arXiv:2602.01523v1 [cs.LG] 2 Feb 2026

大语言模型推理中基于可验证奖励的强化学习相对预算理论

Akifumi Wachi [1] **Hirota Kinoshita** [2†] **Shokichi Takakura** [1] **Rei Higuchi** [3 4] **Taiji Suzuki** [3 4]

摘要

强化学习（RL）是提升大语言模型推理能力的主流范式，但其效果会因任务和计算预算而异。我们提出一种相对预算 理论，用一个称为相对预算$\xi := H/\mathbb{E}[T]$的量来解释这种差异，其中 H 是生成视界（令牌预算）， T 表示在基础策略下首次生成正确解答所需的令牌数。我们证明， ξ 通过控制奖励方差和信息性轨迹出现的可能性，决定了样本效率。我们的分析揭示了三种状态：在不足 状态（$\xi \to 0$）下，信息性轨迹十分罕见，样本复杂度急剧上升；在平衡状态（$\xi = \Theta(1)$）下，信息性轨迹以不可忽略的概率出现，强化学习的样本效率达到最高；在充足状态（$\xi \to \infty$）下，学习仍然稳定，但每次迭代带来的边际收益会逐渐减小。我们还为在线强化学习提供了有限样本保证，用以刻画不同状态下的学习进展。具体而言，在一个采用理想化分布假设的案例研究中，我们证明相对预算会随迭代次数线性增长。我们的实证结果在现实设置中验证了这些预测，并确定了一个能够最大限度提升学习效率、且与推理性能峰值相吻合的预算 $\xi \in [1.5, 2.0]$ 。

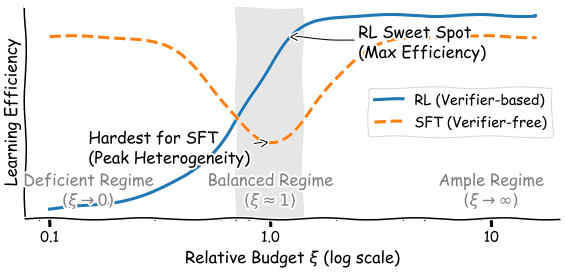


图 *1*。相对预算理论的概念示意图。均衡相对预算状态（$\xi \approx 1$）充当相变点。在这一状态下，强化学习的样本效率达到最大，而监督微调则会出现性能退化。

1. 引言

具有可验证奖励的强化学习（可验证奖励强化学习）已成为提升大语言模型（大语言模型）推理能力的强大范式。当奖励可以自动验证时，可验证奖励强化学习支持长时域规划与迭代式自我修正，例如 OpenAI o 系列（Jaech 等人，2024年）、DeepSeek-R1（Guo 等人，2025年）以及 Kimi K1.5（Kimi 团队等人，2025年）。值得注意的是，Jaech 等人（2024年）报告称，持续强化学习训练与增加测试时计算所带来的收益展现出不同于预训练的缩放行为（Kaplan 等人，2020年）。

尽管实证研究进展迅速，但在计算约束下对可验证奖励强化学习的理论理解仍不充分。现有研究主要关注算法的实证发展（Yu等人，2025年）或数据方案（Ye等人，2024年；Zhou等人，2025年），而对扩展保证的关注较少。一个值得注意的例外是Setlur等人（2025年），他们提供了奠基性分析，表明基于验证器的方法（例如强化学习）在测试时计算扩展方面优于无验证器方法（例如监督微调）。然而，该框架既未考虑任务难度与计算预算之间的关系，也未考虑在线学习的动态变化。缺乏此类指导，实践者可能会将大量计算资源投入到理论上学习效率低下的领域。

这种关系决定了可验证奖励强化学习的有效性。从经验上看，性能增益在“能力边缘”（Zhang 等人，2025年）处达到峰值，此时任务对于基础模型而言既不简单，也并非难以解决。至关重要的是，测试时计算扩展（Snell 等人，2024年）使难度变得动态化，因为任务的可解决性会随预算的增加而提升。这一转变要求建立一个统一能力与计算约束的理论框架，以解释强化学习在不同机制下的成功。

[†]部分工作是在 LY Corporation 实习期间完成的。[1]LY Corporation [2]芝加哥丰田技术研究所 [3]东京大学 [4]理化学研究所人工智能项目。通信作者：Akifumi Wachi <akifumi.wachi@lycorp.co.jp>。

预印本。*2026*年*2*月*3*日。

我们的贡献在于通过相对预算 $\xi := H/\mathbb{E}[T]$分析可验证奖励强化学习的动态，其中 H 是令牌预算，T 表示直到首次得到正确解之前的令牌数。利用 ξ 将计算预算与任务难度联系起来，我们提出以下理论结果：

