差距。有鉴于对上下文策略优化的理论理解不足，以及如何在推理任务的测试时扩展中迭代利用这些上下文内信息方面的缺失，我们想问：

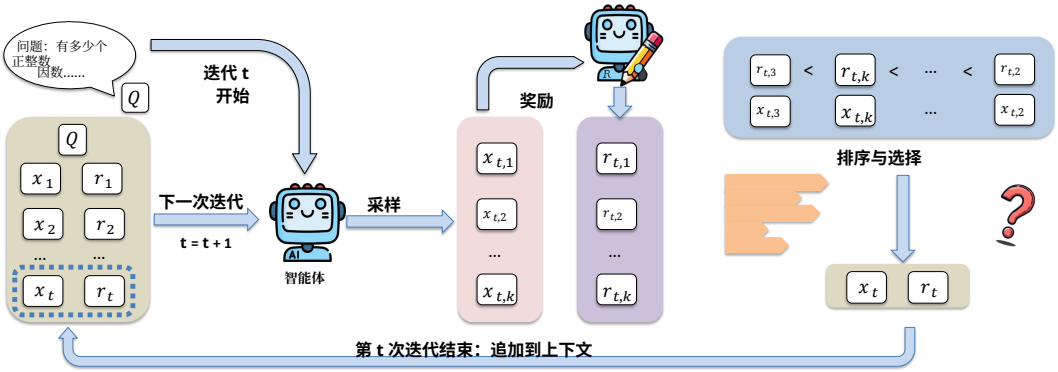


图 1：上下文策略优化（ICPO）框架。在每一轮 t中，智能体利用其过去尝试的历史记录及老虎机反馈 $\{(\mathbf{x}_1, r_1) \ldots, (\mathbf{x}_t, r_t)\}$ 来改进其响应 $\mathbf{x}_{t+1}$，以最大化获得的奖励 r_t。

我们能否从启发推理测试时扩展的上下文学习出发，理解大语言模型的自我反思过程？

本文通过提出上下文策略优化（ICPO）框架，对这一问题给出了肯定的回答。该框架关注大语言模型如何利用上下文动作和响应来改进其响应$\mathbf{x}$，而非预测某些特定结果。如图 1所示，ICPO 过程考虑 Transformer（大语言模型）生成响应 $\mathbf{x}_t$，并接收用户给出的奖励或自我评估奖励 y_t，随后通过生成 $\mathbf{x}_{t+1}$来改进响应。理论上，我们证明，经过充分预训练后，单层 Transformer 足以执行策略优化框架，并利用观测到的奖励逐步改进其响应$\mathbf{x}$ r。将 ICPO 框架应用于实际数学推理时，我们通过经验结果表明，该框架足够稳健，能够将自获取奖励纳入策略优化过程，并逐步改进响应。结合关于 ICPO 过程的扎实理论结果，以及精心设计的实用算法 ME-ICPO，我们提出了一个可证明且实用的上下文学习框架，用于数学推理的测试时扩展。我们的贡献总结如下：

• 我们将多轮自我反思机制表述为上下文策略优化（ICPO）框架，其中智能体根据收到的反馈生成并改进其响应。ICPO 框架将当前的上下文学习框架从监督学习扩展到带有老虎机反馈的策略优化。ICPO 为理解大语言模型推理中的自我反思与自我改进奠定了理论基础。

• 我们证明，在线性自注意力（LSA）Transformer 的显式设计下，当 LSA 在由特定策略优化框架生成的轨迹上得到充分预训练时，即使面对先前未见的奖励函数，它也能可证明地模拟底层策略优化。据我们所知，这是首个直接推导得出的、对上下文策略优化给出详细结构刻画的机制性解释。

• 在实验上，我们提出 ME-ICPO，这是一种植根于我们理论的实用算法，在数学推理任务上相较基础模型取得了显著改进。ME-ICPO 表明，ICPO 框架可以利用自我评估反馈，并通过熵正则化响应选择来确保稳健的策略更新。

总而言之，我们的工作展示了首个上下文优化机制，有助于理解大语言模型如何通过自我反思改进回答，并在各种数学推理任务中取得了出色的经验值表现。

符号记法。在本文中，我们使用普通字母（如 x）表示标量，使用小写粗体字母（如$\mathbf{x}$）表示向量，并使用大写粗体字母（如 $\mathbf{A}$）表示矩阵。函数由粗体符号表示，如 $\boldsymbol{f}$。集合和类由书法体表示，如 $\mathcal{F}$。对于向量 $\mathbf{x}$，$\|\mathbf{x}\|_2$ 表示其 ℓ_2-范数。对于矩阵 $\mathbf{A}$，$\|\mathbf{A}\|_{\mathrm{op}}$ 表示其算子（谱）范数，即 $\|\mathbf{A}\|_{\mathrm{op}} := \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \sigma_{\max}(\mathbf{A})$。$\mathbf{F}(\mathbf{p}) := \text{对角矩阵}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$。对于向量 $\mathbf{x} \in \mathbb{R}^K$，$\text{对角矩阵}(\mathbf{x}) \in \mathbb{R}^{K\times K}$ 表示对角矩阵，其 $[\text{对角矩阵}(\mathbf{x})]_{ii} = x_i$，且非对角元为零。对于正整数 N，我们使用 $[N]$ 表示 $\{1, 2, \ldots, N\}$。

2 相关工作

We 本节介绍测试时扩展、自我反思和上下文学习方面的工作。

测试时扩展。 测试时扩展（或推理时扩展）指的是这样一种现象：在测试时分配更多资源可以提升大语言模型的推理能力，并且已在实践中得到广泛应用（Jaech 等人，2024年；Guo 等人，2025）。最早的测试时扩展可追溯至少样本思维链（Wei 等人，2022年），其中提供的少样本示例能够提升大语言模型的推理能力。其他测试时扩展研究则侧重于后训练搜索算法，包括蒙特卡洛树搜索（Zhang 等人，2024a）、Best-of-N（Wang 等人，2022年；Huang 等人，2025）、思维树（Yaoet al.，2023a）。在这些研究的基础上，TTRL（Zuo 等人，2025）基于测试时的自我评估进行基于梯度的更新，并通过更新参数来改进大语言模型的回答。

自我反思与自我评估。测试时扩展的核心在于自我反思与自我评估，即大语言模型通过自我奖励来评估自身的回答（Madaan 等，2023；Shinn 等，2023）。具体而言，大语言模型评审和多数评判（MJ）能够以较低成本提供偏好信号，但这些信号也带有噪声；自一致性则可以转化为偏好或排序（Prasad 等，2024）。然而，自我评估容易受到提示词／位置敏感性和风格偏差的影响，因此需要进行校准（采用对称提示、位置打乱以及可执行性／格式检查）（Wang 等，2025）。与此同时，过程监督（过程奖励模型（Wang 等，2023a；Chen 等，2024b）、逐步验证器（Lightman 等，2023））将监督重点从结果转向中间步骤，从而能在多个轮次中可靠地过滤错误（Lightman 等，2023）。近期对生成—验证（GV）差距的分析表明，只有当验证显著比生成更容易时，迭代改进才能奏效，这也说明了鲁棒过滤和反馈的重要性（Song 等，2024）。除了这些经验研究，近期的理论研究还关注大语言模型的后验采样（Bai 等，2023；Von Oswald 等，2023），直接假设大语言模型具备估计后验分布的能力。

上下文学习与上下文强化学习。除了经验研究方面的进展，一系列理论工作将回归确立为研究上下文学习的核心试验场。对于岭线性回归，经过训练的线性自注意力能够在上下文中实现预条件梯度下降，其模型深度与隐式更新步骤的数量相匹配，并且在标准假设下呈几何收敛（Von Oswald 等人，2023年）。从训练动态的角度来看，已有研究表明，训练完成后，单头线性注意力层实际上会在上下文最小二乘目标上执行一步梯度下降（Zhang 等人，2024b）。在稀疏场景中，多头构造可以恢复稀疏信号，并在上下文中执行稀疏线性回归（Chen 等人，2024a）；近期研究进一步揭示了一种分层机制：第一层注意力头对上下文进行预处理，后续层则执行简单的迭代优化，二者共同带来优于朴素梯度下降和岭回归基线的超额风险保证（Chen 等人，2024a）。除了这些旨在理解回归的研究，近期工作还推动了对上下文学习作为一种元强化学习过程的理解。具体而言，（Lin 等人，2023年）证明，多层 Transformer 结构可以通过在轨迹上进行预训练来模仿老虎机／强化学习式更新；而（Wang 等人，2024年）表明，上下文学习中的线性回归可以扩展到强化学习中使用的时序差分学习。尽管如此，近期鲜有文献探讨策略优化，即根据历史信息直接优化输出 $\mathbf{x}$ 。

3 预备知识

我们考虑一种多臂老虎机抽象，它是序贯决策的标准理论框架。我们考虑一个 K 臂老虎机，在每一轮 $t \in [T]$ 中，智能体选择一个动作 $A_t \in [K]$ ，并执行相应的动作，该动作表示为独热向量 $\mathbf{x}_t = \mathbf{e}_{A_t} \in \mathbb{R}^K$。随后，智能体会收到一个标量奖励 r_t ，该奖励由一个未知任务向量$\mathbf{w} \in \mathbb{R}^K$ 的线性模型生成，其形式为 $r_t = \langle \mathbf{w}, \mathbf{x}_t \rangle + \epsilon_t$，其中 ϵ_t 是一个均值为零的 σ_ξ 次高斯随机变量。智能体的总体目标是通过最大化期望回报 $\langle \mathbf{w}, \mathbf{x}_t \rangle$来优化其策略 $\mathbf{x}_t$ 。

策略优化框架。 我们认为，预训练数据集是通过元强化学习中的策略优化过程生成的。具体而言，我们从一种类似于跟随正则化领导者（FTRL，Shalev-Shwartz 2007；McMahan2011a）框架的镜像下降方法开始，其形式为

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p} \in \Delta^K} \sum_{s=1}^{t} \left\langle \frac{r_s}{p_{s,A_s}} \mathbf{x}_s, \mathbf{p} \right\rangle - R(\mathbf{p}),$$

在 ICPO 框架中，我们考虑一种实用的方法来优化由$\mathbf{s} \propto \log \mathbf{p}$定义的策略的对数似然。在本文接下来的部分中，我们将策略优化写作

$$\mathbf{s}_{t+1} = \arg\max_{\mathbf{s}\in\mathbb{R}^K} \sum_{s=1}^{t} \langle r_s \mathbf{x}_s, \mathbf{s}_t \rangle - \lambda \sum_{s=1}^{t} \langle \mathbf{x}_s, \mathbf{s}_t \rangle - \frac{1}{2\eta_t} \mathbf{s}^\top \mathbf{H} \mathbf{s}, \tag{3.1}$$

其中，我们实现熵正则项 $R(\mathbf{p}) \approx \mathbf{s}^\top \mathbf{H} \mathbf{s}$，并使用拉格朗日乘子$r_s/p_{s,\ A_s}$替换 FTRL 中使用的无偏估计量 $\lambda \sum_{s=1}^{t} \langle \mathbf{x}_s,\ \mathbf{s}_t \rangle$ 以惩罚频繁访问的臂。方程 3.1 的闭式解在 $\mathbf{s}$ 上呈线性结构，可写为

$$\mathbf{s}_{t+1} = \eta_t (\mathbf{U}\mathbf{g}_t + \mathbf{V}\mathbf{n}_t), \text{ where } \mathbf{U} = \mathbf{H}^{-1}, \mathbf{V} = -\lambda \mathbf{H}^{-1}, \mathbf{g}_t = \sum_{s=1}^{t} r_s \mathbf{x}_s, \mathbf{n} = \sum_{s=1}^{t} \mathbf{x}_s \tag{3.2}$$

其中我们设定 $\eta_t = c/t$，优化后的策略是将 softmax 策略与γ-贪婪随机探索相结合

$$\mathbf{p}_{t+1} = \mathrm{softmax}(\mathbf{s}_{t+1}),\ \mathbf{p}_{t+1}^{\mathrm{PO}} = (1-\gamma)\mathbf{p}_{t+1} + \gamma \tfrac{\mathbf{1}}{K},\ \gamma \in [0, 1). \tag{3.3}$$

监督预训练数据生成。我们通过运行专家策略优化算法来生成预训练数据集。我们采样 B 条相互独立的轨迹。对于每条轨迹 $\tau \in [B]$，从先验分布中采样一个任务向量 $\mathrm{w}_\tau \sim \mathcal{N}(0, \tau_w^2 \mathrm{I}_K)$。随后执行教师 N 步，生成完整的交互历史 $\mathcal{H}_{\tau,N} = \{(\mathrm{x}_\tau,1, r_\tau,1) \ldots, (\mathrm{x}_{\tau,N}, r_{\tau,N})\}$，以及相应的专家 logit 向量序列 $\{\mathrm{s}_\tau^{\mathrm{PO}},_t\}_{t=1}^{N}$，该序列在方程 3.2 中更新。基于这些轨迹，我们构建一个监督训练数据集 $\mathcal{D}$，其中包含历史前缀与教师下一步逻辑值组成的配对。最终数据集由所有此类配对组成 $\mathcal{D} = \{(\mathcal{H}_{\tau,t}, \mathrm{s}_\tau^{\mathrm{PO}},_{t+1})\}_{\tau\in[B]}^{t\in[N-1]}$，其中$\mathcal{H}_{\tau,t}$ 是轨迹 τ 截至并包括第 t 轮的历史前缀。

对监督式预训练数据中的数据覆盖作出如下假设 $\mathcal{D}$，这是线性老虎机中的标准多样性假设（Papini 等人，2021；Hao 等人，2020年；Wu 等人，2020年）。

假设 3.1（数据覆盖与信号主导性）。我们假设，在预训练数据集中，任务 τ_w^2 和 FTRL 探索参数 γ 的覆盖范围足以覆盖奖励噪声。具体而言，定义学习率 $\eta_t = c/t$，我们假设覆盖率严格为正

$$c_\lambda := \tau_w \gamma / K - (1-\gamma) c \|\mathbf{U}\|_{\mathrm{op}} \sigma_\xi^2 / 2 > 0.$$

线性自注意力（LSA）。线性自注意力（LSA）是标准自注意力机制的一种简化变体，已被确立为分析 Transformer 和上下文学习的有用模型（Von Oswald 等人，2023年；Zhang 等人，2024b）。LSA 层接收一系列输入嵌入，以矩阵 $\mathbf{E} \in \mathbb{R}^{d\times N}$表示，其中 d 是嵌入维度，N 是序列长度。它通过以下计算形式生成一个维度相同的输出矩阵：

$$f_{\mathrm{lsa}}(\mathbf{E}; \boldsymbol{\theta}) = \mathbf{E} + \mathbf{W}^{PV}\mathbf{E} \cdot \left(\mathbf{E}^\top \mathbf{W}^{KQ} \mathbf{E}/\rho\right), \tag{3.4}$$

其中 $\boldsymbol{\theta} = (\mathbf{W}^{KQ}, \mathbf{W}^{PV})$ 是可学习参数（矩阵），ρ是注意力矩阵的归一化因子。该运算允许输入序列中的所有元素通过格拉姆矩阵项 $\mathbf{E}^\top \mathbf{W}^{KQ} \mathbf{E}$ 进行交互，从而更新初始嵌入。

4 ICPO的理论框架

在本节中，我们基于线性自注意力（LSA）网络中的启发性分析，为上下文策略优化提供理论依据。通过这一极简 LSA 模型，我们从理论上证明，预训练的 LSA 能够利用上下文内数据模仿专家策略优化算法。随后，我们给出主要理论保证，表明这种学习不仅在原则上可行，而且在有限数据下高效，并且对测试时的扰动具有鲁棒性。

ICPO 前向传播。 我们首先介绍 ICPO 框架的前向传播。由 $\boldsymbol{\theta}$ 参数化的 LSA 模型从一个空嵌入 $\mathbf{E}^{(0)} = (\mathbf{q}_x, q_r)^\top$ 开始，其中 $\mathbf{q}_x = \mathbf{1}_K, q_r = 0$ 是下一个词元生成的占位符。对于每个时间步 $t \in [T]$，LSA 模型根据下文所述的 LSA 下一个词元生成过程中的逻辑值更新来更新其策略

$$\widehat{\mathbf{s}}_t := \left[f_{\mathrm{LSA}}(\mathbf{E}^{(t-1)}; \boldsymbol{\theta})\right]_{1:K,\,t}, \qquad \mathbf{p}_t = (1-\gamma)\,\mathrm{softmax}\big(\widehat{\mathbf{s}}_t\big) + \tfrac{\gamma}{K}\mathbf{1},$$

其中 $\left[f_{\mathrm{LSA}}(\mathbf{E}^{(t-1)};\boldsymbol{\theta}^*)\right]_{1:K,t}$ 表示索引为 t 的新生成词元所对应的 K 维度。γ 与 FTRL 实现中使用的探索因子相同。采用这一策略 $\mathbf{p}_t \in \mathbb{R}^K$ 后，LSA 模型会选择其动作 A_t 并通过以下方式获得奖励

$$A_t \sim \mathbf{p}_t, \quad \mathbf{x}_t = \mathbf{e}_{A_t}, \quad r_t = \langle \mathbf{w}, \mathbf{x}_t \rangle + \epsilon_t.$$

Fin地，通过插入观测到的奖励 r_t 和嵌入 $\mathbf{x}_t$，更新词元序列

$$\mathbf{E}^{(t)} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_t & \boldsymbol{q}_x \\ r_1 & \cdots & r_t & q_r \end{pmatrix}$$

并将 LSA 中的归一化因子更新为 $\rho = t$，以确保注意力矩阵的上界为 1。随后，ICPO 的前向传播进入下一轮 $t \leftarrow t+1$。

训练目标。监督式预训练在数据集 $\mathcal{D}$ 上进行，通过最小化投影加权损失，使模型输出的逻辑值 $\widehat{\ }\mathbf{s}_t = f_{\mathrm{LSA}}(\mathbf{E}^{(t-1)}, \boldsymbol{\theta})$ 与策略优化的逻辑值 $\mathbf{s}_t^{\mathrm{PO}}$相匹配

$$\mathcal{L}(\boldsymbol{\theta}) = \tfrac{1}{2}\mathbb{E}_{\tau\in\mathcal{D}}\left[\sum_{t=1}^{N-1} \left\|\mathrm{Proj}(\widehat{\mathbf{s}}_{\tau,t+1} - \mathbf{s}^{\mathrm{PO}}_{\tau,t+1})\right\|^2_{\boldsymbol{\Gamma}}\right], \tag{4.1}$$

其中，投影算子 Proj 定义为 $\boldsymbol{\Pi}_\perp := \mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ 它从逻辑值中去除常数偏置 $\mathbf{1}^\top\mathbf{s}$，因为这种偏移不会影响策略 $\mathbf{p} \propto \exp(\mathbf{s})$。期望矩阵 $\boldsymbol{\Gamma}$ 的设计灵感来自自然策略优化（Kakade,2001），其定义如下

$$\boldsymbol{\Gamma} = \tfrac{1}{N-1}\mathbb{E}_{\tau\in\mathcal{D}}\left[\sum_{t=1}^{N-1}\mathrm{Diag}(\mathbf{p}_{\tau,t}) - \mathbf{p}_{\tau,t}\mathbf{p}_{\tau,t}^\top\right].$$

费舍尔加权损失为监督式预训练提供了一种新的损失函数。我们通过以下定理证明，预训练数据 $\mathbf{p}^{\mathrm{PO}}_{t+1}$ 与 LSA 输出 $\widehat{\ }\mathbf{p}^{t+1}$ 之间常见的 KL 损失被损失 $\mathcal{L}(\boldsymbol{\theta})$ 夹在中间，二者相差至多一个常数因子。

定理 4.1（混合策略 KL 由费舍尔投影二次损失控制）。假设教师和学生都采用 γ-混合探索，其中 $\gamma \in$（0，1），如方程 3.3所述，并令N 表示期望中样本的轨迹长度。那么，

$$\frac{(1-\gamma)^2}{4}\mathcal{L}(\boldsymbol{\theta}) \le \mathbb{E}\Big[\tfrac{1}{N-1}\textstyle\sum_{t=1}^{N-1} D_{\mathrm{KL}}\big(\mathbf{p}^{\mathrm{PO}}_{t+1} \,\|\, \widehat{\mathbf{p}}_{t+1}\big)\Big] \le \frac{K}{4\gamma}\mathcal{L}(\boldsymbol{\theta}).$$

定理 4.1 表明，广泛使用的 KL 损失是费舍尔加权损失的良好替代指标；它还解释了，即使在线性自注意力层中，使用 KL 损失也能使 Transformer 学会自我反思并改进其响应。

4.1 ICPO 的理论保证

现在，我们给出基于如下定义的经验 Fisher 加权损失的理论结果：

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}) := \tfrac{1}{2B(N-1)}\textstyle\sum_{\tau\in\mathcal{D}}^{t\in[T]} \left\|\mathrm{Proj}\big(\widehat{\mathbf{s}}_{\tau,t+1} - \mathbf{s}^{\mathrm{PO}}_{\tau,t+1}\big)\right\|^2_{\widehat{\boldsymbol{\Gamma}}}, \ \widehat{\boldsymbol{\Gamma}} := \tfrac{1}{B(N-1)}\textstyle\sum_{\tau\in\mathcal{D}}^{t\in[T]}\Big(\mathrm{Diag}(\mathbf{p}_{\tau,t}) - \mathbf{p}_{\tau,t}\mathbf{p}_{\tau,t}^\top\Big).$$

我们将总体优化器记为 $\boldsymbol{\theta}^* = \mathrm{argmin}_{\boldsymbol{\theta}}\,\mathcal{L}(\boldsymbol{\theta})$，其经验解 $\boldsymbol{\theta} = \widehat{\ }\arg\min \mathcal{L}(\boldsymbol{\theta})$。那么，第一个定理表明，总体优化器 $\boldsymbol{\theta}^*$ 恰好模仿了我们在方程 3.3 中描述的策略优化算法。

定理 4.2（总体等价性）。在假设 3.1下，考虑一个参数为 $\boldsymbol{\theta}^*$、使总体损失 $\mathcal{L}$（$\boldsymbol{\theta}$）最小化的单层 LSA，它将对所有可能的历史轨迹 $\mathcal{H}_t$ 模仿策略优化行为，即 $\widehat{\ }\mathbf{p}^{t+1}$（$\mathcal{H}_t$；$\boldsymbol{\theta}^\star$）$= \mathbf{p}^{\mathrm{PO}}_{t+1}$（$\mathcal{H}_t$）。

定理 4.2 表明，总体损失 $\mathcal{L}$（$\boldsymbol{\theta}$）足够精确且信息充分，能够保证参数 $\boldsymbol{\theta}^*$ 利用任意上下文内数据，驱动 LSA 精确地模仿策略优化框架。接着，简单的集中性分析表明，经验估计也能以高概率得到类似结果：

定理 4.3（有限样本结果）。在假设 3.1 下，令单层 LSA 在由式 3.3 中的策略优化过程生成的 B 独立同分布轨迹 $\{\mathcal{H}_{\tau,N}\}_{\tau=1}^B$ 上进行训练。定义$M := B(N-1)$。利用每条轨迹中的所有前缀 $t \in \{1,\ldots,N-1\}$（允许轨迹内的$\widehat{\ }$轨迹依赖），构造经验 Fisher 加权损失 $\mathcal{L}(\boldsymbol{\theta})$，并令全局优化器为$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \widehat{\mathcal{L}}(\boldsymbol{\theta})$。对于任意 $\delta \in (0,1)$，以至少 $1-\delta$的概率，若 $M \gtrsim t^2\,(2K + \log(1/\delta))/c_\lambda^2$,

那么对于任意固定的测试历史 $\mathcal{H}_t$，都有 $\widehat{\mathbf{p}}_{t+1}(\mathcal{H}_t;\widehat{\boldsymbol{\theta}}) = \mathbf{p}^{\mathrm{PO}}_{t+1}(\mathcal{H}_t)$。此外，期望行为不匹配程度的上界为

$$\mathbb{E}_{\mathrm{train}}\Big[\ \big\|\widehat{\mathbf{p}}_{t+1}(\mathcal{H}_t;\widehat{\boldsymbol{\theta}}) - \mathbf{p}^{\mathrm{PO}}_{t+1}(\mathcal{H}_t)\big\|_2^2\ \Big]\ \le\ 2(1-\gamma)^2\delta,$$

其中，$\mathbb{E}_{\mathrm{train}}$ 是对 B 轨迹的随机性取期望。

我们通过以下注释总结理论结果。

注释 4.4. 定理 4.3 表明，为保证 LSA 能够良好地模仿预训练策略优化轨迹，监督式预训练数据所需的样本复杂度为 $\widetilde{\mathcal{O}}\left(N^2K/c_\lambda^2\right)$。这种常数级样本复杂度源于假设 3.1，该假设与 Papini 等人（2021年）；Hao 等人（2020年）；Wu 等人（2020年）所采用的γ-贪心探索足以在线性老虎机中实现常数遗憾这一多样性假设相似。

与Lin 等人（2023年）；Park 等人（2024年）相比，我们的分析和框架有共同之处，也面临着显著的困难。

注释 4.5. Park 等人（2024年）在假设大语言模型能够进行后验采样的前提下，分析了大语言模型的遗憾，但未进行结构分析。相比之下，我们的分析建立在一个略作修改的策略优化框架之上；该框架受通过引入拉格朗日函数求解 FTRL 镜像下降的思路启发。借助新设计的监督式学习损失，我们证明，线性自注意力 Transformer 能够在结构上模仿策略优化框架。与Park 等人（2024年）提出的无监督损失相比，我们通过定理4.1 证明，新提出的费舍尔加权损失是实用 KL 损失的一种良好替代指标，从而更有力地说明 Transformer 在充分预训练后能够学会策略优化。

注释 4.6。 我们注意到，Lin 等人 (2023年) 研究了算法蒸馏，并证明 $\widetilde{\mathcal{O}}(\sqrt{T})$层 ReLU 激活变换器能够模拟（软）线性 UCB 和线性汤普森采样。相比之下，我们旨在理解策略优化，而非其基于价值的方法，这更适合分析大语言模型改进其响应的行为。此外，我们的分析基于单层线性自注意力框架，因此网络无需随上下文变长而改变，也更贴近实际的长上下文场景；而 Lin 等人 (2023年) 要求网络层数按 $\sqrt{T}$ 的量级增长，其中T是上下文序列的长度。

此外，正如先前文献中经常讨论的那样（McMahan，2011b；Shani 等人，2020年），FTRL 等策略优化方法已知能够抵御对抗性奖励或设定错误的奖励，这凸显了它们在采用自我评估、带噪或扰动奖励的 ICPO 中的适用性。实际自我改进的一项关键属性是，学习到的策略也应保持稳定。我们通过考察学习到的 ICPO 循环在测试时对单次奖励冲击的响应，分析其鲁棒性；相关内容将在下一条定理中介绍。我们将从定义 CRN 耦合设置（Glasserman 与 Yao，1992）开始。

定义 4.7（s-共同随机数耦合轨迹）。如果两条轨迹共享一个共同随机数（CRN），并且在轨迹 $\widetilde{\mathcal{H}_t}$ 和 $\mathcal{H}_t$ 中截至第 $s-1$ 轮之前完全相同，则称它们是 s-共同随机数耦合的。我们将 $\mathcal{F}_s$ 表示为在观察第 s 轮的奖励之前发生的所有事件所构成的过滤。在第 s 轮，奖励偏移了 δ_r。我们将由该奖励偏移引起的漂移定义为

$$\Delta\mathbf{p}^s_{t+1}(\boldsymbol{\theta}) := \widehat{\mathbf{p}}_{t+1}(\widetilde{\mathcal{H}}_t;\boldsymbol{\theta}^*) - \widehat{\mathbf{p}}_{t+1}(\mathcal{H}_t;\boldsymbol{\theta}^*) \in \mathbb{R}^K.$$

接下来，我们将给出以下定理，表明在精心设计的学习率 $\eta_t = c/t$ 下，任何一次性奖励扰动的影响都会逐渐减小。

定理 4.8（对单步奖励扰动的稳定性）。在假设 3.1下，假设测试时的 ICPO 循环以固定的总体最优参数 $\boldsymbol{\theta}^\star$ 在 s-共同随机数耦合轨迹（任务$\mathbf{w}$）中运行，定义

$$a := \frac{c(1-\gamma)}{2}\|\mathbf{U}\|_{\mathrm{op}},\quad b := \frac{c(1-\gamma)}{2}\sqrt{\frac{K}{2}}\Big(\|\mathbf{V}+\mathbf{U}\,\mathrm{Diag}(\mathbf{w})\|_{\mathrm{op}} + \sqrt{\tfrac{2}{\pi}}\sigma_\xi\|\mathbf{U}\|_{\mathrm{op}}\Big),$$

该常数不随 s 或 t 增长。令 $C_b = f(b)$ 为另一个绝对常数，则对于任意 $1 \le s \le t$，

$$\mathbb{E}\left[\|\Delta\widehat{\mathbf{p}}^s_{t+1}\|_2|\mathcal{F}_{s-1}\right] \le \tfrac{a(1+C_b)}{s}\left(\tfrac{t}{s}\right)^{b-1}|\delta_r|.$$

特别地，令学习率 $\eta_t = c/t$ 足够小，使得 $b<1$，则 s-共同随机数耦合轨迹中的一次性奖励偏移递减至0，即$\lim_{t\to\infty}\mathbb{E}\left[\,\|\Delta\widehat{\mathbf{p}}_{t+1}\|_2 \mid \mathcal{F}_{s-1}\right] = 0$。

算法 1 ME-ICPO：最小熵上下文策略优化

输入： 问题 Q；轮次数 N；每轮候选项数 k；
输入： 系统提示 SysPrompt；摘要器（提示）Summ
1: $\mathcal{H}_0 \leftarrow \{\text{SysPrompt}\ \ Q\}$
2: **对** $t = 1$ **到** n **执行**
3: 示例回答 $A_{1\ k}^{(t)}$: $\sim p_t\ (\cdot \mid \mathcal{H}_{t-1})$ 及其回答 $a_j^{(t)}$ 对于所有 $j \in [k]$
4: 评估 $\widehat{a}_t \leftarrow$ 多数投票$\{a_j^{(k)}\}_{j=1}^{K}$，令 $r_j^{(t)} \leftarrow \mathbb{1}[a_j^{(t)} = \widehat{a}_t\}$ 对于所有 $j \in [k]$
5: 总结 x $_j^{(t)} =$ 总结（$A_j^{(}t)$ ）以及 $\widetilde{\mathcal{H}}_j^{(t)} = \mathcal{H}_{t-1} \cup$（$x_j^{(t)}$，$r_j^{(t)}$）对于所有 $j \in [k]$
6: 选择熵最小的响应 $j^* \leftarrow \arg\min_{j\in[k]} H(\widetilde{\mathcal{H}}_j^{(t)})$
7: 更新上下文内列表 $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1}$ $\cup (x_{j^*}^{(t)}, r_{j^*}^{(t)})$
8: **循环结束**
输出： 从 $p_{n+1}(\cdot \mid H_n)$采样的响应

5 最小熵上下文策略优化

基于对ICPO框架的理论理解，我们现在提出一种实用算法，利用上下文策略优化与自获取奖励，通过测试时扩展提升其推理能力。首先，我们采用理论分析中给出的ICPO符号记法。智能体从仅包含问题$\mathbf{Q}$的历史提示 $\mathcal{H}_0 = \{Q\}$ 开始。对于每个时间 $t \in [T]$，模型输出其响应 $\mathbf{x}_t \sim p(\cdot|\mathcal{H}_{t-1})$，然后通过自获取奖励观测奖励 r_t。随后，智能体更新其历史 $\mathcal{H}_t = \{\mathbf{Q}, (\mathbf{x}_1, r_1) \cdots, (\mathbf{x}_t, r_t)\}$，并进入下一个时间步 $t \leftarrow t + 1$。然而，直接应用ICPO框架会带来两个重大挑战。第一个挑战与上下文长度有关：过长的上下文会阻碍智能体通过上下文内策略优化过程累积并开展有效的推理过程。第二个挑战是自评奖励的可信度，因为自我评估可能存在噪声且不准确。为应对这两个挑战，我们提出测试时上下文扩展算法：最小熵上下文策略优化（ME-ICPO）。ME-ICPO算法按以下三个主要步骤运行。

响应生成与自我评估。 在第3行，智能体采样 k 个响应 $A_k^{(t)}$，使用历史上下文内数据 $\mathcal{H}_{t-1}$，其最终答案 $a_k^{(t)}$ 出现在最终的 `boxed{}`块中（Hendrycks 等人，2021年）。随后，按照 Wang 等人 （2022年）的做法，对第4 行中的答案进行多数投票，以评估响应的准确率。

思维链摘要。 为了压缩输出响应中的信息并精简上下文内文本，我们根据其思维链（思维链）对响应进行摘要 $x_j^{(t)}$，并忽略第5行中详细的问题求解过程。其动机在于，详细的数值处理预期比全局思维链更容易。

最小熵响应选择。与Yao 等人（2023a）的树搜索式算法类似，我们选择一个响应 $x_{j^*}^{(t)}$，并将其及其响应放入第7行的上下文内历史 $\mathcal{H}_{t+1}$ 中。具体而言，不同于面向多步推理而设计的树搜索Yao 等人（2023a）或Best-of-N Huang 等人（2025）算法——在那些算法中 x_t 是单步响应——我们将 x_t 视为求解整个问题的思维链描述。因此，我们并非选择奖励最高的 x_t，而是遵循离线强化学习中的“悲观主义”，选择能使未来响应熵最小的 x_t，即第6行中所执行的 $j^* \leftarrow \arg\min_{j\in[k]} H(\widetilde{\mathcal{H}}_j^{(t)})$。直观上，这种最小熵选择可以避免智能体选择受损响应 x，否则可能驱使智能体产生随机答案。此外，该过程还会鼓励智能体选择多样化的响应 x_{j^*}，从而有助于进一步降低熵。

需要将我们的测试时方法与在固定教师算法生成的轨迹上训练学生模型的方法区分开来，这一点至关重要。由于 ME-ICPO 在测试时不会进行参数更新，我们并不声称所部署的大语言模型与任何特定的策略优化算法唯一等价。相反，我们的理论分析表明，LSA 架构具有很强的归纳偏置，能够执行此类更新。ME-ICPO 旨在通过奖励感知提示和有原则的反馈选择提供实用的接口，从而有效利用这一

模型在测试时固有的计算能力，而无需梯度（Madaan 等人，2023年；Shinn 等人，2023年）。复杂度推导和前因子讨论详见附录C。

6 实验

我们展示实验结果，以验证第4节中的理论主张，以及 ME-ICPO 在本节中的表现。

6.1 验证实验

为验证定理 4.2 和定理 4.8 的结果，我们在一个单一LSA 上进行了两项受控检验。图 2分别展示了师生策略匹配误差（上图）以及混合策略在一次性奖励冲击下的稳定性，并给出了实例化的解析界（下图）。

师生策略匹配。我们固定元配置 $K = 10, N = 30, B = 100,$ $\gamma = 0.2$，步长 $\eta_t = c/t$，其中 $c = 1.0$；任务先验 $\mathrm{w} \sim \mathcal{N}(0, \tau_w^2 I)$，其中 $\tau_w = 1.0$；奖励噪声 $\epsilon_t \sim \mathcal{N}(0, \sigma_\xi^2)$，其中 $\sigma_\xi = 0.5$。测试时，我们从同一生成过程中抽取 $B_{\text{test}} = 64$ 个新任务。每一轮，我们固定已实现历史 $\mathcal{H}_t$，计算教师混合策略 $\mathrm{p}_{t+1}^{\mathrm{PO}}$ 和模型的混合策略 $\widehat{\mathrm{p}}_{t+1}$（定义见第 4 节），然后使用模型策略继续闭环。对任务进行汇总后，我们报告 $\mathbb{E}\|\widehat{\mathrm{p}}_t - \mathrm{p}_t^{\mathrm{PO}}\|_2$ 的均值和一个标准差范围。如图 2 所示，差距迅速降至数值精度范围内，并随 t 增大而减小，这与总体等价性和有限样本保证一致。

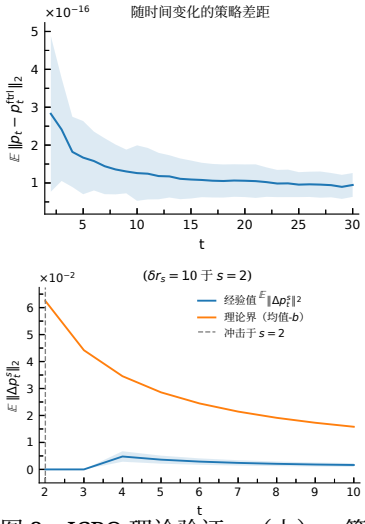


图 2：ICPO 理论验证。（上）：策略匹配。（下）：奖励冲击稳定性。

单次奖励冲击下的稳定性。我们使用通过监督模仿在 PO轨迹上训练的 LSA，其参数为 $K = 5$，$N = 5$，$\gamma = 0.8$，$c = 0.5$，$\tau_w = 0.5$和 $\sigma_\xi = 0.1$。测试时，我们保持数据模型和 γ 不变，将时域扩展至$N = 10$，并在 $B_{\text{test}} = 256$ 个任务上进行评估。对于每个任务，我们运行两条 s条由共同随机数耦合的轨迹：一条基线轨迹，以及一条在 $\delta r_s = 1.0$ 处注入单次奖励冲击的扰动轨迹 $s = 2$。按照定义4.7，我们记录 $\Delta\widehat{\mathbf{p}}_t^s$，并绘制均值和的一个标准差带 $\mathbb{E}\|\Delta\widehat{\mathbf{p}}_t^s\|_2$。我们可以计算 $b \in [0.1236, 0.2127]$，从而检验非放大条件。图2显示，冲击后短暂出现峰值，随后稳定下降，随时间推移没有放大；分析曲线则给出了保守的上界。

6.2 大语言模型实验设置

我们在标准数学问答基准（AIME 2024、AMC 和 MATH-500，按照 TTRL 划分为五个难度等级）（aop，2024；美国数学协会，2025；Li 等人，2024；Hendrycks 等人，2021）上，使用具有代表性的骨干模型（Qwen2.5-Math-1.5B 和 Qwen2.5-Math-7B）（Yang 等人，2025）评估 ME-ICPO。作为参考，我们还报告了 Llama-3.1-8B-Instruct（Grattafiori 等人，2024）和 DeepSeek-R1-Distill-Llama-8B（Guo 等人，2025）在 AIME 2024 上的表现。完整的超参数、基线配置、硬件/软件环境以及提示词详情见附录 B.2、B.3 和 D。遵循 Guo 等人（2025）的做法，我们针对每个问题生成 k=16 个回答（T=0.6，top-p=0.95）。我们报告 Mean@k= $\frac{1}{|\mathcal{D}|}\sum_{q\in\mathcal{D}}\frac{1}{k}\sum_{i=1}^{k} c_i$（$q$），其中 $c_i^-(q) = \mathbf{1}[a_i(q) = a^\star(q)]$，，以及准确率，即单次作答正确的概率。

6.3 主要结果

我们的主要结果见表1，报告了*Mean@16* (%) 和准确率(%) 在所有任务和模型上的均值及标准差（基于 5 个随机种子）。结果表明，ME-ICPO 通过其无梯度的上下文优化过程，持续提升了性能。这些提升同时适用于较大的模型（**Qwen2.5-Math-7B**）和较小的模型（**Qwen2.5-Math-1.5B**），这表明 ME-ICPO 在不同模型规模下均有效。附录D中提供了经过定性筛选的提示示例。我们还评估了预言机奖励变体，结果汇总于表2。

我们进一步报告了其他模型 **Llama-3.1-8B-Instruct** (Grattafiori 等人，2024年）在 **DeepSeek-R1-Distill-Llama-8B** (Guo 等人，2025）在 **AIME 2024** 上的表现，见图 3。

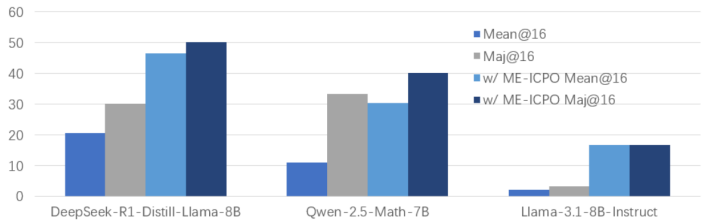


图 3：ME-ICPO 前后骨干模型的性能比较。

表1：不同模型的性能比较，指标为 *Mean@16* 和标准准确率（%）。

模型 \ 基准测试	AIME 2024	AMC	MATH-L1	MATH-L2	MATH-L3	MATH-L4	MATH-L5
			***Mean@16* (%)**				
Qwen2.5-Math-7B	11.04 ± 1.65	41.42 ± 0.99	52.62 ± 1.38	50.76 ± 0.95	49.88 ± 0.88	43.60 ± 0.80	30.58 ± 0.78
使用 ME-ICPO	30.42 ± 1.81	47.06 ± 0.84	62.35 ± 1.27	64.31 ± 1.23	58.87 ± 0.78	49.31 ± 0.97	38.71 ± 0.86
Δ	+19.38	+5.64	+9.73	+13.55	+8.99	+5.71	+8.13
Qwen2.5-Math-1.5B	6.46 ± 0.96	30.42 ± 0.58	49.27 ± 0.81	48.54 ± 0.67	45.42 ± 0.86	40.28 ± 0.57	25.23 ± 0.45
使用 ME-ICPO	9.79 ± 1.11	33.58 ± 0.63	61.19 ± 0.77	54.93 ± 0.71	52.08 ± 0.69	46.44 ± 0.64	29.85 ± 0.70
Δ	+3.33	+3.16	+11.92	+6.39	+6.66	+6.16	+4.62
			准确率 (%)				
Qwen2.5-Math-7B	11.13 ± 3.27	41.33 ± 1.97	46.98 ± 2.71	42.67 ± 1.86	43.71 ± 1.75	37.79 ± 1.65	26.13 ± 1.54
使用 ME-ICPO	30.05 ± 3.02	47.20 ± 2.26	57.32 ± 2.37	54.74 ± 2.04	51.90 ± 1.67	40.84 ± 1.78	31.71 ± 1.40
Δ	+18.92	+5.87	+10.34	+12.07	+8.19	+3.05	+5.58
Qwen2.5-Math-1.5B	6.51 ± 1.94	30.25 ± 1.07	44.68 ± 1.83	39.95 ± 1.32	39.77 ± 0.99	34.97 ± 0.68	20.89 ± 1.35
使用 ME-ICPO	9.82 ± 2.15	33.73 ± 0.96	57.06 ± 1.70	47.60 ± 1.00	47.81 ± 1.17	41.55 ± 0.72	24.83 ± 1.29
Δ	+3.31	+3.48	+12.38	+7.65	+8.04	+6.58	+3.94

我们还考察了 Maj@k 指标（Wang 等人，2023b），该指标通过多数投票聚合 k 个回答 $\hat{a}(q)$ $=$ 模式$(a_1(q)\ldots,a_k(q))$（平票时均匀随机打破平局），并计算 Maj@k $= \frac{1}{|\mathcal{D}|}\sum_{q\in\mathcal{D}} \mathbf{1}\{\hat{a}(q) = a^\star(q)\}$。如图 3 所示，我们的实验表明，ME-ICPO 的平均表现（Mean@16）可以超过基础模型多数投票所预期的上界，这与 TTRL 中的观察结果类似。此外，在 ME-ICPO 的输出上再进行多数投票，还能带来额外提升。

6.4 分析与消融实验

我们进一步使用 **Qwen2.5-Math-7B** 分析 AIME 2024 基准，以评估 ME-ICPO 核心组件的贡献及其对超参数的敏感度，详情见附录 B.1；附录 C中则提供了计算成本分析。

消融实验。为单独评估核心组件——基于熵的选择和显式奖励信号——的贡献，我们进行了消融实验，结果见表2。结果清楚地表明，最小熵选择标准是

表 2：ME-ICPO 的消融实验。（Oracle 结果使用真实标签作为奖励，仅供参考。）

方法	准确率 (%)	*Mean@16* (%)
不含奖励	19.30	19.17
不含熵	5.77	5.83
不含熵和奖励	6.21	6.46
ME-ICPO (完整版本)	**30.05**	**30.42**
ME-ICPO Oracle	38.19	38.12

算法中最关键的组件。移除这一贪心选择机制（无熵）会导致性能急剧下降。由系统提示明确呈现的显式奖励信号同样发挥着至关重要的作用。在保留熵选择的情况下，移除奖励标签（无奖励）仍会导致性能显著下降。

7 结论

在这项工作中，我们通过引入一个有理论依据的上下文策略优化（ICPO）框架，研究了大语言模型推理中的测试时扩展和自我反思现象。我们证明，在 ICPO 框架下，单层线性自注意力Transformer 可以在理论上模拟策略优化算法，为如何在大语言模型中实现自我反思提供了理论概念验证。基于 ICPO 框架，我们提出了一种实用且有效的算法——最小熵上下文策略优化（ME-ICPO），该算法提供了一套结合自评奖励与上下文内响应选择的测试时扩展流程。大量经验研究表明，我们的方法在多个不同基准上提升了性能。

可复现性声明

我们为确保结果可复现付出了大量努力。我们在附录中详细介绍了所有数据集、模型配置、超参数和训练流程 B。为便于忠实复现，我们在以下地址公开了完整的代码库、实验脚本和配置文件：https://github.com/UNCSciML/ICPO。该代码仓库包含准确的随机种子和环境规格。我们在补充材料中提供了源代码，以便忠实复现我们的实验。

伦理声明

我们已仔细审阅《伦理准则》，认为我们的工作不存在任何重大伦理问题。我们的研究不涉及人类受试者、敏感数据或可能造成危害的应用。我们相信，我们的方法和贡献符合公平性、透明度和科研诚信的原则。

致谢

我们感谢匿名审稿人提出的宝贵意见。本研究的部分工作由 ZW 在微软研究院实习期间完成。本研究还获得了北卡罗来纳大学教堂山分校数据科学与社会学院为 WZ 提供的创业启动资金支持。

参考文献

美国数学邀请赛题目及解答，2024年。访问日期：2025年9月13日。

Yu Bai、Fan Chen、Huan Wang、Caiming Xiong 和 Song Mei。作为统计学家的 Transformer：可证明的上下文学习与上下文算法选择。神经信息处理系统进展，36：57125–57211，2023年。

Mislav Balunovi´c、Jasper Dekoninck、Ivo Petrov、Nikola Jovanovi´c 和 Martin Vechev。MathArena：在未污染的数学竞赛上评估大语言模型。*arXiv* 预印本 *arXiv:2505.23281*，2025。

Maciej Besta、Nils Blach、Ales Kubicek、Robert Gerstenberger、Michal Podstawski、Lukas Gianinazzi、Joanna Gajda、Tomasz Lehmann、Hubert Niewiadomski、Piotr Nyczyk 等。思维图：使用大语言模型解决复杂问题。载于*AAAI*人工智能大会论文集，第38卷，第17682–17690页，2024年。

Xingwu Chen、Lei Zhao 和 Difan Zou。Transformer 如何在上下文学习中利用多头注意力？一项关于稀疏线性回归的案例研究。神经信息处理系统进展，37：119573–119613，2024a。

Zhaorun Chen、Zhuokai Zhao、Zhihong Zhu、Ruiqi Zhang、Xiang Li、Bhiksha Raj 和 Huaxiu Yao。Autoprm：通过可控问题分解实现多步推理程序监督自动化。*arXiv* 预印本 *arXiv:2402.11452*，2024b。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、MarcelBlistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gemini 2.5：凭借高级推理、多模态、长上下文和新一代智能体能力，推动前沿发展。*arXiv* 预印本 *arXiv:2507.06261*，2025。

Shivam Garg、Dimitris Tsipras、Percy S Liang 和 Gregory Valiant。Transformer 能在上下文内学到什么？简单函数类案例研究。神经信息处理系统进展，35:30583–30598，2022年。

Paul Glasserman 和 David D Yao。共同随机数的一些准则和保证。《管理科学》，38(6):884–908，1992。

Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等。Llama 3 模型群。*arXiv* 预印本 *arXiv:2407.21783*，2024年。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang 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。

Botao Hao、Tor Lattimore 和 Csaba Szepesvari。线性上下文赌博机中的自适应探索。载于人工智能与统计国际会议，第 3536–3545 页。PMLR，2020年。Dan Hendrycks、Collin Burns、Saurav Kadavath、Akul Arora、Steven Basart、Eric Tang、Dawn Song 和 Jacob Steinhardt。使用 MATH 数据集评估数学问题求解能力。*arXiv* 预印本 *arXiv:2103.03874*，2021。Audrey Huang、Adam Block、Qinghua Liu、Nan Jiang、Akshay Krishnamurthy 和 Dylan J Foster。Best-of-N 是否真的最好？推理时对齐中的覆盖率、扩展性与最优性。*arXiv* 预印本 *arXiv:2503.21878*，2025。Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、Ahmed El-Kishky、Aiden Low、Alec Helyar、Aleksander Madry、Alex Beutel、Alex Carney 等。OpenAI o1 系统卡。*arXiv* 预印本 *arXiv:2412.16720*，2024年。Sham M Kakade。自然策略梯度。神经信息处理系统进展，14，2001。Jia Li、Edward Beeching、Lewis Tunstall、Ben Lipkin、Roman Soletskyi、Shengyi Huang、Kashif Rasul、Longhui Yu、Albert Q Jiang、Ziju Shen 等。NuminaMath：AI4Maths 中规模最大的公开数据集，包含 86 万组竞赛数学题及其解答。*Hugging Face* 仓库，13(9):9，2024年。Hunter Lightman、Vineet Kosaraju、Yuri Burda、Harrison Edwards、Bowen Baker、Teddy Lee、Jan Leike、John Schulman、Ilya Sutskever 和 Karl Cobbe。让我们逐步验证。载于第十二届国际学习表征会议，2023。Licong Lin、Yu Bai 和 Song Mei。Transformer 作为决策者：通过监督式预训练实现可证明的上下文内强化学习。*arXiv* 预印本 *arXiv:2310.08566*，2023。Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang 等。Self-Refine：通过自我反馈进行迭代改进。神经信息处理系统进展，36:46534–46594，2023。美国数学协会。美国数学竞赛，2025。访问日期：2025年9月13日。Brendan McMahan。跟随正则化领导者与镜像下降：等价定理及 L1 正则化。载于第十四届人工智能与统计国际会议论文集，第 525–533 页。JMLR 研讨会与会议论文集，2011a。H. Brendan McMahan。跟随正则化领导者与镜像下降：等价定理及 L1 正则化。载于 *AISTATS*，2011b。Matteo Papini、Andrea Tirinzoni、Marcello Restelli、Alessandro Lazaric 和 Matteo Pirotta。在线性上下文赌博机中利用优质表征。载于国际机器学习会议，第 8371–8380 页。PMLR，2021。Chanwoo Park、Xiangyu Liu、Asuman Ozdaglar 和 Kaiqing Zhang。大语言模型智能体会产生遗憾吗？在线学习与博弈中的案例研究。*arXiv* 预印本 *arXiv:2403.16843*，2024年。Archiki Prasad、Weizhe Yuan、Richard Yuanzhe Pang、Jing Xu、Maryam Fazel-Zarandi、Mohit Bansal、Sainbayar Sukhbaatar、Jason Weston 和 Jane Yu。自洽偏好优化。*arXiv* 预印本 *arXiv:2411.04109*，2024年。David Rein、Betty Li Hou、Asa Cooper Stickland、Jackson Petty、Richard Yuanzhe Pang、Julien Dirani、Julian Michael 和 Samuel R Bowman。GPQA：研究生水平的 Google-proof 问答基准。载于首届语言建模会议，2024年。Shai Shalev-Shwartz。在线学习：理论、算法与应用。希伯来大学，2007。Lior Shani、Yonathan Efroni、Aviv Rosenberg 和 Shie Mannor。基于老虎机反馈的乐观策略优化。载于国际机器学习会议，第 8604–8613 页。PMLR，2020年。

Noah Shinn、Federico Cassano、Beck Labash、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao。《Reflexion：采用语言强化学习的语言智能体》，2023 年。*URL https://arxiv.org/abs/2303.11366*，1，2023 年。

Yuda Song、Hanlin Zhang、Carson Eisenach、Sham Kakade、Dean Foster 和 Udaya Ghai。《注意差距：审视大语言模型的自我改进能力》。*arXiv* 预印本 *arXiv:2412.02674*，2024 年。

Johannes Von Oswald、Eyvind Niklasson、Ettore Randazzo、Jo˜ao Sacramento、Alexander Mordvintsev、Andrey Zhmoginov 和 Max Vladymyrov。《Transformer 通过梯度下降进行上下文内学习》。发表于国际机器学习会议，第 35151–35174 页。PMLR，2023 年。

Jiuqi Wang、Ethan Blaser、Hadi Daneshmand 和 Shangtong Zhang。Transformer 能够学习用于上下文强化学习的时序差分方法。*arXiv* 预印本 *arXiv:2405.13861*，2024年。

Peiyi Wang、Lei Li、Zhihong Shao、RX Xu、Damai Dai、Yifei Li、Deli Chen、Yu Wu 和 Zhifang Sui。Math-shepherd：无需人工标注，逐步验证并强化大语言模型。*arXiv* 预印本 *arXiv:2312.08935*，2023a。

Ruiqi Wang、Jiyu Guo、Cuiyun Gao、Guodong Fan、Chun Yong Chong 和 Xin Xia。大语言模型能否取代人工评估者？软件工程领域大语言模型评审的经验研究。*ACM* 软件工程论文集，2（ISSTA）：1955–1977，2025。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdh- ery 和 Denny Zhou。自一致性能够提升语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*，2022年。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdh- ery 和 Denny Zhou。自一致性提升了语言模型中的思维链推理。发表于*ICLR*，2023b。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示能够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年。

Weiqiang Wu、Jing Yang 和 Cong Shen。具有多样化上下文的随机线性上下文老虎机。发表于人工智能与统计国际会议，第2392–2401页。PMLR，2020年。

An Yang、Beichen Zhang、Binyuan Hui、Bofei Gao、Bowen Yu、Chengpeng Li、Dayiheng Liu、Jian- hong Tu、Jingren Zhou、Junyang Lin 等。Qwen2.5-math 技术报告：通过自我改进迈向数学专家模型。*arXiv* 预印本 *arXiv:2409.12122*，2024a。

An Yang、Beichen Zhang、Binyuan Hui、Bofei Gao、Bowen Yu、Chengpeng Li、Dayiheng Liu、Jian- hong Tu、Jingren Zhou、Junyang Lin等。Qwen2.5-math 技术报告：通过自我改进迈向数学专家模型。*arXiv* 预印本 *arXiv:2409.12122*，2024b。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、ChengenHuang、Chenxu Lv等。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025。

Shunyu Yao、Dian Yu、Jeffrey Zhao、Izhak Shafran、Tom Griffiths、Yuan Cao和 KarthikNarasimhan。思维树：借助大语言模型进行审慎的问题求解。神经信息处理系统进展，36：11809–11822，2023a。

Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。ReAct：在语言模型中协同推理与行动。发表于国际学习表征会议（*ICLR*），2023b。

Di Zhang、Xiaoshui Huang、Dongzhan Zhou、Yuqiang Li 和 Wanli Ouyang。通过使用 Llama-3 8B 进行蒙特卡洛树自我优化，获得达到 GPT-4 水平的数学奥林匹克解答。*arXiv* 预印本 *arXiv:2406.07394*，2024a。Ruiqi Zhang、Spencer Frei 和 Peter L Bartlett。经过训练的 Transformer 能在上下文内学习线性模型。机器学习研究杂志，25(49):1–55，2024b。

Pei Zhou、Jay Pujara、Xiang Ren、Xinyun Chen、Heng-Tze Cheng、Quoc V Le、Ed Chi、Denny Zhou、Swaroop Mishra 和 Huaixiu Steven Zheng。Self-discover：大语言模型自行组合推理结构。神经信息处理系统进展，37:126032–126058，2024年。

Yuxin Zuo、Kaiyan Zhang、Li Sheng、Shang Qu、Ganqu Cui、Xuekai Zhu、Haozhan Li、Yuchen Zhang、Xinwei Long、Ermo Hua 等。TTRL：测试时强化学习。*arXiv* 预印本 *arXiv:2504.16084*，2025。

附录

A 第 4 节的证明与技术细节 14
A.1 预备知识与关键引理 .14A.2 奖励冲击鲁棒性 .21A.3 KL 散度与费 Fisher 加权二次项 .22A.4 策略优化的闭环模仿 .23A.5 Fisher 训练的双通道 LSA 的收敛性 .24

B 实验细节 25
B.1 超参数敏感性。 .25B.2 超参数与实现细节 .25B.3 硬件与环境配置 .26B.4 补充实验结果 .26

C 详细复杂度分析 27
C.1 理论时间与显存复杂度推导 .27C.2 实证成本 .31

D 提示模板与定性案例 31
D.1 数据集专用系统提示 .31D.2 摘要提示 .32D.3 定性案例研究 .33

附录 P证明与技术细节第4 节

A.1 预备知识与关键引理

在本小节中，我们将建立后续证明所需的最少工具和引理。我们采用正文中的所有符号记法和设定。

引理 A.1 （$\mathbf{1}^\perp$ 上的混合曲率与费舍尔界）。 如果教师采用混合探索 $\mathbf{p} = (1-\gamma)\tilde{\mathbf{p}} + \gamma\,\mathbf{1}/K$，那么费舍尔矩阵 $\mathbf{F}(\mathbf{p}) := $ 对角矩阵$(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ 在 $\mathbf{1}^\perp$ 上有界：

$$\frac{\gamma}{K}\,\mathbf{I} \;\preceq\; \mathbf{F}(\mathbf{p})\Big|_{\mathbf{1}^\perp} \;\preceq\; \frac{1}{2}\,\mathbf{I}.$$

证明。 利用对角矩阵$(\cdot)$的线性性质，并展开 $\mathbf{p}\mathbf{p}^\top$，

$$\begin{aligned}\mathbf{F}(\mathbf{p}) &= \mathrm{Diag}\big((1-\gamma)\tilde{\mathbf{p}} + \gamma\,\tfrac{\mathbf{1}}{K}\big) - \big((1-\gamma)\tilde{\mathbf{p}} + \gamma\,\tfrac{\mathbf{1}}{K}\big)\big((1-\gamma)\tilde{\mathbf{p}} + \gamma\,\tfrac{\mathbf{1}}{K}\big)^\top \\ &= (1-\gamma)\mathbf{F}(\tilde{\mathbf{p}}) \;+\; \gamma\,\mathbf{F}\big(\tfrac{\mathbf{1}}{K}\big) \;+\; \gamma(1-\gamma)\,(\tilde{\mathbf{p}} - \tfrac{\mathbf{1}}{K})(\tilde{\mathbf{p}} - \tfrac{\mathbf{1}}{K})^\top\end{aligned}$$

最后的秩一项和$\mathbf{F}(\tilde{\mathbf{p}})$均为半正定，因此

$$\mathbf{F}(\mathbf{p}) \;\succeq\; \gamma\,\mathbf{F}\big(\tfrac{\mathbf{1}}{K}\big).$$

在 $\mathbf{1}^\perp$ 上，$\mathbf{F}(\frac{\mathbf{1}}{k}) =$ 对角矩阵$(\frac{\mathbf{1}}{k}) = (\frac{\mathbf{1}}{k})(\frac{\mathbf{1}}{k})^\top$的作用等同于 $(\frac{\mathbf{1}}{k})\mathbf{I}$，因此可得下界：

$$\mathbf{F}(\mathbf{p})\Big|_{\mathbf{1}^\perp} \;\succeq\; \frac{\gamma}{K}\,\mathbf{I}.$$

对于上界，任意$\mathbf{x} \in \mathbf{1}^\perp$ 满足 $\|\mathbf{x}\|_2 = 1$,

$$\mathbf{x}^\top\mathbf{F}(\mathbf{p})\,\mathbf{x} \;=\; \mathrm{Var}_{i\sim\mathbf{p}}[x_i] \;\le\; \frac{(\max_i x_i - \min_i x_i)^2}{4} \;\le\; \frac{\big(\sqrt{2}\big)^2}{4} \;=\; \frac{1}{2},$$

这里用到了波波维丘不等式，并且在满足 $\mathbf{x}^\top\mathbf{1} = 0$ 和 $\|\mathbf{x}\|_2 = 1$的向量中，取值范围在两个坐标上放置质量 $\pm 1/\sqrt{2}$ 时达到最大。对这类 $\mathbf{x}$取上确界，得到$\mathbf{F}(\mathbf{p})\big|_{\mathbf{1}^\perp} \preceq \frac{1}{2}\,\mathbf{I}$。 □

引理 A.2（softmax 在 $\mathbf{1}^\perp$ 上是 1/2-利普希茨的）。对于任意 $\mathbf{u},\ \mathbf{v} \in \mathbb{R}^K$，满足 $\mathbf{u}-\mathbf{v} \in \mathbf{1}^\perp$，

$$\|\mathrm{softmax}(\mathbf{u}) - \mathrm{softmax}(\mathbf{v})\|_2 \ \le\ \tfrac{1}{2}\,\|\mathbf{u}-\mathbf{v}\|_2.$$

证明。根据向量映射的中值定理，存在 ξ，位于线段 $[v,u]$ 上，使得 at

$$\mathrm{softmax}(\mathbf{u}) - \mathrm{softmax}(\boldsymbol{b}) \ =\ J(\xi)\,(\mathbf{u}-\mathbf{v}),$$

其中 $J(\xi)$ 是 softmax 的雅可比矩阵。众所周知，$J(\xi) = \mathbf{F}(\mathbf{p})$ 其中 $\mathbf{p} = \mathrm{softmax}(\xi)$，且 $J(\xi)\mathbf{1} = \mathbf{0}$。由于 $\mathbf{u}-\mathbf{v} \in \mathbf{1}^\perp$，我们可以将范围限制为 $\mathbf{1}^\perp$ 并应用引理 A.1：

$$\|\mathrm{softmax}(\mathbf{u}) - \mathrm{softmax}(\mathbf{v})\|_2 \ =\ \|J(\xi)(\mathbf{u}-\mathbf{v})\|_2 \ \le\ \|J(\xi)\big|_{\mathbf{1}^\perp}\|_{\mathrm{op}}\,\|\mathbf{u}-\mathbf{v}\|_2 \ \le\ \tfrac{1}{2}\,\|\mathbf{u}-\mathbf{v}\|_2.$$

□

引理 A.3（单层 LSA 的查询列闭式）。下一步的 logit**向量**可以写成闭式

$$\widehat{\mathbf{s}}_{t+1} \ =\ \mathbf{q}_x \ +\ \frac{1}{t}\mathbf{R}\,\mathbf{G}_t\,\boldsymbol{b}, \tag{A.1}$$

其中 $\mathbf{R} := [\mathbf{W}^{PV}]_{1:K,:}$ 是行选择器**矩阵**，$\boldsymbol{b} := \mathbf{W}^{KQ}\boldsymbol{q}$ 是变换后的**查询向量**，$\boldsymbol{q} := \begin{pmatrix}\mathbf{q}_x \\ q_r\end{pmatrix} \in \mathbb{R}^{K+1}$，其中 $\mathbf{q}_x \in \mathbb{R}^K$ 和 $q_r \in \mathbb{R}$，且 $\mathbf{G}_t := \mathbf{E}^{(t)}(\mathbf{E}^{(t)})^\top$ 是历史格拉姆矩阵。

证明。根据方程 3.4 中的 LSA 定义，

$$\begin{aligned}
\widehat{\mathbf{s}}_{t+1} &:= \big[f_{\mathrm{lsa}}(\mathbf{E}^{(t)};\boldsymbol{\theta})\big]_{1:K,\,t+1} \\
&= \left[\mathbf{E}^{(t)} + \mathbf{W}^{PV}\mathbf{E}^{(t)}\cdot\frac{(\mathbf{E}^{(t)})^\top\mathbf{W}^{KQ}\mathbf{E}^{(t)}}{t}\right]_{1:K,\,t+1} \\
&= \big[\mathbf{E}^{(t)}\big]_{1:K,\,t+1} \ +\ \frac{1}{t}\left[\mathbf{W}^{PV}\mathbf{E}^{(t)}\big((\mathbf{E}^{(t)})^\top\mathbf{W}^{KQ}\mathbf{E}^{(t)}\big)\right]_{1:K,\,t+1} \\
&\overset{(a)}{=} \mathbf{q}_x \ +\ \frac{1}{t}\left[\mathbf{W}^{PV}\mathbf{E}^{(t)}\big((\mathbf{E}^{(t)})^\top\mathbf{W}^{KQ}\mathbf{E}^{(t)}\mathbf{e}_{t+1}\big)\right]_{1:K} \\
&\overset{(b)}{=} \mathbf{q}_x \ +\ \frac{1}{t}\left[\mathbf{W}^{PV}\mathbf{E}^{(t)}\big((\mathbf{E}^{(t)})^\top\mathbf{W}^{KQ}(\mathbf{E}^{(t)}\mathbf{e}_{t+1})\big)\right]_{1:K} \\
&\overset{(c)}{=} \mathbf{q}_x \ +\ \frac{1}{t}\left[\mathbf{W}^{PV}\big(\mathbf{E}^{(t)}(\mathbf{E}^{(t)})^\top\big)\big(\mathbf{W}^{KQ}\boldsymbol{q}\big)\right]_{1:K} \\
&\overset{(d)}{=} \mathbf{q}_x \ +\ \frac{1}{t}\,\mathbf{R}\,\mathbf{G}_t\,\boldsymbol{b}.
\end{aligned}$$

理由：(a) 按列切片等同于右乘标准基 $\mathbf{e}_{t+1} \in \mathbb{R}^{t+1}$，且$[\mathbf{E}^{(t)}]_{1:K,\,t+1} = \mathbf{q}_x$；(b) 利用结合律分离出 $\mathbf{E}^{(t)}\mathbf{e}_{t+1}$；(c) 代入 $\mathbf{E}^{(t)}\mathbf{e}_{t+1} = \boldsymbol{q} = (\mathbf{q}_x^\top,\, q_r)^\top$并重新组合 $\mathbf{E}^{(t)}(\mathbf{E}^{(t)})^\top$；(d) 应用简洁定义 $\mathbf{R} = [\mathbf{W}^{PV}]_{1:K,:}$、$\mathbf{G}_t = \mathbf{E}^{(t)}(\mathbf{E}^{(t)})^\top$，以及 $\boldsymbol{b} = \mathbf{W}^{KQ}\boldsymbol{q}$。

□

引理 A.4（双通道投影 logits：精确相等）。假设满足架构标准形式：$q_r = 0$，$\mathbf{q}_x \in$ 张成$\{\mathbf{1}\}$，且动作–logit 投影的最后一列平行于$\mathbf{1}$（即，$\mathbf{w}_{12}^{PV} \parallel \mathbf{1}$）。设截至第 t 轮的历史统计量为**计数向量**

$$\mathbf{n}_t := \sum_{s=1}^{t}\mathbf{x}_s \ =\ \sum_{s=1}^{t}\mathbf{e}_{A_s},$$

以及**奖励向量**

$$\mathbf{g}_t := \sum_{s=1}^{t} r_s\mathbf{x}_s \ =\ \sum_{s=1}^{t} r_s\,\mathbf{e}_{A_s}.$$

设 $\mathbf{W}^{PV}, \mathbf{W}^{KQ} \in \mathbb{R}^{(K+1)\times(K+1)}$ 按块划分为

$$\mathbf{W}^{PV} = \begin{pmatrix}\mathbf{W}_{11}^{PV} & \mathbf{w}_{12}^{PV} \\ (\mathbf{w}_{21}^{PV})^\top & w_{22}^{PV}\end{pmatrix}, \qquad \mathbf{W}^{KQ} = \begin{pmatrix}\mathbf{W}_{11}^{KQ} & \mathbf{w}_{12}^{KQ} \\ (\mathbf{w}_{21}^{KQ})^\top & w_{22}^{KQ}\end{pmatrix},$$

其中 $\mathbf{W}_{11}^{PV}, \mathbf{W}_{11}^{KQ} \in \mathbb{R}^{K\times K}$ 以及 $\mathbf{w}_{12}^{PV}, \mathbf{w}_{21}^{PV}, \mathbf{w}_{12}^{KQ}, \mathbf{w}_{21}^{KQ} \in \mathbb{R}^{K}$。对于变换后的查询

$$\boldsymbol{b} \;:=\; \mathbf{W}^{KQ}\boldsymbol{q} \;=\; \begin{pmatrix}\boldsymbol{\phi}_1\\ \phi_2\end{pmatrix}, \qquad \boldsymbol{\phi}_1 \in \mathbb{R}^K,\ \phi_2 \in \mathbb{R},$$

投影后的 LSA 对数几率满足：对于任意非零的 $\boldsymbol{\phi}_1,\ \phi_2$,

$$\mathrm{Proj}\,(\widehat{\mathbf{s}}_{t+1}) \;=\; \frac{1}{t}\Big(\mathbf{W}_n\,\mathbf{n}_t \;+\; \mathbf{W}_g\,\mathbf{g}_t\Big), \tag{A.2}$$

其中有效算子为

$$\mathbf{W}_n \;:=\; \mathrm{Proj}\big(\mathbf{W}_{11}^{PV}\ \mathrm{Diag}(\boldsymbol{\phi}_1)\big), \qquad \mathbf{W}_g \;:=\; \mathrm{Proj}\big(\phi_2\,\mathbf{W}_{11}^{PV}\big).$$

证明。 利用引理 A.3 可得 $\widehat{\ }\mathbf{s}_{t+1} = \mathbf{q}_x + \frac{1}{t}\mathbf{R}\,\mathbf{G}_t\,\boldsymbol{b}$,– 其中 $\mathbf{R} = [\mathbf{W}^{PV}]_{1:K,\ :} = [\,\mathbf{W}_{11}^{PV}\mathbf{w}_{12}^{PV}\,]$，并且

$$\mathbf{G}_t \;=\; \mathbf{E}^{(t)}(\mathbf{E}^{(t)})^\top \;=\; \begin{pmatrix}\mathbf{C}_t + \mathbf{q}_x\mathbf{q}_x^\top & \mathbf{g}_t\\ \mathbf{g}_t^\top & r^\top r\end{pmatrix}, \qquad \mathbf{C}_t := \mathrm{Diag}(\mathbf{n}_t), \quad r := (r_1,\ldots,r_t)^\top.$$

投影到 $\mathbf{1}^\perp$ 并展开分块乘法后,

$$\begin{aligned}
\mathrm{Proj}(\widehat{\mathbf{s}}_{t+1}) &\overset{(a)}{=} \mathrm{Proj}\left(\mathbf{q}_x + \frac{1}{t}\,\mathbf{R}\,\mathbf{G}_t\,\boldsymbol{b}\right)\\
&\overset{(b)}{=} \frac{1}{t}\,\mathrm{Proj}(\mathbf{R}\,\mathbf{G}_t\,\boldsymbol{b})\\
&\overset{(c)}{=} \frac{1}{t}\,\mathrm{Proj}\left(\begin{bmatrix}\mathbf{W}_{11}^{PV} & \mathbf{w}_{12}^{PV}\end{bmatrix}\begin{bmatrix}\mathbf{C}_t + \mathbf{q}_x\mathbf{q}_x^\top & \mathbf{g}_t\\ \mathbf{g}_t^\top & r^\top r\end{bmatrix}\begin{bmatrix}\boldsymbol{\phi}_1\\ \phi_2\end{bmatrix}\right)\\
&\overset{(d)}{=} \frac{1}{t}\,\mathrm{Proj}\big(\mathbf{W}_{11}^{PV}\big((\mathbf{C}_t + \mathbf{q}_x\mathbf{q}_x^\top)\boldsymbol{\phi}_1 + \mathbf{g}_t\,\phi_2\big) \;+\; \mathbf{w}_{12}^{PV}\big(\mathbf{g}_t^\top\boldsymbol{\phi}_1 + (r^\top r)\phi_2\big)\big)\\
&\overset{(e)}{=} \frac{1}{t}\,\mathrm{Proj}\big(\mathbf{W}_{11}^{PV}\,\mathbf{C}_t\,\boldsymbol{\phi}_1 \;+\; \mathbf{W}_{11}^{PV}\,\mathbf{g}_t\,\phi_2\big)\\
&\overset{(f)}{=} \frac{1}{t}\,\big(\mathrm{Proj}\big(\mathbf{W}_{11}^{PV}\ \mathrm{Diag}(\boldsymbol{\phi}_1)\big)\,\mathbf{n}_t \;+\; \mathrm{Proj}\big(\phi_2\,\mathbf{W}_{11}^{PV}\big)\,\mathbf{g}_t\big)\\
&\overset{(g)}{=} \frac{1}{t}\Big(\mathbf{W}_n\,\mathbf{n}_t \;+\; \mathbf{W}_g\,\mathbf{g}_t\Big).
\end{aligned}$$

步骤说明：(a) 将 Proj 应用于闭式表达式；(b) Proj$\mathbf{q}_x = 0$，因为 $\mathbf{q}_x \in$张成$\{\mathbf{1}\}$；(c) 与 $\mathbf{R}$、$\mathbf{G}_t$、$\boldsymbol{b}$进行分块乘法；(d) 显式计算该乘积；(e) Proj$(\mathbf{W}_{11}^{PV}\mathbf{q}_x\mathbf{q}_x^\top\boldsymbol{\phi}_1)$ 和 Proj$(\mathbf{w}_{12}^{PV})$ 均为零，因为 $\mathbf{q}_x$，$\mathbf{w}_{12}^{PV} \parallel \mathbf{1}$；(f) 使用 ($A$对角矩阵($u$))$v$ =(A对角矩阵(v))u，其中 $u = \mathbf{n}_t$、$v = \boldsymbol{\phi}_1$；(g) 按照陈述识别 $\mathbf{W}_n$，$\mathbf{W}_g$。□

引理 A.5（双通道算子中的 Fisher 加权二次型）。 设归一化历史 $^-$统计量为 $^-\mathbf{n}_t := \mathbf{n}_t/t$ 和 $^-\mathbf{g}_t := \mathbf{g}_t/t$。定义拼接算子 $\mathbf{W} := [\mathbf{W}_n\,\mathbf{W}_g\,] \in \mathbb{R}^{K\times 2K}$ 以及拼接后的归一化统计量 $^-\mathbf{z}_t := (^-\mathbf{n}_t^\top, {}^-\mathbf{g}_t^\top)^\top \in \mathbb{R}^{2K}$。设总体二阶矩矩阵为

$$\bar{\boldsymbol{\Sigma}} := \mathbb{E}[\bar{\mathbf{z}}_t\bar{\mathbf{z}}_t^\top] = \begin{pmatrix}\boldsymbol{\Sigma}_{nn} & \boldsymbol{\Sigma}_{ng}\\ \boldsymbol{\Sigma}_{gn} & \boldsymbol{\Sigma}_{gg}\end{pmatrix}, \qquad \boldsymbol{\Sigma}_{y\bar{z}} := \mathbb{E}[\mathbf{y}_{t+1}\bar{\mathbf{z}}_t^\top],\ \text{with}\ \mathbf{y}_{t+1} := \mathrm{Proj}\,(\mathbf{s}_{t+1}^{\mathrm{PO}}).$$

With $\boldsymbol{\Gamma} := \mathbb{E}[\mathbf{F}(\mathbf{p}_{t+1}^{\mathrm{PO}})]$, 由方程 4.1 定义的费舍尔加权损失 可写成二次型

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\,\mathrm{tr}(\boldsymbol{\Gamma}\,\bar{\mathbf{W}}\,\bar{\boldsymbol{\Sigma}}\,\bar{\mathbf{W}}^\top) \;-\; \mathrm{tr}(\boldsymbol{\Gamma}\,\boldsymbol{\Sigma}_{y\bar{z}}\,\bar{\mathbf{W}}^\top) \;+\; \frac{1}{2}\,\mathrm{tr}(\boldsymbol{\Gamma}\,\boldsymbol{\Sigma}_{yy}),$$

其中 $\boldsymbol{\Sigma}_{yy} := \mathbb{E}[\mathbf{y}^{t+1}\mathbf{y}_{t+1}^\top]$。

证明。 根据方程 4.1，对训练对 (τ, t) 进行均匀平均,

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\,\mathbb{E}\Big[\big\|\mathrm{Proj}\,(\widehat{\mathbf{s}}_{t+1} - \mathbf{s}_{t+1}^{\mathrm{PO}})\big\|_{\boldsymbol{\Gamma}}^2\Big].$$

根据引理A.4，投影后的学生 logits 可表示为双通道形式

$$\mathrm{Proj}\,(\widehat{\mathbf{s}}_{t+1}) = \frac{1}{t}\big(\mathbf{W}_n\mathbf{n}_t + \mathbf{W}_g\mathbf{g}_t\big) = \mathbf{W}_n\bar{\mathbf{n}}_t + \mathbf{W}_g\bar{\mathbf{g}}_t = \bar{\mathbf{W}}\,\bar{\mathbf{z}}_t.$$

令 $\mathbf{y}_{t+1} :=$ 投影算子 $(\mathbf{s}^{\mathrm{PO}}_{t+1})$。利用 $\|\boldsymbol{b}\|^2_{\mathbf{\Gamma}} = \boldsymbol{b}^\top \mathbf{\Gamma} \boldsymbol{b}$、期望的线性性以及 $\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$，我们得到

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2}\,\mathbb{E}\Big[\big\|\bar{\mathbf{W}}\bar{\mathbf{z}}_t - \mathbf{y}_{t+1}\big\|^2_{\mathbf{\Gamma}}\Big] \\
&= \frac{1}{2}\,\mathbb{E}\big[(\bar{\mathbf{W}}\bar{\mathbf{z}}_t - \mathbf{y}_{t+1})^\top \mathbf{\Gamma} (\bar{\mathbf{W}}\bar{\mathbf{z}}_t - \mathbf{y}_{t+1})\big] \\
&= \frac{1}{2}\,\mathbb{E}\big[\bar{\mathbf{z}}_t^\top \bar{\mathbf{W}}^\top \mathbf{\Gamma} \bar{\mathbf{W}} \bar{\mathbf{z}}_t\big] \;-\; \mathbb{E}\big[\bar{\mathbf{z}}_t^\top \bar{\mathbf{W}}^\top \mathbf{\Gamma} \mathbf{y}_{t+1}\big] \;+\; \frac{1}{2}\,\mathbb{E}\big[\mathbf{y}_{t+1}^\top \mathbf{\Gamma} \mathbf{y}_{t+1}\big] \\
&= \frac{1}{2}\,\mathrm{tr}\big(\mathbf{\Gamma}\,\bar{\mathbf{W}}\,\mathbb{E}[\bar{\mathbf{z}}_t \bar{\mathbf{z}}_t^\top]\,\bar{\mathbf{W}}^\top\big) \;-\; \mathrm{tr}\big(\mathbf{\Gamma}\,\mathbb{E}[\mathbf{y}_{t+1}\bar{\mathbf{z}}_t^\top]\,\bar{\mathbf{W}}^\top\big) \;+\; \frac{1}{2}\,\mathrm{tr}\big(\mathbf{\Gamma}\,\mathbb{E}[\mathbf{y}_{t+1}\mathbf{y}_{t+1}^\top]\big) \\
&= \frac{1}{2}\,\mathrm{tr}(\mathbf{\Gamma}\,\bar{\mathbf{W}}\,\bar{\mathbf{\Sigma}}\,\bar{\mathbf{W}}^\top) \;-\; \mathrm{tr}(\mathbf{\Gamma}\,\mathbf{\Sigma}_{y\bar{z}}\,\bar{\mathbf{W}}^\top) \;+\; \frac{1}{2}\,\mathrm{tr}(\mathbf{\Gamma}\,\mathbf{\Sigma}_{yy}).
\end{aligned}
$$

□

引理 A.6（经验 Fisher 加权二次型）。设 $M := B(N-1)$ 为训练对的数量，并用 $(\tau,\ t)$ 对其进行索引，其中 $\tau \in [B]$ 且 $t \in [N-1]$。将经验 Fisher 权重和经验二阶矩定义为

$$
\widehat{\mathbf{\Gamma}} \;:=\; \frac{1}{M}\sum_{\tau,t}\Big(\mathrm{Diag}(\mathbf{p}^{\mathrm{PO}}_{\tau,t}) - \mathbf{p}^{\mathrm{PO}}_{\tau,t}(\mathbf{p}^{\mathrm{PO}}_{\tau,t})^\top\Big), \quad \widehat{\bar{\mathbf{\Sigma}}} \;:=\; \frac{1}{M}\sum_{\tau,t}\bar{\mathbf{z}}_{\tau,t}\bar{\mathbf{z}}_{\tau,t}^\top,
$$

$$
\widehat{\mathbf{\Sigma}}_{y\bar{z}} \;:=\; \frac{1}{M}\sum_{\tau,t}\mathbf{y}_{\tau,t+1}\bar{\mathbf{z}}_{\tau,t}^\top, \qquad \widehat{\mathbf{\Sigma}}_{yy} \;:=\; \frac{1}{M}\sum_{\tau,t}\mathbf{y}_{\tau,t+1}\mathbf{y}_{\tau,t+1}^\top,
$$

其中 $\bar{\mathbf{z}}_{\tau,t} := (\bar{\mathbf{n}}_{\tau,t}^\top, \bar{\mathbf{g}}_{\tau,t}^\top)^\top$，$\bar{\mathbf{n}}_{\tau,t} = \mathbf{n}_{\tau,t}/t$，$\bar{\mathbf{g}}_{\tau,t} = \mathbf{g}_{\tau,t}/t$，且 $\mathbf{y}_{\tau,t+1} :=$ 投影算子$(\mathbf{s}^{\mathrm{PO}}_{\tau,t+1})$。设 $\bar{\mathbf{W}} = [\mathbf{W}_n\ \mathbf{W}_g]$ 如引理 A.5 中所示。则式 equation 4.1 中的经验 Fisher 加权损失具有如下二次型

$$
\widehat{\mathcal{L}}(\boldsymbol{\theta}) \;=\; \frac{1}{2}\,\mathrm{tr}\big(\widehat{\mathbf{\Gamma}}\,\bar{\mathbf{W}}\,\widehat{\bar{\mathbf{\Sigma}}}\,\bar{\mathbf{W}}^\top\big) \;-\; \mathrm{tr}\big(\widehat{\mathbf{\Gamma}}\,\widehat{\mathbf{\Sigma}}_{y\bar{z}}\,\bar{\mathbf{W}}^\top\big) \;+\; \frac{1}{2}\,\mathrm{tr}\big(\widehat{\mathbf{\Gamma}}\,\widehat{\mathbf{\Sigma}}_{yy}\big),
$$

特别地，最后一项与 $\boldsymbol{\theta}$ 无关。

证明。根据正文中的定义，

$$
\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{2M}\sum_{\tau,t}\Big\|\mathrm{Proj}\Big(\widehat{\mathbf{s}}_{\tau,t+1} - \mathbf{s}^{\mathrm{PO}}_{\tau,t+1}\Big)\Big\|^2_{\widehat{\mathbf{\Gamma}}}.
$$

根据引理 A.4，投影算子$(\widehat{\mathbf{s}}_{\tau,t+1}) = \mathbf{W}\,\bar{\mathbf{z}}_{\tau,t}$，因此

$$
\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{2M}\sum_{\tau,t}\big(\bar{\mathbf{W}}\bar{\mathbf{z}}_{\tau,t} - \mathbf{y}_{\tau,t+1}\big)^\top \widehat{\mathbf{\Gamma}}\,\big(\bar{\mathbf{W}}\bar{\mathbf{z}}_{\tau,t} - \mathbf{y}_{\tau,t+1}\big).
$$

展开二次项，并利用迹的线性性与 $\boldsymbol{b}^\top A \boldsymbol{b} = \mathrm{tr}(A\boldsymbol{b}\boldsymbol{b}^\top)$，

$$
\begin{aligned}
\widehat{\mathcal{L}}(\boldsymbol{\theta}) =& \frac{1}{2}\,\mathrm{tr}\Big(\widehat{\mathbf{\Gamma}}\,\bar{\mathbf{W}}\,\frac{1}{M}\sum_{\tau,t}\bar{\mathbf{z}}_{\tau,t}\bar{\mathbf{z}}_{\tau,t}^\top\,\bar{\mathbf{W}}^\top\Big) \;-\; \mathrm{tr}\Big(\widehat{\mathbf{\Gamma}}\,\frac{1}{M}\sum_{\tau,t}\mathbf{y}_{\tau,t+1}\bar{\mathbf{z}}_{\tau,t}^\top\,\bar{\mathbf{W}}^\top\Big) \\
&+\; \frac{1}{2}\,\mathrm{tr}\Big(\widehat{\mathbf{\Gamma}}\,\frac{1}{M}\sum_{\tau,t}\mathbf{y}_{\tau,t+1}\mathbf{y}_{\tau,t+1}^\top\Big) \\
=& \frac{1}{2}\,\mathrm{tr}\big(\widehat{\mathbf{\Gamma}}\,\bar{\mathbf{W}}\,\widehat{\bar{\mathbf{\Sigma}}}\,\bar{\mathbf{W}}^\top\big) \;-\; \mathrm{tr}\big(\widehat{\mathbf{\Gamma}}\,\widehat{\mathbf{\Sigma}}_{y\bar{z}}\,\bar{\mathbf{W}}^\top\big) \;+\; \frac{1}{2}\,\mathrm{tr}\big(\widehat{\mathbf{\Gamma}}\,\widehat{\mathbf{\Sigma}}_{yy}\big).
\end{aligned}
$$

□

引理 A.7（总体二阶矩在 S 上正定）。设 $\bar{\mathbf{z}}_t := \begin{pmatrix}\bar{\mathbf{n}}_t \\ \bar{\mathbf{g}}_t\end{pmatrix} \in \mathbb{R}^{2K}$，其中包含 $\bar{\mathbf{n}}_t = \frac{1}{t}\sum_{s=1}^{t}\mathbf{x}_s$，且 $\bar{\mathbf{g}}_t = \frac{1}{t}\sum_{s=1}^{t} r_s \mathbf{x}_s$，其中 $r_s = \mathbf{w}^\top \mathbf{x}_s + \epsilon_s$。假设 $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\ \tau_w^2 I_K)$ 且与 $\{\mathbf{x}_s\}$ 独立，并且 $\{\epsilon_s\}$，而 $\{\epsilon_s\}$ 是条件零均值的 σ_ϵ-次高斯鞅差序列。智能体采用 γ-混合探索。定义总体二阶矩

$$
\bar{\mathbf{\Sigma}} := \mathbb{E}[\bar{\mathbf{z}}_t \bar{\mathbf{z}}_t^\top].
$$

假设教师的参数（步长 c_0、奖励预条件器 $\mathbf{U}$）和问题参数满足

$$\frac{(1-\gamma)}{2}\, c_0\, \|\mathbf{U}\|_{\mathrm{op}}\, \sigma_\epsilon^2 \;<\; \tau_w\, \frac{\gamma}{K}. \tag{A.3}$$

那么，对于任意 $t \geq 2$，限制 $\bar{\mathbf{\Sigma}}\big|_S$ 在 $S := \mathbf{1}^\perp \oplus \mathbf{1}^\perp$ 上严格正定；特别地，

$$\lambda_{\min}\big(\bar{\mathbf{\Sigma}}\big|_S\big) \;>\; 0,$$

并且 $\bar{\mathbf{\Sigma}}\big|_S$ 在 S 上可逆。

证明。 我们在 $S := \mathbf{1}^\perp \oplus \mathbf{1}^\perp$ 上进行讨论。记

$$\bar{\mathbf{\Sigma}} = \mathbb{E}[\bar{\mathbf{z}}_t \bar{\mathbf{z}}_t^\top] = \begin{pmatrix} \mathbf{\Sigma}_{nn} & \mathbf{\Sigma}_{ng} \\ \mathbf{\Sigma}_{gn} & \mathbf{\Sigma}_{gg} \end{pmatrix}.$$

Set

$$a := \lambda_{\min}\big(\mathbf{\Sigma}_{nn}\big|_{\mathbf{1}^\perp}\big), \qquad c := \lambda_{\min}\big(\mathbf{\Sigma}_{gg}\big|_{\mathbf{1}^\perp}\big), \qquad b := \big\|\mathbf{\Sigma}_{ng}\big|_{\mathbf{1}^\perp}\big\|_{\mathrm{op}}.$$

将 $\mathbf{x}_s = \mathbf{p}_s + \mathbf{e}_s$ 分解为 $\mathbf{p}_s := \mathbb{E}[\mathbf{x}_s \mid \mathcal{F}_{s-1}]$，且 $\mathbb{E}[\mathbf{e}_s \mathbf{e}_s^\top \mid \mathcal{F}_{s-1}] = \mathbf{F}(\mathbf{p}_s)$。对于任意 $u \in \mathbf{1}^\perp$，

$$\mathbb{E}\big[(u^\top \bar{\mathbf{n}}_t)^2\big] \;\geq\; \frac{1}{t^2} \sum_{s=1}^{t} \mathbb{E}\big[u^\top \mathbf{F}(\mathbf{p}_s) u\big].$$

根据引理 A.1 以及混合探索，$\mathbf{F}(\mathbf{p}_s)\big|_{\mathbf{1}^\perp} \succeq (\gamma/K)\, I$，因此

$$a \;\geq\; \frac{1}{t^2} \sum_{s=1}^{t} \frac{\gamma}{K} \;=\; \frac{\gamma}{K\, t}. \tag{A.4}$$

将 $\bar{}\mathbf{g}_t = \bar{}\mathbf{g}_t^{(\mathrm{sig})} + \bar{}\mathbf{g}_t^{(\mathrm{noise})}$ 与 $\bar{}\mathbf{g}_t^{(\mathrm{sig})} = \frac{1}{t}\sum_{s=1}^{t} (\mathbf{w}^\top \mathbf{x}_s)\mathbf{x}_s =$ 对角矩阵($\bar{}\mathbf{n}_t$)w分开。对于任意 $v \in \mathbf{1}\ \perp$，

$$\mathbb{E}\big[(v^\top \bar{\mathbf{g}}_t^{(\mathrm{sig})})^2\big] = \tau_w^2\, \mathbb{E}\big[\|\operatorname{Diag}(\bar{\mathbf{n}}_t) v\|_2^2\big] = \tau_w^2 \sum_{i=1}^{K} \mathbb{E}[\bar{n}_{t,i}^2]\, v_i^2.$$

令 $N_{t,i} := \sum_{s=1}^{t} \mathbf{1}\{A_s = i\}$。由于 $N_{t,i}^2 \geq N_{t,i}$, $\mathbb{E}[\bar{n}_{t,i}^2] = \mathbb{E}[N_{t,i}^2]/t^2 \geq \mathbb{E}[N_{t,i}]/t^2 = \mathbb{E}[\bar{n}_{t,i}]/t$。混合探索可得 $\mathbb{E}[\bar{n}_{t,i}] \geq \gamma/K$，因此

$$c \;\geq\; \tau_w^2\, \frac{\gamma}{K\, t}. \tag{A.5}$$

加入半正定噪声协方差只会增大 $\mathbf{\Sigma}_{gg}$，因此等式 A.5 成立。我们有

$$\mathbf{\Sigma}_{ng} = \mathbb{E}[\bar{\mathbf{n}}_t \bar{\mathbf{g}}_t^\top] = \underbrace{\mathbb{E}\big[\bar{\mathbf{n}}_t (\bar{\mathbf{g}}_t^{(\mathrm{sig})})^\top\big]}_{=0} + \mathbb{E}\big[\bar{\mathbf{n}}_t (\bar{\mathbf{g}}_t^{(\mathrm{noise})})^\top\big].$$

第一项为零，因为 $\mathbf{w}$ 与 $(\bar{}\mathbf{n}_t, \{\epsilon_s\})$ 独立，且 $\mathbb{E}[\mathbf{w}] = 0$。对于第二项，

$$\mathbb{E}\big[\bar{\mathbf{n}}_t (\bar{\mathbf{g}}_t^{(\mathrm{noise})})^\top\big] = \frac{1}{t^2} \sum_{u=1}^{t} \sum_{s=1}^{t} \mathbb{E}[\mathbf{x}_u\, \epsilon_s\, \mathbf{x}_s^\top] = \frac{1}{t^2} \sum_{s=1}^{t-1} \sum_{u=s+1}^{t} \mathbb{E}[\mathbf{x}_u\, \epsilon_s\, \mathbf{x}_s^\top],$$

whe我们使用 $\mathbb{E}[\epsilon_s \mid \mathcal{F}_{s-1}] = 0$ 来消除 $u \leq s$。固定 $s < u$。考虑“世界 0”，其中 ϵ_s 是 se t to 0，而“世界 1”是真实世界。于是

$$\mathbb{E}[\mathbf{x}_u\, \epsilon_s\, \mathbf{x}_s^\top] = \mathbb{E}\big[(\mathbf{p}_u^{(1)} - \mathbf{p}_u^{(0)})\, \epsilon_s\, \mathbf{x}_s^\top\big],$$

由于 $\mathbb{E}[\mathbf{p}_u^{(0)}\, \epsilon_s \mathbf{x}_s^\top] = 0$。根据 $\eta_u = c_0/u$，投影 logits 满足（单步标准形式）

$$\mathbf{y}_u = \frac{c_0}{u-1}\big(\mathbf{V}\, \mathbf{n}_{u-1} + \mathbf{U}\, \mathbf{g}_{u-1}\big) \;\Rightarrow\; \Delta \mathbf{y}_u = \frac{c_0}{u-1}\, \mathbf{U}\, (\epsilon_s\, \mathbf{x}_s)$$

仅当 ϵ_s 受到扰动时。根据引理 A.2 以及方程 $(1-\gamma)$ 中的因子 3.3，

$$\|\mathbf{p}_u^{(1)} - \mathbf{p}_u^{(0)}\|_2 \;\leq\; (1-\gamma) \cdot \tfrac{1}{2}\, \|\Delta \mathbf{y}_u\|_2 \;\leq\; (1-\gamma) \cdot \tfrac{1}{2} \cdot \frac{c_0}{u-1}\, \|\mathbf{U}\|_{\mathrm{op}}\, |\epsilon_s|.$$

因此

$$\begin{aligned}\left\|\mathbb{E}[\mathbf{x}_u\,\epsilon_s\,\mathbf{x}_s^\top]\right\|_{\mathrm{op}} &\le \mathbb{E}\big[\|\mathbf{p}_u^{(1)}-\mathbf{p}_u^{(0)}\|_2\,|\epsilon_s|\big]\\ &\le \frac{(1-\gamma)}{2}\,\frac{c_0}{u-1}\,\|\mathbf{U}\|_{\mathrm{op}}\,\mathbb{E}[\epsilon_s^2]\\ &\le \frac{(1-\gamma)}{2}\,\frac{c_0}{u-1}\,\|\mathbf{U}\|_{\mathrm{op}}\,\sigma_\epsilon^2.\end{aligned}$$

将 $u=s+1,\dots,t$ 和 $s=1,\dots,t-1$ 相加，得到

$$\begin{aligned}b \;=\; \left\|\boldsymbol{\Sigma}_{ng}\big|_{\mathbf{1}^\perp}\right\|_{\mathrm{op}} &\le \frac{1}{t^2}\sum_{s=1}^{t-1}\sum_{u=s+1}^{t}\frac{(1-\gamma)}{2}\,\frac{c_0}{u-1}\,\|\mathbf{U}\|_{\mathrm{op}}\,\sigma_\epsilon^2\\ &\le \frac{(1-\gamma)}{2}\,c_0\,\|\mathbf{U}\|_{\mathrm{op}}\,\sigma_\epsilon^2\cdot\frac{t-1}{t^2}\\ &\le \frac{C_b}{t},\end{aligned}\tag{A.6}$$

其中 $C_b := \frac{(1-\gamma)}{2}\;\; c_0\,\|\mathbf{U}\|_{\mathrm{op}}\,\sigma_\epsilon^2$。

对于任意单位 $(u,v)\in S$,

$$\begin{pmatrix}u\\v\end{pmatrix}^\top \bar{\boldsymbol{\Sigma}}\Big|_S\begin{pmatrix}u\\v\end{pmatrix} \;\ge\; a\|u\|^2-2b\|u\|\,\|v\|+c\|v\|^2 \;\ge\; \lambda_{\min}\begin{pmatrix}a & -b\\ -b & c\end{pmatrix}.$$

Thus

$$\lambda_{\min}\big(\bar{\boldsymbol{\Sigma}}\big|_S\big) \;\ge\; \frac{a+c-\sqrt{(a-c)^2+4b^2}}{2}.\tag{A.7}$$

根据公式A.4、公式A.5和公式 A.6,

$$\sqrt{ac} \;=\; \tau_w\,\frac{\gamma}{K\,t},\qquad b\;\le\;\frac{C_b}{t}.$$

小交叉块条件公式 A.3 意味着 $b<\sqrt{ac}$(对所有 $t\ge 2$) ¯, 因此 2×2 矩阵($\begin{smallmatrix}a & -b\\ -b & c\end{smallmatrix}$) 是正定的, 并且公式 A.7 的右侧严格为正。因此$\bar{\boldsymbol{\Sigma}}\big|_S \succ 0$。 □

引理 A.8（ $^{-}\mathbf{z}\mathcal{N}\;\mathbf{0}\;\tau_w^2\;\;K$ $\quad$ $^{-}{}_t \;=\; {}^{-\top}_{t}\;{}^{-\top}_{t}\;^{\top}\in\mathbb{R}^{2K}$ $\quad$ $\sim$ 的样本二阶矩集中）。在假设 3.1 与 w (, I) 下，归一化统计量 z : (n , g) 是次高斯的，其与维度无关的 ψ_2–范数 L : $=\|^{-}\mathbf{z}_t\|_{\psi_2}$ 仅依赖于 $(\tau_w,\sigma_\epsilon,\gamma)$（而不依赖于 K）。设

$$\widehat{\bar{\boldsymbol{\Sigma}}} \;:=\; \frac{1}{M}\sum_{m=1}^{M}\bar{\mathbf{z}}_t^{(m)}\bar{\mathbf{z}}_t^{(m)\top},\qquad \bar{\boldsymbol{\Sigma}} \;:=\; \mathbb{E}[\bar{\mathbf{z}}_t\bar{\mathbf{z}}_t^\top],$$

其中 $\{^{-}\mathbf{z}_t^{(m)}\}_{m=1}^{M}$ 是 $^{-}\mathbf{z}_t$ 的独立同分布副本，由同一总体过程的独立轨迹采样生成（固定 γ）。令 $\sigma>0$ 为引理 A.7 中的总体下界，即 $\lambda_{\min}\big(\bar{\boldsymbol{\Sigma}}\big|_S\big)\ge\sigma$在 $S:=\mathbf{1}^\perp\oplus\mathbf{1}^\perp$。则存在普适常数 $C>0$，使得对任意 $\delta\in(0,1)$，以至少 $1-\delta$ 的概率,

$$\left\|\widehat{\bar{\boldsymbol{\Sigma}}}-\bar{\boldsymbol{\Sigma}}\right\|_{\mathrm{op}} \;\le\; C\,L^2\left(\sqrt{\frac{2K+\log(1/\delta)}{M}}\;+\;\frac{2K+\log(1/\delta)}{M}\right).\tag{A.8}$$

具体而言，如果

$$M\;\ge\;C_{\mathrm{mc}}\,\frac{2K+\log(1/\delta)}{\underline{\sigma}^2}\qquad\text{with}\qquad C_{\mathrm{mc}}:=16\,C^2\,L^4,$$

~然后是 $\left\|\bar{\boldsymbol{\Sigma}}-\underline{\bar{\boldsymbol{\Sigma}}}\right\|_{\mathrm{op}}\le\frac{1}{2}\,\sigma$，因此

$$\lambda_{\min}\Big(\widehat{\bar{\boldsymbol{\Sigma}}}\Big|_S\Big) \;\ge\; \tfrac{1}{2}\,\underline{\sigma}.\tag{A.9}$$

证明。 $\bar{\mathbf{z}}_t$ 的次高斯性可由以下几点得出：(i) $\bar{\mathbf{n}}_t = \frac{1}{t}\sum_{s=1}^{t}\mathbf{x}_s$ 是来自某个 γ-混合策略的独热向量的平均值，因此其各坐标均为次高斯，且 ψ_2–范数有界于一个仅取决于 γ的常数；(ii) $\bar{\mathbf{g}}_t = \frac{1}{t}\sum_{s=1}^{t} r_s\mathbf{x}_s$ 与 $r_s = \mathbf{w}^\top\mathbf{x}_s + \epsilon_s$有关，其中 $\mathbf{w}$ 是与 $\{\mathbf{x}_s\}$ 和 $\{\epsilon_s\}$独立的高斯变量，且 $\{\epsilon_s\}$ 在条件下是次高斯的。因此， $\bar{\mathbf{g}}_t$ 是次高斯向量之和，其 ψ_2–范数由 $(\tau_w, \sigma_\epsilon, \gamma)$ 控制；拼接后仍保持次高斯性，且 $L = \|\bar{\mathbf{z}}_t\|_{\psi_2} = O(1)$ 属于 $(\tau_w, \sigma_\epsilon, \gamma)$。

对于 $\mathbb{R}^d$ 中满足 $d = 2K$的独立同分布次高斯样本，标准的样本协方差算子范数偏差界（例如矩阵伯恩斯坦不等式／次高斯协方差集中）给出方程 A.8：

$$\big\|\widehat{\bar{\mathbf{\Sigma}}} - \bar{\mathbf{\Sigma}}\big\|_{\mathrm{op}} \;\le\; C\,L^2\left(\sqrt{\frac{d+\log(1/\delta)}{M}} \;+\; \frac{d+\log(1/\delta)}{M}\right).$$

若 $M \ge C_{\mathrm{mc}}(2K + \log(1/\delta))/\sigma^2$，则 $\|\bar{\mathbf{\Sigma}} - \bar{\mathbf{\Sigma}}\|_{\mathrm{op}} \le \sigma/2$。根据韦尔不等式，在限制到 S 的情况下，

$$\lambda_{\min}\Big(\widehat{\bar{\mathbf{\Sigma}}}\Big|_S\Big) \;\ge\; \lambda_{\min}\Big(\bar{\mathbf{\Sigma}}\Big|_S\Big) \;-\; \big\|\widehat{\bar{\mathbf{\Sigma}}} - \bar{\mathbf{\Sigma}}\big\|_{\mathrm{op}} \;\ge\; \underline{\sigma} - \tfrac{1}{2}\underline{\sigma} \;=\; \tfrac{1}{2}\underline{\sigma},$$

即方程 A.9。 □

引理 A.9（费舍尔双通道二次型：梯度与海森矩阵）。 考虑总体费舍尔加权二次型，来自方程 4.1 ，并以双通道参数化形式表示

$$\mathcal{L}(\bar{\mathbf{W}}) \;:=\; \frac{1}{2}\,\mathrm{tr}\big(\mathbf{\Gamma}\,\bar{\mathbf{W}}\,\bar{\mathbf{\Sigma}}\,\bar{\mathbf{W}}^\top\big) \;-\; \mathrm{tr}\big(\mathbf{\Gamma}\,\mathbf{\Sigma}_{y\bar{z}}\,\bar{\mathbf{W}}^\top\big) \;+\; \mathrm{const}, \tag{A.10}$$

其中 $\bar{\mathbf{W}} \in \mathbb{R}^{K\times 2K}$, $\bar{\mathbf{\Sigma}} = \mathbb{E}[\bar{\mathbf{z}}_t\bar{\mathbf{z}}_t^\top]$, $\mathbf{\Sigma}_{y\bar{z}} = \mathbb{E}[\mathbf{y}_{t+1}\bar{\mathbf{z}}_t^\top]$, 且 $\mathbf{\Gamma} = \mathbb{E}[\mathbf{F}(\mathbf{p}_{t+1}^{\mathrm{PO}})]$。那么

$$\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}}) \;=\; \mathbf{\Gamma}\big(\bar{\mathbf{W}}\,\bar{\mathbf{\Sigma}} - \mathbf{\Sigma}_{y\bar{z}}\big) \;=:\; \mathbf{E}(\bar{\mathbf{W}}) \in \mathbb{R}^{K\times 2K}, \tag{A.11}$$

$$\nabla^2\mathcal{L}(\bar{\mathbf{W}})[\Delta] \;=\; \mathbf{\Gamma}\,\Delta\,\bar{\mathbf{\Sigma}} \qquad (\Delta \in \mathbb{R}^{K\times 2K}). \tag{A.12}$$

证明。 令 $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathrm{const}$ 为

$$\mathcal{L}_1(\bar{\mathbf{W}}) := \tfrac{1}{2}\,\mathrm{tr}(\mathbf{\Gamma}\,\bar{\mathbf{W}}\,\bar{\mathbf{\Sigma}}\,\bar{\mathbf{W}}^\top), \qquad \mathcal{L}_2(\bar{\mathbf{W}}) := -\,\mathrm{tr}(\mathbf{\Gamma}\,\mathbf{\Sigma}_{y\bar{z}}\,\bar{\mathbf{W}}^\top).$$

使用弗罗贝尼乌斯内积 $\langle A\ B\rangle = A^\top B$ $d\bar{\mathbf{\Sigma}}^\top = d\bar{\mathbf{\Sigma}}^\top +$, tr( ) 和 (W W) (W) WW $\bar{\Sigma}(d\bar{\mathbf{W}})^\top$ 其中 $\bar{\mathbf{\Sigma}}^\top = \bar{\mathbf{\Sigma}}, \mathbf{\Gamma}^\top = \mathbf{\Gamma}$,

$$d\mathcal{L}_1 = \big\langle \mathbf{\Gamma}\,\bar{\mathbf{W}}\,\bar{\mathbf{\Sigma}},\, d\bar{\mathbf{W}}\big\rangle, \qquad d\mathcal{L}_2 = \big\langle -\,\mathbf{\Gamma}\,\mathbf{\Sigma}_{y\bar{z}},\, d\bar{\mathbf{W}}\big\rangle,$$

因此 $d\mathcal{L} = \langle\mathbf{\Gamma}(\mathbf{W}\mathbf{\Sigma} - \mathbf{\Sigma}_{y\bar{z}})\ d\mathbf{W}\rangle$，这就证明了方程 A.11。对于海森矩阵，令 $\mathbf{W}(\epsilon) = \bar{\mathbf{W}} + \epsilon\Delta$,

$$\frac{d}{d\epsilon}\,\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}}(\epsilon)) = \mathbf{\Gamma}\,\Delta\,\bar{\mathbf{\Sigma}},$$

从而得到方程 A.12。等价地， $d^2\mathcal{L}[\Delta,\Delta] = \mathrm{tr}(\Delta^\top\mathbf{\Gamma}\,\Delta\,\bar{\mathbf{\Sigma}}) = \|\mathbf{\Gamma}^{1/2}\Delta\bar{\mathbf{\Sigma}}^{1/2}\|^2_F \ge 0$. □

引理 A.10 （ 上的 PL 不等式）。 在混合探索（引理A.1）下，使得引理 A.7成立，总体费舍尔加权二次型

$$\mathcal{L}(\bar{\mathbf{W}}) \;=\; \tfrac{1}{2}\,\mathrm{tr}\big(\mathbf{\Gamma}\,\bar{\mathbf{W}}\,\bar{\mathbf{\Sigma}}\,\bar{\mathbf{W}}^\top\big) \;-\; \mathrm{tr}\big(\mathbf{\Gamma}\,\mathbf{\Sigma}_{y\bar{z}}\,\bar{\mathbf{W}}^\top\big) \;+\; \mathrm{const}$$

满足：对于任意全局极小值点$\mathbf{W}^\star$,

$$\mathcal{L}(\bar{\mathbf{W}}) - \mathcal{L}(\bar{\mathbf{W}}^\star) \;\le\; \frac{1}{2\mu}\,\big\|\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}})\big\|_F^2, \qquad \mu \;:=\; \lambda_{\min}^{+}\big(\mathbf{\Gamma}\big|_{\mathbf{1}^\perp}\big)\,\lambda_{\min}^{+}\big(\bar{\mathbf{\Sigma}}\big|_S\big), \tag{A.13}$$

其中 $S := \mathbf{1}^\perp \oplus \mathbf{1}^\perp$。此外，根据引理 A.1 和引理 A.7,

$$\lambda_{\min}^{+}\big(\mathbf{\Gamma}\big|_{\mathbf{1}^\perp}\big) \;\ge\; \frac{\gamma}{K}, \qquad \lambda_{\min}^{+}\big(\bar{\mathbf{\Sigma}}\big|_S\big) \;\ge\; \underline{\sigma} > 0, \qquad \Rightarrow \qquad \mu \;\ge\; \frac{\gamma}{K}\cdot\underline{\sigma}. \tag{A.14}$$

证明。 令 $\Delta := \bar{\mathbf{W}} - \bar{\mathbf{W}}^\star$，并定义 $\mathbf{X} := \mathbf{\Gamma}^{1/2}\,\Delta\,\bar{\mathbf{\Sigma}}^{1/2}$。根据引理 A.9, $\nabla\bar{\mathbf{W}}\mathcal{L}(\bar{\mathbf{W}}) = \mathbf{\Gamma}\,\Delta\,\bar{\mathbf{\Sigma}} = \mathbf{\Gamma}^{1/2}\mathbf{X}\,\bar{\mathbf{\Sigma}}^{1/2}$, $\mathcal{L}(\bar{\mathbf{W}}) - \mathcal{L}(\bar{\mathbf{W}}^\star) = \frac{1}{2}\mathrm{tr}(\mathbf{\Gamma}\,\Delta\,\bar{\mathbf{\Sigma}}\,\Delta^\top) = \frac{1}{2}\|\mathbf{X}\|_F^2$.对于任意半正定矩阵 A、B 以及任意矩阵 X, $\|\mathrm{A}^{1/2}\mathrm{XB}^{1/2}\|_F^2 = \mathrm{tr}(\mathrm{X}^\top\mathrm{AXB}) \ge \lambda_{\min}(\mathrm{A})\lambda_{\min}(\mathrm{B})\|\mathrm{X}\|_F^2$。将 $\mathrm{A} = \mathbf{\Gamma}\big|_{\mathbf{1}^\perp}$ 和 $\mathrm{B} = \bar{\mathbf{\Sigma}}\big|_S$代入此式，可得 $\big\|\nabla\bar{\mathbf{W}}\mathcal{L}(\bar{\mathbf{W}})\big\|_F^2 = \big\|\mathbf{\Gamma}^{1/2}\mathbf{X}\,\bar{\mathbf{\Sigma}}^{1/2}\big\|_F^2 \ge \lambda^{+}\min\big(\mathbf{\Gamma}\big|_{\mathbf{1}^\perp}\big)\,\lambda^{+}\min\big(\bar{\mathbf{\Sigma}}\big|_S\big)\,\|\mathrm{X}\|_F^2 = 2\,\mu\,(\mathcal{L}(\bar{\mathbf{W}}) - \mathcal{L}(\bar{\mathbf{W}}^\star))$, 这等价于 PL 不等式方程 A.13。界限方程 A.14 可由引理 A.1（给出 $\mathbf{\Gamma}\big|_{\mathbf{1}^\perp} \succeq (\gamma/K)I$）和引理 A.7（给出 $\bar{\mathbf{\Sigma}}\big|_S \succeq \sigma I$）推出。 □

A.2 奖励冲击鲁棒性

定理的证明 4.8。 在 s-CRN 耦合（定义 4.7）下，基准运行和扰动运行共享相同的任务$\mathbf{w}$、相同的采样均匀随机变量 $\{U_t\}$ 以及相同的噪声序列 $\{\epsilon_t\}$；它们的区别仅在于在轮次 s输入受冲击奖励 $\widetilde{r}_s = r_s + \delta_r$。定义

$$\Delta \mathbf{p}_{t+1} := \widetilde{\mathbf{p}}_{t+1} - \mathbf{p}_{t+1}, \quad \Delta \mathbf{n}_t := \widetilde{\mathbf{n}}_t - \mathbf{n}_t, \quad \Delta \mathbf{g}_t := \widetilde{\mathbf{g}}_t - \mathbf{g}_t, \quad a_{t+1} := \mathbb{E}[\|\Delta \mathbf{p}_{t+1}\| \mid \mathcal{F}_{s-1}],$$

以及归一化累积量 $\bar{\mathbf{n}}_t := \mathbf{n}_t/t$, $\bar{\mathbf{g}}_t := \mathbf{g}_t/t$ （其增量也类似）。计数和奖励的路径关系。对于 $\mathbf{x}_t = \mathbf{e}_{A_t}$ 和 $r_t = \langle \mathbf{w}, \mathbf{x}_t \rangle + \epsilon_t$,

$$\mathbf{n}_t = \sum_{u=1}^{t} \mathbf{x}_u, \qquad \mathbf{g}_t = \sum_{u=1}^{t} r_u \, \mathbf{x}_u = \mathrm{Diag}(\mathbf{n}_t)\, \mathbf{w} + \sum_{u=1}^{t} \epsilon_u \, \mathbf{x}_u.$$

令 $h_t := \sum_{u=1}^{t} \epsilon_u \, \mathbf{x}_u$ 和 $\widetilde{h}_t := \sum_{u=1}^{t} \epsilon_u \, \widetilde{\mathbf{x}}_u$。对于扰动运行，只有第 s 轮输入的奖励发生变化，因此

$$\widetilde{\mathbf{g}}_t = \mathrm{Diag}(\widetilde{\mathbf{n}}_t)\, \mathbf{w} + \widetilde{h}_t + \mathbf{1}\{t \geq s\}\, \delta_r \, \widetilde{\mathbf{x}}_s.$$

因此，对于 $t \geq s$,

$$\Delta \bar{\mathbf{g}}_t = \frac{1}{t}\, \mathrm{Diag}(\mathbf{w})\, \Delta \mathbf{n}_t + \frac{1}{t}\, \Delta h_t + \frac{1}{t}\, \delta_r \, \widetilde{\mathbf{x}}_s, \qquad \Delta h_t := \widetilde{h}_t - h_t. \tag{A.15}$$

投影 logits 的双通道线性化。采用与方程 3.2 对齐的（总体最优）单步标准形式，并投影消去 softmax 规范自由度,

$$\Delta \mathbf{y}_{t+1} := \mathrm{Proj}\Big(\widehat{\mathbf{s}}_{t+1}^{(\mathrm{pert})} - \widehat{\mathbf{s}}_{t+1}^{(\mathrm{base})}\Big) = \frac{c}{t}\, \Big(\mathbf{V}\, \Delta \mathbf{n}_t + \mathbf{U}\, \Delta \mathbf{g}_t\Big) = c\Big(\mathbf{V}\, \Delta \bar{\mathbf{n}}_t + \mathbf{U}\, \Delta \bar{\mathbf{g}}_t\Big). \tag{A.16}$$

Softmax 在 $1/2$ 上是 $\mathbf{1}^{\perp}$-利普希茨的，且 γ-混合方程 3.3 将敏感度缩放为 $(1-\gamma)$，因此

$$\|\Delta \mathbf{p}_{t+1}\| \;\leq\; (1-\gamma)\, \frac{1}{2}\, \|\Delta \mathbf{y}_{t+1}\|. \tag{A.17}$$

结合方程A.15–方程A.17 并对 $\mathcal{F}_{s-1}$ 进行条件化，得到

$$\begin{aligned} a_{t+1} \leq & \frac{c(1-\gamma)}{2}\, \mathbb{E}\Big[\, \big\|\big(\mathbf{V} + \mathbf{U}\, \mathrm{Diag}(\mathbf{w})\big)\, \Delta \bar{\mathbf{n}}_t + \mathbf{U}\, \Delta \bar{h}_t + \mathbf{U}\, \frac{\delta_r}{t}\, \widetilde{\mathbf{x}}_s \big\| \;\Big|\; \mathcal{F}_{s-1} \Big] \\ \leq & \frac{c(1-\gamma)}{2t}\, \Big(\|\mathbf{V} + \mathbf{U}\, \mathrm{Diag}(\mathbf{w})\|_{\mathrm{op}}\, \mathbb{E}[\|\Delta \mathbf{n}_t\| \mid \mathcal{F}_{s-1}] \\ & + \|\mathbf{U}\|_{\mathrm{op}}\, \mathbb{E}[\|\Delta h_t\| \mid \mathcal{F}_{s-1}] + \|\mathbf{U}\|_{\mathrm{op}}\, |\delta_r| \Big). \end{aligned} \tag{A.18}$$

CRN 耦合控制 $\mathbb{E}\|\Delta \mathbf{n}_t\|$ 和 $\mathbb{E}\|\Delta h_t\|$。 通过逆累积分布函数耦合，

$$\mathbb{P}\big(\widetilde{A}_u \neq A_u \mid \mathcal{F}_{u-1}\big) = \tfrac{1}{2}\|\Delta \mathbf{p}_u\|_1 \;\leq\; \tfrac{\sqrt{K}}{2}\, \|\Delta \mathbf{p}_u\|.$$

由于 $\|\Delta \mathbf{x}_u\| \leq \sqrt{2}$,

$$\mathbb{E}\big[\|\Delta \mathbf{x}_u\| \mid \mathcal{F}_{u-1}\big] \;\leq\; \sqrt{\tfrac{K}{2}}\, \|\Delta \mathbf{p}_u\|. \tag{A.19}$$

将方程 A.19 从 $u = s$到 t 求和，并对 $\mathcal{F}_{s-1}$进行条件化，

$$\mathbb{E}\big[\|\Delta \mathbf{n}_t\| \mid \mathcal{F}_{s-1}\big] \;\leq\; \sqrt{\tfrac{K}{2}} \sum_{u=s}^{t} a_u. \tag{A.20}$$

对于噪声累积量，利用零均值的 σ_ϵ 次高斯假设（因此 $\mathbb{E}[|\epsilon_u| \,|\, \mathcal{F}_{u-1}] \leq C_\epsilon := \sqrt{2/\pi}\, \sigma_\epsilon$）可得

$$\mathbb{E}\big[\|\Delta h_t\| \mid \mathcal{F}_{s-1}\big] \;\leq\; C_\epsilon \sqrt{\tfrac{K}{2}} \sum_{u=s}^{t} a_u. \tag{A.21}$$

一种 Gr¨onwall 型递推及其解。将方程代入方程A.20–方程A.21 中，代入方A.18，定义

$$a \;:=\; \frac{c(1-\gamma)}{2}\, \|\mathbf{U}\|_{\mathrm{op}}, \qquad b \;:=\; \frac{c(1-\gamma)}{2} \sqrt{\frac{K}{2}} \Big(\|\mathbf{V} + \mathbf{U}\, \mathrm{Diag}(\mathbf{w})\|_{\mathrm{op}} \;+\; C_\epsilon\, \|\mathbf{U}\|_{\mathrm{op}} \Big),$$

并得到，对于 $t \geq s$,

$$a_{t+1} \;\leq\; \frac{a}{t}\,|\delta_r| \;+\; \frac{b}{t}\sum_{u=s}^{t} a_u. \tag{A.22}$$

设 $S_t := \sum_{u=s}^{t} a_u$ 且满足 $S_{s-1} = 0$。则

$$S_{t+1} \;\leq\; \Big(1 + \frac{b}{t}\Big)\, S_t \;+\; \frac{a}{t}\,|\delta_r|. \tag{A.23}$$

展开等式 A.23 并使用标准的伽马比值界，可得一个常数 $C_b > 0$(仅取决于 b)，使得

$$S_t \;\leq\; a\,C_b\,|\delta_r| \sum_{u=s}^{t-1} \frac{1}{u}\left(\frac{t}{u}\right)^{b} \;\leq\; \frac{a\,C_b}{b}\left(\frac{t}{s}\right)^{b}|\delta_r|.$$

回到等式A.22,

$$a_{t+1} \;\leq\; \frac{a}{t}\,|\delta_r| \;+\; \frac{b}{t}\,S_t \;\leq\; \frac{a}{t}\,|\delta_r| \;+\; \frac{a\,C_b}{t}\left(\frac{t}{s}\right)^{b}|\delta_r| \;=\; \frac{a(1+C_b)}{s}\left(\frac{t}{s}\right)^{b-1}|\delta_r|.$$

因此，对于任意 $1 \leq s \leq t$,

$$\mathbb{E}\big[\,\|\Delta\widehat{\mathbf{p}}^{s}_{t+1}\|_2 \;\big|\; \mathcal{F}_{s-1}\big] \;\leq\; \frac{a(1+C_b)}{s}\left(\frac{t}{s}\right)^{b-1}|\delta_r|. \tag{A.24}$$

特别地，如果学习率常数 c 足够小，使得 $b < 1$，那么当 $t \to \infty$ 时，$\big(\frac{t}{s}\big)^{b-1} \to 0$，且单次影响衰减至零:

$$\lim_{t\to\infty} \mathbb{E}\big[\,\|\Delta\widehat{\mathbf{p}}^{s}_{t+1}\|_2 \;\big|\; \mathcal{F}_{s-1}\big] = 0.$$

证明至此完成。 □

A.3 KL 散度与费 Fisher 加权二次项

定理证明*4.1*。 回顾一下， N 表示第 4 节数据集构造中的轨迹长度，下面的期望对 $(\tau,\ t)$ 取平均，其中 $t \in \{1,\ldots,N-1\}$ 和 $\tau \in \mathcal{D}$。令$\widetilde{\mathbf{p}}(\mathbf{s}) := \mathrm{softmax}(\mathbf{s})$，并定义投影对数几率误差

$$\Delta_{t+1} \;:=\; \mathrm{Proj}\big(\widehat{\mathbf{s}}_{t+1} - \mathbf{s}^{\mathrm{PO}}_{t+1}\big) \in \mathbf{1}^{\perp}, \qquad \mathrm{Proj} := \mathbf{I} - \tfrac{1}{K}\mathbf{1}\mathbf{1}^{\top}.$$

将（softmax）费舍尔矩阵写为 $\mathbf{F}(\mathbf{p}) :=$ 对角矩阵$(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\top}$ 并回顾公式 equation 4.1 可得

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \mathbb{E}\left[\frac{1}{N-1}\sum_{t=1}^{N-1} \Delta_{t+1}^{\top}\boldsymbol{\Gamma}\,\Delta_{t+1}\right], \qquad \boldsymbol{\Gamma} \;:=\; \frac{1}{N-1}\mathbb{E}\left[\sum_{t=1}^{N-1}\mathbf{F}\big(\mathbf{p}^{\mathrm{PO}}_{t}\big)\right].$$

根据引理 A.1 （混合曲率），沿着教师的 γ-混合策略，我们在 $\mathbf{1}^{\perp}$ 上有谱夹逼:

$$\frac{\gamma}{K}\,\mathbf{I} \;\preceq\; \mathbf{F}\big(\mathbf{p}^{\mathrm{PO}}_{t}\big)\Big|_{\mathbf{1}^{\perp}} \;\preceq\; \frac{1}{2}\,\mathbf{I}. \tag{A.25}$$

由半正定锥的凸性，同样的界可转移到 $\boldsymbol{\Gamma}$在 $\mathbf{1}^{\perp}$ 上。

上界。 与均匀分布混合是一种马尔可夫核；根据f-散度的数据处理性质,

$$D_{\mathrm{KL}}\big(\mathbf{p}^{\mathrm{PO}}_{t+1}\,\|\,\widehat{\mathbf{p}}_{t+1}\big) \;\leq\; D_{\mathrm{KL}}\big(\widetilde{\mathbf{p}}(\mathbf{s}^{\mathrm{PO}}_{t+1})\,\|\,\widetilde{\mathbf{p}}(\widehat{\mathbf{s}}_{t+1})\big) = D_{\mathrm{KL}}\big(\widetilde{\mathbf{p}}(\mathbf{s}^{\mathrm{PO}}_{t+1})\,\|\,\widetilde{\mathbf{p}}(\mathbf{s}^{\mathrm{PO}}_{t+1} + \Delta_{t+1})\big).$$

令 $\phi(\mathbf{s}) := \log\sum_i e^{s_i}$; 则 $\nabla\phi = \mathrm{softmax}$ 且 $\nabla^2\phi(\mathbf{s}) = \mathbf{F}(\widetilde{\mathbf{p}}(\mathbf{s}))$。KL散度的布雷格曼积分形式给出，对任意 $\Delta \in \mathbf{1}^{\perp}$,

$$D_{\mathrm{KL}}\big(\widetilde{\mathbf{p}}(\mathbf{s})\,\|\,\widetilde{\mathbf{p}}(\mathbf{s}+\Delta)\big) = \int_0^1 (1-\tau)\,\Delta^{\top}\mathbf{F}\big(\widetilde{\mathbf{p}}(\mathbf{s}+\tau\Delta)\big)\,\Delta\,d\tau. \tag{A.26}$$

利用 $\|\mathbf{F}(\cdot)\big|_{\mathbf{1}^{\perp}}\|_{\mathrm{op}} \leq \frac{1}{2}$和 $\int_0^1 (1-\tau)\,d\tau = \frac{1}{2}$,

$$D_{\mathrm{KL}}\big(\mathbf{p}^{\mathrm{PO}}_{t+1}\,\|\,\widehat{\mathbf{p}}_{t+1}\big) \;\leq\; \tfrac{1}{4}\,\|\Delta_{t+1}\|_2^2.$$

由方程 A.25， $\Delta_{t+1}^{\top}\mathbf{\Gamma}\,\Delta_{t+1} \geq \frac{\gamma}{K}\|\Delta_{t+1}\|_2^2$，因此 $\|\Delta_{t+1}\|_2^2 \leq \frac{K}{\gamma}\,\Delta_{t+1}^{\top}\mathbf{\Gamma}\,\Delta_{t+1}$。对 t取平均并求期望，

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{t=1}^{N-1} D_{\mathrm{KL}}\big(\mathbf{p}_{t+1}^{\mathrm{PO}} \,\|\, \widehat{\mathbf{p}}_{t+1}\big)\right] \;\leq\; \frac{K}{4\gamma}\,\mathcal{L}(\boldsymbol{\theta}).$$

下界。 Pinsker不等式意味着 $D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) \geq \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1^2 \geq \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_2^2$. 令 $f(\mathbf{s}) := (1-\gamma)\,\widetilde{\mathbf{p}}(\mathbf{s}) + \gamma\,\mathbf{1}/K$，使得 $\mathbf{p}_{t+1}^{\mathrm{PO}} = f(\mathbf{s}_{t+1}^{\mathrm{PO}})$ 且 $\widehat{\mathbf{p}}_{t+1} = f(\widehat{\mathbf{s}}_{t+1})$。根据均值积分，

$$\mathbf{p}_{t+1}^{\mathrm{PO}} - \widehat{\mathbf{p}}_{t+1} = \int_0^1 \nabla f(\mathbf{s}_{t+1}^{\mathrm{PO}} + \tau\Delta_{t+1})\,\Delta_{t+1}\,d\tau = (1-\gamma)\int_0^1 \mathbf{F}\big(\widetilde{\mathbf{p}}(\mathbf{s}_{t+1}^{\mathrm{PO}} + \tau\Delta_{t+1})\big)\,\Delta_{t+1}\,d\tau,$$

由此 $\|\mathbf{p}_{t+1}^{\mathrm{PO}} - \widehat{\mathbf{p}}_{t+1}\|_2 \leq (1-\gamma)\cdot\frac{1}{2}\|\Delta_{t+1}\|_2$，因此

$$\|\Delta_{t+1}\|_2^2 \;\geq\; \frac{4}{(1-\gamma)^2}\,\|\mathbf{p}_{t+1}^{\mathrm{PO}} - \widehat{\mathbf{p}}_{t+1}\|_2^2.$$

利用方程 A.25 的上界， $\Delta_{t+1}^{\top}\mathbf{\Gamma}\,\Delta_{t+1} \leq \frac{1}{2}\,\|\Delta_{t+1}\|_2^2$，我们得到

$$\mathcal{L}(\boldsymbol{\theta}) \;\leq\; \frac{1}{2}\,\mathbb{E}\left[\frac{1}{N-1}\sum_{t=1}^{N-1}\|\Delta_{t+1}\|_2^2\right] \;\leq\; \frac{2}{(1-\gamma)^2}\,\mathbb{E}\left[\frac{1}{N-1}\sum_{t=1}^{N-1}\|\mathbf{p}_{t+1}^{\mathrm{PO}} - \widehat{\mathbf{p}}_{t+1}\|_2^2\right].$$

最后，结合 Pinsker 不等式可得

$$\mathbb{E}\left[\frac{1}{N-1}\sum_{t=1}^{N-1} D_{\mathrm{KL}}\big(\mathbf{p}_{t+1}^{\mathrm{PO}} \,\|\, \widehat{\mathbf{p}}_{t+1}\big)\right] \geq \frac{1}{2}\,\mathbb{E}\left[\frac{1}{N-1}\sum_{t=1}^{N-1}\|\mathbf{p}_{t+1}^{\mathrm{PO}} - \widehat{\mathbf{p}}_{t+1}\|_2^2\right] \geq \frac{(1-\gamma)^2}{4}\,\mathcal{L}(\boldsymbol{\theta}).$$

这就证明了双侧界。 □

A.4 策略优化的闭环模仿

定理证明*4.2*。 根据引理 A.4，投影后的学生的逻辑值是归一化统计量的线性函数：

$$\mathrm{Proj}\,\widehat{\mathbf{s}}_{t+1} \;=\; \bar{\mathbf{W}}\,\bar{\mathbf{z}}_t, \qquad \bar{\mathbf{W}} := [\,\mathbf{W}_n \;\; \mathbf{W}_g\,] \in \mathbb{R}^{K\times 2K}, \quad \bar{\mathbf{z}}_t := \begin{pmatrix}\bar{\mathbf{n}}_t \\ \bar{\mathbf{g}}_t\end{pmatrix} \in \mathbb{R}^{2K}.$$

对于由方程 3.2 生成的教师，且满足 $\eta_t = c/t$，同样有

$$\mathbf{y}_{t+1} := \mathrm{Proj}\,\mathbf{s}_{t+1}^{\mathrm{PO}} \;=\; \frac{c}{t}\,\mathrm{Proj}\big(\mathbf{V}\,\mathbf{n}_t + \mathbf{U}\,\mathbf{g}_t\big) \;=\; \bar{\mathbf{W}}_\star\,\bar{\mathbf{z}}_t, \qquad \bar{\mathbf{W}}_\star := \mathrm{Proj}\,[\,\mathbf{V} \;\; \mathbf{U}\,] \in \mathbb{R}^{K\times 2K},$$

因此，标签可由相同的双通道结构实现。
引理 A.5 给出了总体费舍尔加权二次型

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\,\mathrm{tr}(\mathbf{\Gamma}\,\bar{\mathbf{W}}\,\bar{\mathbf{\Sigma}}\,\bar{\mathbf{W}}^{\top}) - \mathrm{tr}(\mathbf{\Gamma}\,\mathbf{\Sigma}_{y\bar{z}}\,\bar{\mathbf{W}}^{\top}) + \mathrm{const}, \qquad \mathbf{\Gamma} := \mathbb{E}[\mathbf{F}(\mathbf{p}_{t+1}^{\mathrm{PO}})], \;\; \bar{\mathbf{\Sigma}} := \mathbb{E}[\bar{\mathbf{z}}_t\bar{\mathbf{z}}_t^{\top}].$$

利用可实现性 $\mathbf{y}_{t+1} = \bar{\mathbf{W}}_\star\bar{\mathbf{z}}_t$，我们有 $\mathbf{\Sigma}_{y\bar{z}} = \mathbb{E}[\mathbf{y}_{t+1}\bar{\mathbf{z}}_t^{\top}] = \mathbb{E}[\,\bar{\mathbf{W}}_\star\bar{\mathbf{z}}_t\bar{\mathbf{z}}_t^{\top}] = \bar{\mathbf{W}}_\star\bar{\mathbf{\Sigma}}$。代入后得到配方后的 形式

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\,\mathrm{tr}\Big(\mathbf{\Gamma}\,(\bar{\mathbf{W}} - \bar{\mathbf{W}}_\star)\,\bar{\mathbf{\Sigma}}\,(\bar{\mathbf{W}} - \bar{\mathbf{W}}_\star)^{\top}\Big) + \mathrm{const}.$$

根据引理 A.1（混合曲率），$\mathbf{\Gamma}\big|_{\mathbf{1}^{\perp}} \succeq (\gamma/K)\,I$；根据引理 A.7， $\bar{\mathbf{\Sigma}}\big|_S \succeq \sigma\,I$在 $S := \mathbf{1}^{\perp} \oplus \mathbf{1}^{\perp}$在假设 3.1下。因此，该二次型在 强凸 于 $\mathbf{W}$ 在 S 且具有唯一的 极小值$\mathbf{W} = \mathbf{W}_\star$。因此，对于任意历史前缀 $\mathcal{H}_t$，

$$\mathrm{Proj}\,\widehat{\mathbf{s}}_{t+1}(\mathcal{H}_t;\boldsymbol{\theta}^\star) = \bar{\mathbf{W}}_\star\,\bar{\mathbf{z}}_t = \mathrm{Proj}\,\mathbf{s}_{t+1}^{\mathrm{PO}}(\mathcal{H}_t).$$

由于 softmax 对加性常数不变（投影后的 logits 恰好消除了 $\mathrm{span}\{\mathbf{1}\}$ 中的规范自由度），因此诱导策略相同： $\widehat{\mathbf{p}}_{t+1}(\mathcal{H}_t;\boldsymbol{\theta}^\star) = \mathbf{p}_{t+1}^{\mathrm{PO}}(\mathcal{H}_t)$。 □

定理 4.3 的证明。根据引理 A.4，投影后的学生 logits 是归一化统计量的线性函数：$\mathrm{Proj}\hat{\mathbf{s}}_{t+1} = \bar{\mathbf{W}}\,\bar{\mathbf{z}}_t$，其中 $\bar{\mathbf{W}} = [\mathbf{W}_n\,\mathbf{W}_g]$，且 $\bar{\mathbf{z}}_t = (\bar{\mathbf{n}}_t^\top, \bar{\mathbf{g}}_t^\top)^\top$。令 $\mathbf{y}_{t+1} := \mathrm{Proj}\,(\mathbf{s}_{t+1}^{\mathrm{PO}})$。引理 A.5 和引理 A.6 给出了总体和经验 Fisher 加权二次型，而定理 4.2（在假设 3.1 下）保证了可实现性：存在唯一的 $\mathbf{W}_\star$，使得对所有前缀都有 $\mathbf{y}_{t+1} = \bar{\mathbf{W}}_\star \bar{\mathbf{z}}_t$。此外，引理 A.1 推出 $\bar{\mathbf{\Gamma}}\big|_{\mathbf{1}^\perp} \succeq (\gamma/K)I$，引理 A.7 则推出 $\bar{\mathbf{\Sigma}}\big|_S \succeq \sigma I$，其中 $\sigma \asymp c_\lambda/t$ 在 $S := \mathbf{1}^\perp \oplus \mathbf{1}^\perp$ 上，因此，关于 $\bar{\mathbf{W}}$ 的总体二次型在 S 上强凸，并具有唯一极小值点 $\mathbf{W}_\star$。对于经验问题，引理 A.8 给出（针对独立同分布的 rollout）

$$\|\hat{\bar{\mathbf{\Sigma}}} - \bar{\mathbf{\Sigma}}\|_{\mathrm{op}} \;\le\; C_1 L^2 \Big(\sqrt{\tfrac{2K+\log(1/\delta)}{M}} + \tfrac{2K+\log(1/\delta)}{M} \Big).$$

取 $M \ge C\,t^2(2K+\log(1/\delta))/c_\lambda^2$ (取足够大的绝对常数 C，其取值依赖于 C_1，L) 可保证 $\hat{\bar{\mathbf{\Sigma}}}\big|_S \succeq \frac{1}{2}\underline{\sigma} I$，而 $\hat{\bar{\mathbf{\Gamma}}}\big|_{\mathbf{1}^\perp} \succeq (\gamma/K)I$ 则由混合曲率得到。将可实现标签 $\mathbf{y}_{t+1} = \bar{\mathbf{W}}_\star \bar{\mathbf{z}}_t$ 代入引理 A.6可得

$$\hat{\mathcal{L}}(\bar{\mathbf{W}}) = \tfrac{1}{2}\,\mathrm{tr}\big(\hat{\bar{\mathbf{\Gamma}}}\,(\bar{\mathbf{W}} - \bar{\mathbf{W}}_\star)\,\hat{\bar{\mathbf{\Sigma}}}\,(\bar{\mathbf{W}} - \bar{\mathbf{W}}_\star)^\top\big) + \mathrm{const},$$

因此，经验极小化解满足 $\bar{\mathbf{W}} = \bar{\mathbf{W}}_\star$，只要 $\bar{\mathbf{\Sigma}}\big|_S \succ 0$ (此结论以概率$\ge 1-\delta$成立)。因此 $\mathrm{Proj}\hat{\mathbf{s}}_{t+1} = \mathrm{Proj}\,\mathbf{s}_{t+1}^{\mathrm{PO}}$，进而 $\hat{\mathbf{p}}^{t+1}(\mathcal{H}_t;\,\widehat{\boldsymbol{\theta}}) = \mathbf{p}_{t+1}^{\mathrm{PO}}(\mathcal{H}_t)$，其中测试历史 $\mathcal{H}_t$ 固定。对于期望失配，在高概率事件上误差为 0；在其补集上，粗略界结合 $(1-\gamma)$ 混合因子可得 $\|\hat{\mathbf{p}}^{t+1} - \mathbf{p}_{t+1}^{\mathrm{PO}}\|_2^2 \le 2(1-\gamma)^2$,因此

$$\mathbb{E}_{\mathrm{train}}\big[\|\widehat{\mathbf{p}}_{t+1} - \mathbf{p}_{t+1}^{\mathrm{PO}}\|_2^2\big] \;\le\; 0\cdot(1-\delta) + 2(1-\gamma)^2\cdot\delta = 2(1-\gamma)^2\delta.$$

这就证明了该定理。 □

A.5 Fisher 训练的双通道 LSA 的收敛性

我们分析方程$\mathbf{W}$ A.10 中 Fisher 加权二次目标函数在双通道算子A.10上的连续时间梯度流。

定理 A.11 （$\mathbf{W}$流的指数收敛）。假设采用混合探索，且假设3.1、引理A.1和引理A.7 均成立。考虑以下针对总体*Fisher* 加权二次型的梯度流 $\mathcal{L}$:

$$\dot{\bar{\mathbf{W}}}(t) \;=\; -\nabla_{\bar{\mathbf{W}}}\mathcal{L}\big(\bar{\mathbf{W}}(t)\big) \;=\; -\bar{\mathbf{\Gamma}}\big(\bar{\mathbf{W}}(t)\,\bar{\mathbf{\Sigma}} - \mathbf{\Sigma}_{y\bar{z}}\big), \tag{A.27}$$

从任意 $\bar{\mathbf{W}}(0)$ 初始化。令 $\bar{\mathbf{W}}^\star$ 为 $\mathcal{L}$ 的全局极小值点。则沿梯度流，$\mathcal{L}\big(\bar{\mathbf{W}}(t)\big)$ 严格递减，且

$$\mathcal{L}\big(\bar{\mathbf{W}}(t)\big) - \mathcal{L}(\bar{\mathbf{W}}^\star) \;\le\; \exp(-2\mu\,t)\,\big(\mathcal{L}\big(\bar{\mathbf{W}}(0)\big) - \mathcal{L}(\bar{\mathbf{W}}^\star)\big), \tag{A.28}$$

with

$$\mu \;:=\; \lambda_{\min}^{+}\big(\bar{\mathbf{\Gamma}}\big|_{\mathbf{1}^\perp}\big)\;\lambda_{\min}^{+}\big(\bar{\mathbf{\Sigma}}\big|_S\big), \qquad S := \mathbf{1}^\perp \oplus \mathbf{1}^\perp,$$

并且根据引理A.1 和引理A.7,

$$\lambda_{\min}^{+}\big(\bar{\mathbf{\Gamma}}\big|_{\mathbf{1}^\perp}\big) \;\ge\; \frac{\gamma}{K}, \qquad \lambda_{\min}^{+}\big(\bar{\mathbf{\Sigma}}\big|_S\big) \;\ge\; \underline{\sigma} \;>\; 0, \qquad \Rightarrow \qquad \mu \;\ge\; \frac{\gamma}{K}\cdot\underline{\sigma}.$$

定理证明*A.11*. 根据引理A.9，我们有

$$\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}}) = \bar{\mathbf{\Gamma}}(\bar{\mathbf{W}}\bar{\mathbf{\Sigma}} - \mathbf{\Sigma}_{y\bar{z}}),$$

因此，梯度流 $\dot{\bar{\mathbf{W}}}(t) = -\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}}(t))$ 是良定义的。沿着轨迹，

$$\frac{d}{dt}\,\mathcal{L}\big(\bar{\mathbf{W}}(t)\big) = \big\langle \nabla_{\bar{\mathbf{W}}}\mathcal{L}\big(\bar{\mathbf{W}}(t)\big),\, \dot{\bar{\mathbf{W}}}(t)\big\rangle = -\big\|\nabla_{\bar{\mathbf{W}}}\mathcal{L}\big(\bar{\mathbf{W}}(t)\big)\big\|_F^2 \;\le\; 0,$$

因此 $\mathcal{L}(\bar{\mathbf{W}}(t))$ 单调不增，并且当且仅当 $\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}}(t)) = 0$ 时严格递减。

接下来，在受限子空间上应用 Polyak–Łojasiewicz（PL）不等式 $S := \mathbf{1}^\perp \oplus \mathbf{1}^\perp$（引理A.10）。令

$$\mu \;:=\; \lambda_{\min}^{+}\big(\bar{\mathbf{\Gamma}}\big|_{\mathbf{1}^\perp}\big)\;\lambda_{\min}^{+}\big(\bar{\mathbf{\Sigma}}\big|_S\big).$$

¯那么，对于任意全局极小值点$\mathbf{W}^\star$,

$$\mathcal{L}(\bar{\mathbf{W}}) - \mathcal{L}(\bar{\mathbf{W}}^\star) \;\le\; \frac{1}{2\mu}\,\big\|\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}})\big\|_F^2.$$

将其与能量衰减恒等式相结合，对所有 $t \ge 0$，可得

$$\frac{d}{dt}\,\big(\mathcal{L}(\bar{\mathbf{W}}(t)) - \mathcal{L}(\bar{\mathbf{W}}^\star)\big) \;=\; -\,\big\|\nabla_{\bar{\mathbf{W}}}\mathcal{L}(\bar{\mathbf{W}}(t))\big\|_F^2 \;\le\; -\,2\mu\,\big(\mathcal{L}(\bar{\mathbf{W}}(t)) - \mathcal{L}(\bar{\mathbf{W}}^\star)\big).$$

根据 Gr¨onwall 不等式，

$$\mathcal{L}\big(\bar{\mathbf{W}}(t)\big) - \mathcal{L}(\bar{\mathbf{W}}^\star) \;\le\; \exp(-2\mu\, t)\,\big(\mathcal{L}\big(\bar{\mathbf{W}}(0)\big) - \mathcal{L}(\bar{\mathbf{W}}^\star)\big).$$

最后，根据混合探索（引理 A.1）以及 S 上总体二阶矩的下界（引理A.7），

$$\lambda_{\min}^{+}\big(\mathbf{\Gamma}\big|_{\mathbf{1}^\perp}\big) \;\ge\; \frac{\gamma}{K}, \qquad \lambda_{\min}^{+}\big(\bar{\mathbf{\Sigma}}\big|_{S}\big) \;\ge\; \underline{\sigma} \;>\; 0,$$

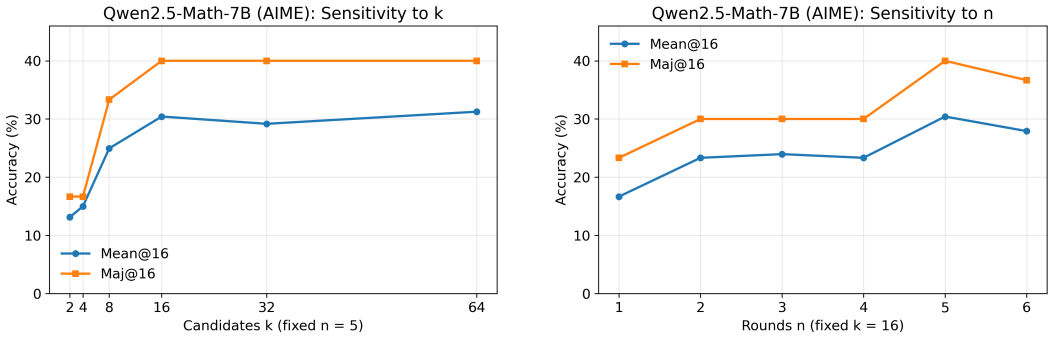
这意味着 $\mu \;\ge\; (\gamma/K)\;\underline{\sigma}$。除驻点外，损失严格下降；根据引理 A.9 以及受限正定性，这些点与 $\mathbf{1}^\perp$上的全局极小点重合。

□

B 实验细节

本节将补充介绍实验设置的详细信息，包括我们的方法和基线方法所用的具体超参数，以及所有实验使用的硬件和软件环境。我们还将呈现补充实验结果，以完善正文内容。B.1 超参数敏感性。

图4 展示了上下文内优化轮数（n）和每轮候选项数（k）变化对 AIME 2024 的影响。如图4a所示，候选项数量（k）从 2 增至 64 时，性能显著提升，Maj@16 准确率增加了一倍以上，之后便趋于饱和。增加轮数（n）总体上也有益，直至在 $n = 5$附近达到峰值。附录C.2（表 8）报告了补充网格及相应的延迟/显存指标。



(a) 改变候选项数量 k（固定 $n = 5$）　　(b) 改变轮数 n（固定 $k = 16$）

图 4：Qwen2.5-Math-7B 在 AIME 2024 上使用 ME-ICPO 的超参数敏感性。

B.2 超参数与实现细节

ME-ICPO（我们的方法）。对于我们的方法，我们采用正文中描述的主要设置（$n = 5$ 优化轮次，$k = 16$ 每轮候选项数）。候选思维链采用温度 0.6 和 top-p 值 0.95 进行采样。摘要步骤是管理上下文长度的关键组件，采用贪心解码，并提示模型生成一份约 100 个词元的简洁摘要，硬性上限为 500 个词元，重点关注推理策略。用于熵计算的预测答案分布同样采用温度 0.6 进行估计。所有系统提示和摘要提示的完整文本见附录D。

B.3 硬件与环境配置

所有实验均在一台配备 8 个 NVIDIA L40S GPU 的服务器节点上进行，每个 GPU 配备 48GB HBM2e 显存。我们的实现使用 PyTorch 2.2、vLLM 0.10.0 和 Transformer 4.55.0。操作系统为 Red Hat Enterprise Linux 9.6，环境通过 Conda 管理，使用 Python 3.11.7 和 CUDA 12.9。

B.4 补充实验结果

Thi本节包含在 rebuttal 阶段补充的结果与比较 t o add回答具体的计算和可扩展性问题。

B.4.1 计算成本与基线比较

为评估效率，我们将 ME-ICPO 与标准的提示式测试时搜索基线——思维树 (ToT) (Yao 等人，2023a) 和蒙特卡洛树精炼 (MCTR) (Zhang 等人，2024a)——以及基于训练的测试时扩展方法 TTRL (Zuo 等人，2025) 进行比较。为确保公平，ME-ICPO 使用固定配置，每轮进行 N=5 轮精炼，每轮采样 k=16 个样本，并对 5 个随机种子的结果取平均。我们报告了 AIME 2024 上的平均推理时间（每道问题耗时，单位为秒）。由于 ToT/MCTR 使用了不同的搜索深度和分支因子，其运行时间无法严格匹配；对于 TTRL，我们控制实验，使其使用与 ME-ICPO 相近的 GPU 小时数。如表 3 和表 4所示，ME-ICPO 在具有竞争力的计算预算下取得了顶尖的准确率和 Mean@16，体现了底层策略优化机制带来的原则性提升。浅层搜索的提示方法 (ToT/MCTR) 在搜索深度受限时速度可能更快，但准确率明显落后；而在实际运行时间预算匹配的情况下，ME-ICPO 的准确率/Mean@16 高于 TTRL，展现出更高效的扩展能力。更重要的是，ICPO 从机制上解释了测试时自我精炼是策略优化，而非对排行榜进行启发式调整。

B.4.2 前沿模型的可扩展性

为展示我们框架的通用性，我们在近期的前沿模型上评估了 ME-ICPO，包括 Qwen3-4B-Instruct (Yang 等人，2024b)（一种长链式思维专用模型），以及 Gemini- 2.5 系列（Pro 和 Flash）(Comanici 等人，2025）。表 5 表明，ME-ICPO 在这些不同架构上均能持续提升性能，证实了其可扩展性。

B.4.3 更难任务泛化

我们进一步评估了模型在难度极高的高中及大学水平数学竞赛任务上的表现，具体包括哈佛-麻省理工数学竞赛（HMMT）（Balunovi´c 等人，2025）和 APEX-shortlist 数据集（Balunovi´c 等人，2025）。表6 和表7 的结果表明，即使在基线解题率较低的任务上，ME-ICPO 也表现出稳健性。

表3：与推理和训练基线方法相比的计算成本与性能比较（Mean@16，Qwen2.5-Math-7B）。

方法	AIME-2024	AMC	MATH（平均）	时间（秒/问题）
ToT（自我评估）	4.38	16.19	12.51	708
ToT（多数投票）	19.58	29.37	35.63	**363**
MCTR	4.60	1.20	17.20	1758
TTRL	27.20	45.18	46.83	1253
ME-ICPO（本文）	**30.42**	**47.06**	**54.71**	1152

表4：计算性能比较（准确率，Qwen2.5-Math-7B）。

方法	AIME-2024	AMC	MATH（平均值）
ToT（自我评估）	4.40	18.10	10.74
ToT（多数投票）	19.30	29.40	35.91
MCTR	23.30	2.40	33.82
TTRL	30.00	43.37	45.11
ME-ICPO（我们的方法）	**30.05**	**47.20**	**47.30**

表 5：前沿/长思维链模型在 AIME 2024 上的结果。

模型	方法	Mean@16 (%)	准确率 (%)
Qwen3-4B-Instruct	Base	20.62	20.59
Qwen3-4B-Instruct	**ME-ICPO（我们的方法）**	**57.71**	**57.67**
Gemini-2.5-Pro	Base	58.54	56.60
Gemini-2.5-Pro	**ME-ICPO（我们的方法）**	**79.17**	**80.00**
Gemini-2.5-Flash	Base	35.21	35.42
Gemini-2.5-Flash	**ME-ICPO（我们的方法）**	**76.46**	**76.47**

表6：使用 Qwen2.5-Math-7B 在更难的基准测试（HMMT / APEX）上的结果。

方法	HMMT Mean@16（%）	HMMT 准确率（%）	APEX Mean@16（%）	APEX 准确率（%）
Base	**1.04**	0.67	2.55	2.61
ME-ICPO（我们的方法）	0.42	**1.33**	**4.59**	**4.57**

表7：使用 Gemini-2.5-Flash 在更难的基准测试（HMMT / APEX）上的结果。

方法	HMMT Mean@16（%）	HMMT 准确率（%）	APEX Mean@16（%）	APEX 准确率（%）
Base	14.79	14.76	14.68	18.33
ME-ICPO（我们的方法）	**43.12**	**43.14**	**17.18**	**20.00**

C 详细复杂度分析

C.1 理论时间与显存复杂度推导

我们形式化定义仅解码器 Transformer 的计算模型，并给出 ME-ICPO（仅前向、摘要感知历史）和 TTRL（Zuo 等人，2025）（基于反向传播的测试时强化学习）的精确渐近推导。下文的论述和证明与正文中的分析完全一致；我们仅补充简短的衔接说明。

设定与基本开销。 我们分析一个具有 L 层、宽度为 d（参数量为 $|\theta| \asymp Ld^2$）的仅解码器 Transformer。对于单个测试实例，初始提示长度（问题陈述+ 模板）为 s_0。每条采样的思维链（CoT）平均长度为 ℓ。每轮采样 k个候选；总轮次数为 n。令 $\beta := \ell + r$ 表示每个（$\mathbf{x}$ r）对附加的词元数（奖励占位符 r 为 $O(1)$，因此 $\beta \asymp \ell$）。因此，第 t 轮开始时的提示长度为 $T_{t-1} := s_0 + \beta(t-1)$。我们用 $\kappa \in [2,3]$ 表示反向/前向浮点运算次数比（一次反向传播的开销是一次前向传播的 κ 倍），用 g 表示 TTRL 中每轮的策略优化步数。基本开销（忽略常数）如下：

$$\text{full-sequence forward at length } T: \quad C_{\mathrm{fwd}}(T) = \Theta\big(L(T^2 d + T d^2)\big),$$

$$\text{从前缀 } T \text{ 自回归解码 } \ell \text{ 词元: } C_{\mathrm{dec}}(T,\ \ell) = \Theta\Big(L\big((T\ell + \ell^2)\ d + \ell d^2\big)\Big),$$

$$\text{在长度为 } T \text{ 时执行一个训练步骤（教师强制前向 + 反向）: } C_{\mathrm{train}}(T) = (1+\kappa)\ C_{\mathrm{fwd}}(T)。$$

对于 ME-ICPO，单步前瞻得分是在刚生成的分支上通过追加固定数量的词元来计算的，因此其代价是一次增量前向计算

$$C^{\mathrm{inc}}_{\mathrm{score}}(T,\ell) = \Theta\big(L\big((T+\ell)\,d + d^2\big)\big),$$

而不是在长度为 $T+\ell$ 时进行完整的重新计算。

C.1.1 ME-ICPO

定理 C.1（ME-ICPO 的时间复杂度）。 每轮进行一次长度为 T_{t-1} 的共享预填充，从该前缀生成 k 个候选解码，并对每个候选进行增量式分支内评分，则经过 n 轮后的总时间满足

$$T_{\mathrm{ME}} = \Theta\Big(Ld\,\beta^2\,n^3 \ +\ k\,Ld\,\beta^2\,n^2 \ +\ Ld^2\,\beta\,n^2 \ +\ k\,Ld^2\,\beta\,n\Big),$$

并且，使用 $\beta \asymp \ell$,

$$T_{\mathrm{ME}} = \Theta\big(Ld\,\ell^2 n^3 + k\,Ld\,\ell^2 n^2\big) \;+\; \Theta\big(Ld^2(\ell n^2 + k\,\ell n)\big).$$

证明。 总时间复杂度 T_{ME} 是对所有 n 轮次中上下文预填充成本（S_{fwd}）、候选项解码成本（S_{dec}）以及候选项评分成本（S_{score}）的总和。

$$T_{\mathrm{ME}} = \sum_{t=1}^{n} C_{\mathrm{fwd}}(T_{t-1}) + \sum_{t=1}^{n} k\,C_{\mathrm{dec}}(T_{t-1},\ell) + \sum_{t=1}^{n} k\,C^{\mathrm{inc}}_{\mathrm{score}}(T_{t-1},\ell) =: S_{\mathrm{fwd}} + S_{\mathrm{dec}} + S_{\mathrm{score}}.$$

我们首先对各轮次求和，然后确定 n 中的最高阶项，以此分析每个组成部分。第 t 轮的提示长度为 $T_{t-1} = s_0 + \beta(t-1)$，我们使用标准求和公式 $\sum_{r=0}^{n-1} r = \frac{n(n-1)}{2} = \Theta(n^2)$ 和 $\sum_{r=0}^{n-1} r^2 = \frac{(n-1)n(2n-1)}{6} = \Theta(n^3)$。

预填充成本 S_{fwd} 为：

$$\begin{aligned}
S_{\mathrm{fwd}} &= \sum_{t=1}^{n} C_{\mathrm{fwd}}(T_{t-1}) \\
&\overset{\text{(a)}}{=} \Theta\left(Ld\sum_{t=1}^{n} T_{t-1}^2 + Ld^2\sum_{t=1}^{n} T_{t-1}\right) \\
&\overset{\text{(b)}}{=} \Theta\left(Ld\left(ns_0^2 + s_0\beta n(n-1) + \beta^2\frac{(n-1)n(2n-1)}{6}\right) + Ld^2\left(ns_0 + \beta\frac{n(n-1)}{2}\right)\right) \\
&\overset{\text{(c)}}{=} \Theta\big(Ld\,\beta^2 n^3\big) + \Theta\big(Ld^2\,\beta\,n^2\big).
\end{aligned}$$

其中，(a) 代入 $C_{\mathrm{fwd}}(T) = \Theta(L(T^2 d + Td^2))$ 的定义，并将常数提出求和符号；(b) 代入线性序列和二次序列求和的精确公式；(c) 确定表达式各部分在 n 中的最高阶项。

解码成本 S_{dec} 为：

$$\begin{aligned}
S_{\mathrm{dec}} &= \sum_{t=1}^{n} k\,C_{\mathrm{dec}}(T_{t-1},\ell) \\
&\overset{\text{(d)}}{=} \Theta\left(kLd\left(\ell\sum_{t=1}^{n} T_{t-1} + n\ell^2\right) + kLd^2(n\ell)\right) \\
&\overset{\text{(e)}}{=} \Theta\left(kLd\left(\ell\left(ns_0 + \beta\frac{n(n-1)}{2}\right) + n\ell^2\right) + kLd^2 n\ell\right) \\
&\overset{\text{(f)}}{=} \Theta\big(kLd\,\beta\ell n^2\big) + \Theta\big(kLd^2\,\ell n\big).
\end{aligned}$$

其中，(d) 代入 $C_{\mathrm{dec}}(T,\ell) = \Theta(L((T\ell + \ell^2)d + \ell d^2))$ 的定义；(e) 代入 T_{t-1} 的求和；(f) 则指出，对于 Ld 分量， n 中的主导项为二次项 $\Theta(kLd\,\beta\ell n^2)$。

增量评分成本 S_{score} 为：

$$\begin{aligned}
S_{\mathrm{score}} &= \sum_{t=1}^{n} k\,C^{\mathrm{inc}}_{\mathrm{score}}(T_{t-1},\ell) \\
&\overset{\text{(g)}}{=} \Theta\left(kLd\left(\sum_{t=1}^{n} T_{t-1} + n\ell\right) + kLd^2 n\right) \\
&\overset{\text{(h)}}{=} \Theta\left(kLd\,\beta n^2\right) + \Theta\big(kLd^2 n\big).
\end{aligned}$$

其中，(g) 代入 $C^{\mathrm{inc}}_{\mathrm{score}}(T,\ell) = \Theta(L((T+\ell)d + d^2))$ 的定义；(h) 则指出主导项为 $\Theta(kLd\,\beta n^2)$。

将三个组成部分的最高阶项合并，我们得到：

$$\begin{aligned}
T_{\mathrm{ME}} &= S_{\mathrm{fwd}} + S_{\mathrm{dec}} + S_{\mathrm{score}} \\
&= \Theta\big(Ld\,\beta^2 n^3 + Ld^2\,\beta\,n^2\big) + \Theta\big(kLd\,\beta\ell n^2 + kLd^2\,\ell n\big) + \Theta\big(kLd\,\beta n^2 + kLd^2 n\big).
\end{aligned}$$

按各项对 d 和 d^2的依赖关系进行分组，并注意到评分成本项被解码成本项主导，或与解码成本项渐近相等（因为 $\ell \geq 1$），总复杂度可简化为:

$$T_{\mathrm{ME}} = \underbrace{\Theta\big(Ld\,\beta^2 n^3\big)}_{\text{from } S_{\mathrm{fwd}}} + \underbrace{\Theta\big(kLd\,\beta\ell n^2\big)}_{\text{from } S_{\mathrm{dec}}} + \underbrace{\Theta\big(Ld^2\,\beta n^2\big)}_{\text{from } S_{\mathrm{fwd}}} + \underbrace{\Theta\big(kLd^2\,\ell n\big)}_{\text{from } S_{\mathrm{dec}}}.$$

利用近似 $\beta = \Theta(\ell)$，我们得到定理中所述的最终表达式。至此证明完成。

□

定理 C.2（ME-ICPO 的显存复杂度）. 若候选按顺序解码（无 k 路并行处理），则在 n 轮次上的峰值内存满足

$$M_{\mathrm{ME}} = \Theta(|\theta|) \;+\; \Theta\big(Ld\,(s_0 + \beta n)\big) = \Theta(|\theta|) \;+\; \Theta\big(Ld\,(s_0 + \ell n)\big).$$

若 $b \leq k$ 个候选并发解码，则第二项乘以 b。

证明。峰值显存复杂度是模型参数的静态内存与注意力键值缓存的最大动态内存之和。由于 ME-ICPO 是仅前向方法，它不需要为梯度或优化器状态分配内存。模型权重占用 $\Theta(|\theta|)$ 空间。对于具有 L 层、宽度为 d 的仅解码器 Transformer，长度为 T 的序列的键值缓存需要 $M_{\mathrm{KV}}(T) = \Theta(LdT)$ 内存。

轮次 t 开始时的上下文长度为 $T_{t-1} = s_0 + \beta(t-1)$。在解码长度为 ℓ 的候选期间，序列最长增长至 $T_{t-1} + \ell$。由于该长度随 t 单调递增，全局峰值出现在最终轮次（$t = n$），从而得到最大序列长度 $T_{n-1} + \ell = s_0 + \beta(n-1) + \ell = \Theta(s_0 + \beta n)$。

因此，顺序解码的总峰值内存是静态分量与最大动态分量之和:

$$M_{\mathrm{ME}} = \Theta(|\theta|) + M_{\mathrm{KV}}\big(\Theta(s_0 + \beta n)\big) = \Theta(|\theta|) + \Theta\big(Ld(s_0 + \beta n)\big).$$

使用近似 $\beta = \Theta(\ell)$可得到等价形式。若 $b \leq k$ 个候选被并发处理，每个并行分支都会维护自己的KV缓存，因此激活内存项随 b线性增长。

□

C.1.2 TTRL

定理 C.3（TTRL的时间复杂度）. 在每一轮次中，TTRL（i）从前缀 s_0采样 k 个候选，并且（ii）对长度为$s_0 + \ell$的序列使用教师强制执行 g 次策略优化步骤。n 个轮次的总时间为

$$T_{\mathrm{TTRL}} = \Theta\Big(n\,g\,k\,(1+\kappa)\,L\big((s_0+\ell)^2 d + (s_0+\ell)d^2\big)\Big)$$

并且每轮次的预填充与采样成本 $\Theta\big(L(s_0^2 d + s_0 d^2)\big)$ 和 $\Theta\big(kL((s_0\ell + \ell^2)d + \ell d^2)\big)$在 $g \geq 1$时均为低阶项。

证明。总时间复杂度 T_{TTRL}是 n 个轮次成本之和。在每一轮次中，该过程对前缀 s_0执行一次共享预填充，随后进行 k 次候选采样操作，最后对 k 个候选中的每一个执行 g 次训练步骤。因此单轮成本为 $C_{\mathrm{fwd}}(s_0) + k\,C_{\mathrm{dec}}(s_0, \ell) + gk\,C_{\mathrm{train}}(s_0 + \ell)$。$n$ 个轮次的总成本为:

$$T_{\mathrm{TTRL}} = n \cdot \Big(C_{\mathrm{fwd}}(s_0) + k\,C_{\mathrm{dec}}(s_0, \ell) + gk\,C_{\mathrm{train}}(s_0 + \ell)\Big).$$

我们代入仅解码器 Transformer 的标准复杂度公式，使用 $C_{\mathrm{train}}(T) = (1+\kappa)C_{\mathrm{fwd}}(T)$，其中 κ 是反向传播与前向传播的浮点运算次数比。这会得到三项:

$$T_{\mathrm{TTRL}} = \underbrace{\Theta\Big(nL(s_0^2 d + s_0 d^2)\Big)}_{\text{Prefill Cost}} + \underbrace{\Theta\Big(nkL((s_0\ell + \ell^2)d + \ell d^2)\Big)}_{\text{Sampling Cost}} + \underbrace{\Theta\Big(ngk(1+\kappa)\,L\big((s_0+\ell)^2 d + (s_0+\ell)d^2\big)\Big)}_{\text{Training Cost}}.$$

为确定紧确的渐近界，我们找出主导项。训练成本项随优化步数 g 和反向传播因子（$1+\kappa$）而变化。此外，其自注意力成本与较长的序列长度 $s_0+\ell$ 呈二次关系。相比之下，采样成本中的注意力部分与前缀长度 $\Theta(s_0\ell)$ 仅呈线性关系；预填充成本则基于较短的前缀 s_0 计算，且不含 g 和 k 这两个乘法因子。

因此，对于任意 $g \geq 1$，训练成本项的阶数都高于预填充成本和采样成本。因此，低阶项被吸收到 Θ 记号中，得到最终的紧确界：

$$T_{\mathrm{TTRL}} = \Theta\Big(n\, g\, k\, (1+\kappa)\, L\big((s_0+\ell)^2 d + (s_0+\ell) d^2\big)\Big). \qquad \square$$

定理 C.4（TTRL 的显存复杂度）。 在序列长度为 $T = s_0 + \ell$ 的训练过程中，TTRL 步骤必须（至少）保存模型权重、梯度、优化器状态（例如 SGD 动量或 Adam 的一阶矩/二阶矩）以及反向传播所需的激活值。因此，对于有效批量大小 batch，峰值显存满足

$$\begin{aligned} M_{\mathrm{TTRL}} &= \underbrace{\Theta(|\theta|)}_{\text{weights}} + \underbrace{\Theta(|\theta|)}_{\text{gradients}} + \underbrace{\Theta(|\theta|)\text{–}\Theta(2|\theta|)}_{\text{optimizer states}} + \underbrace{\Theta\big(L\, d\, T \cdot \mathrm{batch}\big)}_{\text{backward activations}} \\ &= \Theta(|\theta|) \;+\; \Theta\big(L\, d\, (s_0+\ell) \cdot \mathrm{batch}\big). \end{aligned}$$

证明。 TTRL 训练步骤的峰值显存复杂度由两个主要部分构成：参数驻留内存和激活驻留内存。参数驻留部分包括模型权重、相应的梯度以及优化器状态（例如 Adam 的一阶矩和二阶矩）。由于这些部分均随参数数量线性增长，$|\theta|$，其总内存需求可简洁地表示为 $M_{\mathrm{param}} = \Theta(|\theta|)$。

反向传播需要激活驻留内存。对于长度为$T = s_0 + \ell$ 的训练序列，以及具有 L 层、宽度为 d的仅解码器 Transformer，标准（不使用检查点）反向传播过程必须为序列中的每个词元缓存大小为Θ（d）的激活值（例如隐藏状态和 MLP 中间结果）。将其在 L 层上汇总，并考虑有效批量大小为 batch 的情况，激活内存的规模为 $M_{\mathrm{act}} = \Theta$（$L \cdot d \cdot T \cdot$ batch）。

总峰值内存是这两个组成部分之和，即 $M_{\mathrm{TTRL}} = M_{\mathrm{param}} + M_{\mathrm{act}}$。将推导出的复杂度代入后，可得定理中给出的最终表达式：

$$M_{\mathrm{TTRL}} = \Theta(|\theta|) + \Theta\big(L\, d\, (s_0+\ell) \cdot \mathrm{batch}\big). \qquad \square$$

C.1.3 与 TTRL（Zuo 等人，2025）

命题 C.5（时间复杂度阈值）。 当上下文内优化轮数 n低于阈值 $n^\star$时，ME-ICPO 的计算速度快于 TTRL。该阈值为：

$$n^\star \;=\; \begin{cases} \frac{s_0+\ell}{\ell}\sqrt{gk(1+\kappa)} & \text{if } k \leq n \\ g(1+\kappa)\,\frac{(s_0+\ell)^2}{\ell^2} & \text{if } k > n \end{cases}$$

在轮次数量较少的实际场景中（例如 $n \leq 10$），通常能够满足这一条件，因此 ME-ICPO 是时间效率更高的方法。

证明。 阈值 $n^\star$ 通过令定理 C.1和定理 C.3中推导出的时间复杂度最高阶项相等来求得。根据 ME-ICPO 复杂度中的主导项，我们考虑两种情况。

情况 1（$k \leq n$）：ME-ICPO 的主导项是预填充成本 $T_{\mathrm{ME}} \asymp Ld\,\ell^2 n^3$。TTRL 成本为 $T_{\mathrm{TTRL}} \asymp ngk(1+\kappa)\, L(s_0+\ell)^2 d$。令二者相等并求解 n，得到：

$$Ld\,\ell^2 n^3 = ngk(1+\kappa)\, L(s_0+\ell)^2 d \quad\Longrightarrow\quad n^2 = \frac{gk(1+\kappa)(s_0+\ell)^2}{\ell^2},$$

$$n^\star = \frac{s_0+\ell}{\ell}\sqrt{gk(1+\kappa)}.$$

情况 2（$k > n$）：ME-ICPO 的主导项是解码成本 $T_{\mathrm{ME}} \asymp kLd\,\ell^2 n^2$。令其等于 TTRL 成本，得到：

$$kLd\,\ell^2 n^2 = ngk(1+\kappa)\, L(s_0+\ell)^2 d \quad\Longrightarrow\quad \ell^2 n = g(1+\kappa)(s_0+\ell)^2,$$

$$n^\star = g(1+\kappa)\frac{(s_0+\ell)^2}{\ell^2}.$$

在这两种情况下，对于 $n < n^\star$，ME-ICPO 的复杂度都更低。 □

命题C.6（内存复杂度比较）。 设 b_{ME} 为 ME-ICPO 中同时解码的候选数量， b 为 TTRL 的训练批次大小。当轮次数 n低于某个阈值时，ME-ICPO 的峰值显存低于 TTRL。一个充分条件是:

$$n \;<\; \frac{b(s_0+\ell) - b_{\mathrm{ME}}\, s_0}{b_{\mathrm{ME}}\,\ell} \;+\; \frac{\Delta_{param}\,|\theta|}{b_{\mathrm{ME}}\, c_{\mathrm{kv}}\, L\, d\, \ell},$$

其中， $\Delta_{param} > 0$ 表示 TTRL 所需的额外参数驻留内存（梯度、优化器状态）， c_{kv} 为常数。鉴于 TTRL 严格需要更多参数侧内存，且通常使用更大的批次大小 b，这一条件对所有实际取值的 n均成立。

证明。 我们通过比较定理 C.2 给出的 ME-ICPO 内存上界与定理 C.4给出的 TTRL 内存下界来确立该条件。我们要求出满足以下条件的n 其中 $M_{\mathrm{ME}} < M_{\mathrm{TTRL}}$:

$$\underbrace{c_{\mathrm{w}}^{\mathrm{ME}}\,|\theta| + c_{\mathrm{kv}}\, b_{\mathrm{ME}}\, L\, d\,(s_0+\ell n)}_{M_{\mathrm{ME}}} \;<\; \underbrace{(c_{\mathrm{w}}^{\mathrm{T}} + c_{\mathrm{g}}^{\mathrm{T}} + c_{\mathrm{opt}}^{\mathrm{T}})\,|\theta| + c_{\mathrm{kv}}\, b\, L\, d\,(s_0+\ell)}_{M_{\mathrm{TTRL}}}.$$

设 $\Delta_{param} := (c_{\mathrm{w}}^{\mathrm{T}} + c_{\mathrm{g}}^{\mathrm{T}} + c_{\mathrm{opt}}^{\mathrm{T}} - c_{\mathrm{w}}^{\mathrm{ME}}) > 0$ 为 TTRL 所需的额外参数规模张量（梯度、优化器状态）的常数因子。将不等式重新排列，解出 n:

$$c_{\mathrm{kv}}\, b_{\mathrm{ME}}\, L\, d\, \ell n < c_{\mathrm{kv}}\, b\, L\, d\,(s_0+\ell) - c_{\mathrm{kv}}\, b_{\mathrm{ME}}\, L\, d\, s_0 + \Delta_{param}\,|\theta|$$

$$n < \frac{b(s_0+\ell) - b_{\mathrm{ME}} s_0}{b_{\mathrm{ME}}\ell} + \frac{\Delta_{param}\,|\theta|}{c_{\mathrm{kv}}\, b_{\mathrm{ME}}\, L\, d\, \ell}.$$

由于 $\Delta_{param} > 0$，第二项严格为正，进一步扩大了余量。对于顺序解码（$b_{\mathrm{ME}} = 1$）和 $b \geq 1$等典型用例，第一项为较大的正值，因此对于任何实际的轮次数量 n，该条件都成立。□

C.2 实测成本

我们通过在 **AIME 2024** 上使用**Qwen2.5-Math-7B**进行受控的延迟和内存研究，补充我们的推导。我们改变轮次数 n 和每轮候选项数 k，以考察 ME-ICPO 的计算特性，并验证定理 C.1所预测的渐近趋势。我们遍历 $n \in \{1,3,5,7\}$ 和 $k \in \{4,8,16,32\}$。除非另有说明，ME-ICPO 使用我们的主要协议：温度 0。6，top-$p = 0.95$，摘要上限为 500 个词元，并采用熵前瞻，使用 m=16 短样本。所有运行均在配备 bf16 精度的 2×L40S（48GB）图形处理器上进行，候选项按顺序解码（b_{ME}=1）。

met指标。(1)**运行时间**（秒/问题），在测试集上取平均；(2) 每个问题的**令牌用量** n (ge生成的词元数，以千为单位）；(3)**峰值显存**（GB），根据 vLLM 内存统计数据得出。
结果。表8 报告了n、k）网格中各项设置下的运行时间、令牌用量和峰值显存。可以观察到两种趋势。(i) 对于固定的 k，运行时间随 n 超线性增长，其变化规律可由预填充主导的缩放关系很好地描述，且接近 $O(n^3)$，这与定理C.1一致。(ii) 当 k较大时，成本逐渐转向 $O(k\,n^2)$，因为候选项评估/选择成为瓶颈。令牌用量随 n 和 k 一同增加（多轮轨迹更长且候选更多），而在各种设置下，峰值显存始终稳定在81–82.3GB 范围内，这表明在顺序解码情况下，内存主要由基础模型驻留和 KV 缓存大小决定。

D 提示模板与定性示例

D.1 数据集特定的系统提示词

这些系统提示定义了奖励标签的语义，并且只在上下文开头放置一次。

表 8：在 **AIME 2024** 上使用 **Qwen2.5-Math-7B**时的经验成本与（n, k）对比。

n	k	运行时间（秒/问题）	词元数（千个/问题）	峰值显存（GB）
1	4	197.52	99.59	81.50
1	8	228.89	112.25	81.43
1	16	286.74	379.82	82.24
1	32	468.36	852.47	82.27
3	4	406.65	247.76	81.43
3	8	469.79	272.56	81.43
3	16	715.85	1070.98	82.18
3	32	1163.51	2111.85	82.30
5	4	609.75	372.18	81.42
5	8	713.25	506.68	81.43
5	16	1152.67	1613.49	82.24
5	32	1726.70	3412.94	82.30
7	4	820.60	559.20	81.43
7	8	1156.62	566.82	82.30
7	16	1371.78	1753.71	82.27
7	32	2643.74	5123.67	82.29

系统提示：AIME / AMC（数值答案）

你是一位人工智能数学家。你输出的所有内容都必须使用英语。以下是先前尝试中的压缩解题思路；每条思路都标记了奖励 1（正确）或奖励 0（错误）。请利用问题和这些思路推导出正确的数值答案。**完成所有推理后，在新的一行中只输出一个数字（答案），不要输出其他内容**。

你的最终输出必须采用 boxed{<number>} 的格式，其中 <number> 是最终数值答案，不得包含表达式、变量或其他文字。boxed{<number>} 中的内容必须是十进制数，不能是分数或其他形式。

系统提示：GPQA（选择题）

你是一位人工智能数学家。你输出的所有内容都必须是英文。以下是先前尝试中的压缩解题思路；每条思路均标注了奖励 1（正确）或奖励 0（错误）。请根据问题和这些思路推导出正确选项。**完成所有推理后，换到新的一行，只输出一个字母（答案），不要输出其他任何内容**。你的最终输出必须采用 boxed{<letter>} 的格式，其中 <letter> 必须是 A、B、C、D 中的一个。

系统提示：MATH（自由作答）

你是一位人工智能数学家。你输出的所有内容都必须是英文。以下是先前尝试中的压缩解题思路；每条思路均标注了奖励 1（正确）或奖励 0（错误）。请根据问题和这些思路推导出正确答案。**完成所有推理后，换到新的一行，只输出一个答案（答案），不要输出其他任何内容**。 你的最终输出必须采用 boxed{<answer>} 的格式。

它为奖励标签（$r \in \{0, 1\}$）赋予了明确的任务级含义，告诉模型从高奖励的思路中学习，并降低低奖励思路的权重——这样模型就能在测试时利用 反馈，而无需进行任何梯度更新。

D.2 摘要提示

这些提示会将完整的思维链压缩成简短摘要，同时保留高层策略。

摘要提示：美国数学邀请赛 / AMC

请简要概述下方答案中的推理。不要添加任何引言或额外说明。省略所有数值计算。摘要必须独立完整，不超过 100 个词元。

如果有最终数值结果，请在末尾按 boxed{<number>} 的格式给出（仅限小数，不得包含分数、变量或额外文字）。如果没有数值答案，则不要输出 boxed{}。[答案开始] {。 . . 原始模型答案 .. . } [答案结束]摘要:

摘要提示：GPQA

简洁总结下方答案中的推理。不要添加任何引言或额外说明。省略所有数值计算。总结必须自成一体，且不超过 100 个词元。

如果有最终答案选项，请在末尾按 boxed{<letter>} 的格式写出（必须且只能是 A、B、C、D 之一，不得添加其他文字）。如果没有最终答案，则不要输出 boxed{}。[答案开始] {。。。原始模型答案。。。 } [答案结束]总结:

摘要提示：MATH

简洁总结下方答案中的推理。不要添加任何引言或额外说明。除非对理解至关重要，否则省略所有计算。总结必须自成一体，且不超过 100 个词元。如果有最终答案，请在末尾按 boxed{<answer>} 的格式写出。[答案开始] {。。。原始模型答案。。。 } [答案结束]总结:

它们用简短的策略摘要替代冗长的思维链，只保留与决策相关的逻辑，省去算术细节，使历史记录能够放入上下文内，并在多轮次中持续发挥作用。

D.3 定性案例研究

完整提示

你是一位人工智能数学家。你输出的所有内容都必须使用英语。下面是先前尝试中压缩后的解题思路；每条思路都标注了奖励 1（正确）或奖励 0（错误）。请结合问题和这些思路，推导出正确的数值答案。完成所有推理后，在新的一行中仅输出一个数字（答案），不要输出其他内容。最终输出必须采用 boxed{<number>} 格式，其中 <number> 只能是最终数值答案（不能包含算式、变量或其他文字）。boxed{<number>} 中的内容必须是十进制数，不能是分数或其他形式。Aya 每天早上都会步行 9 千米，然后去咖啡店。她以恒定的 s 千米每小时的速度步行时，整个行程需要 4 小时，其中包括在咖啡店停留的 t 分钟。她以每小时 $s+2$ 千米的速度步行时，整个行程需要 2 小时 24 分钟，其中包括在咖啡店停留的 t 分钟。假设 Aya 以每小时 $s+\frac{1}{2}$ 千米的速度步行。求这段行程所需的时间（分钟），其中包括在咖啡店停留的 t 分钟。错误思路（奖励 0）：[0]- Aya 以每小时 $s+\frac{1}{2}$ 千米的速度步行时，包括在咖啡店停留的 t 分钟在内，总共大约需要 348 分钟。因此，答案是 348.0。正确思路（奖励 1）：[0]- 为了确定 Aya 从家步行到公园再返回所需的时间（包括在咖啡店停留的时间），我们根据已知信息列出两个方程。解方程后，得到了速度 s 和在咖啡店停留的时间 t。接着，我们计算了 Aya 以每小时 $s+\frac{1}{2}$ 千米的速度步行所需的时间，结果为 204 分钟。因此，答案是 204.0。[1]- 通过分析已知条件并进行代数变换，我们推导出 x 的值为 3 5，再将其乘以 340，得到最终答案 204。因此，答案是 204.0 - 我[2]们解方程组，求得步行速度 s 和在咖啡店停留的时间 t。然后，计算 Aya 以每小时 $s+\frac{1}{2}$ 千米的速度步行所需的总时间，包括在咖啡店停留的时间。最终答案是 204.0 分钟。因此，答案是 204.0

蓝色 标示 系统指令，这些指令用于强制执行输出模式和奖励语义。 **橙色** 标示 问题陈述。 **红色** 标示 奖励为 *0*（错误）的思路，这些思路作为反例保留。 **绿色**标示 奖励为 *1*（正确）的思路，这些是 ICPO 在选择过程中优先考虑的思路。

ICPO 在本示例中的体现。 ICPO 采样多条思维链，对其进行总结，并基于自一致性分配奖励。这些绿色项表示解题路径上的低熵一致性，从而排除固定咖啡时间 t，从而在 $s + \frac{1}{2}$ 处得到评估结果为 204.0 分钟。基于熵的筛选会降低红色离群值（≈ 348 分钟）的权重，该离群值不当缩放总时间并忽略固定偏移量。最终评估器以模式和精选思路为条件，输出正确的数值答案，204.0 。

大语言模型的使用

我们仅在论文撰写过程中将大语言模型用作写作助手，以进行轻微的语法和措辞修正。大语言模型未参与研究构思、实验设计、数据分析或结果解读。