三种状态下的强化学习收敛性（定理 5.2）。如图 1所示，我们分析了由 ξ：所决定的三种状态下的次优性和样本复杂度

- **不足 ($\xi \to 0$)**：发现高奖励轨迹的概率趋近于零，导致样本复杂度发散。
- **平衡** （$\xi = \Theta(1)$）：这是一个临界分歧点，在此处，由于高奖励轨迹的概率与 0保持一定距离，强化学习实现了最高效率。相反，监督微调由于解的异质性达到峰值而遭受最严重的次优性（第5.3节）。
- **充足($\xi \to \infty$)**：高奖励轨迹的概率始终与零保持一定距离，从而确保高效学习，尽管每次迭代的收益会逐渐下降。

从根本上说，我们的核心理论仅依赖于一般性假设，从而确保其对特定分布具有稳健性。

在线强化学习的有限样本保证 (定理 6.2至 6.4）。将分析扩展到在线设置后，我们推导出实现单调策略改进所需的样本复杂度，并确定平衡区间是样本效率最高的区域。此外，通过伽马分布对推理时间建模，我们证明相对预算 ξ 会随迭代线性增长，因此需要越来越多的轨迹采样。这保证了策略能够单调降低其预期求解时间，从而有效地将任务推向充足区间。

实证验证。 我们通过在推理基准上使用标准大语言模型开展实验，验证了理论发现。我们观察到，预测的相变出现在 $\xi \approx 1$ 附近。此外，学习信号在 $\xi \in [1.5, 2.0]$ 处达到最大值，这与最佳后训练性能相符。这证实了我们的理论能够有效解释现实任务中的最优计算资源分配。

2. 相关工作

推理能力的最新进展使研究重点从单纯扩大模型参数规模，转向优化训练数据分布和测试时计算资源分配（Alomraniet al.，2025年）。这些实证方法与我们对相对预算 ξ的理论分析相一致。

测试时扩展与数据效率。 推理计算量与性能之间的权衡，已被正式表述为测试时扩展定律（Snell 等人、2024年；Welleck 等人，2024年）。例如 *budget forcing* (Muennighoff 等人，2025年) 通过追加“Wait”令牌来调整令牌预算H。在我们的框架中，这会增加相对预算 ξ，从而有效地将问题从 不足 状态（$\xi \ll 1$）——在此状态下不存在学习信号——转移到 平衡 状态（$\xi \approx 1$），在此状态下推理变得可行。另一种做法是，Ye 等人(2025年) 和Muennighoff 等人 (2025年) 关注任务分布本身。他们表明，如果移除简单或不可解的案例，即便只有少量训练样本（例如 ≈1k 个样本），推理能力也会显现。这些研究表明，相对难度比原始规模更能推动数据效率。

强化学习中的难度筛选。在可验证奖励强化学习中（RLVR），使任务难度与模型能力相匹配至关重要（Albalak 等人，2024年；Sun 等人，2025年）。Bae 等人（2025年）通过在线难度筛选解决了这一问题，并通过实证表明，当动态选取成功率为0。5的样本时，训练效果最佳。他们根据难度将提示分为三个区间，并指出，最大化成功率的方差能够带来最佳学习效率。这一实证结果与我们的定理5.2 和6.3密切相关；定理表明，平衡区间（$\xi \approx 1$）是样本效率的最佳区间。类似地，Cui 等人（2025年）提出在线更新过程奖励模型，利用结果标签，并采用筛选策略，仅保留模型准确率处于 $[0.2, 0.8]$范围内的提示。Bae 等人（2025年）和Cui 等人（2025年）均采用动态筛选来应对“移动目标”问题：随着模型能力提升，固定任务会从平衡区间转移到充足区间。

3. 预备知识与问题陈述

3.1. 大语言模型推理作为马尔可夫决策过程

给定提示（或问题） $x \in \mathcal{X}$，大语言模型逐个词元生成答案。遵循 Ke 等人（2025年）和 Kazemnejad 等人（2025年）的做法，我们将这一过程建模为一个有限时域马尔可夫决策过程（MDP），定义为 $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, F, r, H, \rho \rangle$，其中 $\mathcal{S}$是状态空间， $\mathcal{A}$是词元空间， $F: \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ 是确定性状态转移， $r: \mathcal{S} \times \mathcal{A} \to \{0, 1\}$ 是类别 $\mathcal{R}$中的二元奖励函数， $H \in \mathbb{Z}_+$ 是时域（即词元预算）， $\rho \in \Delta(\mathcal{X})$是初始提示分布，其中$\Delta(\cdot)$是概率单纯形。为简单起见，我们假设奖励反馈无噪声。由于解码是自回归的，每个词元对应于从策略π：$\mathcal{S} \to \Delta(\mathcal{A})$中抽取的一个动作。具体而言，对于所有 $h \in \{1, 2, \ldots, H\}$，令 $s_h = (x, a_1, \ldots, a_{h-1})$，并抽取 $a_h \sim \pi(\cdot \mid s_h)$。因此， $\mathcal{S} := \{(x, a_1, \ldots, a_{h-1}) : x \in \mathcal{X}, a_1, a_2, \ldots, a_{h-1} \in \mathcal{A}, h \in [H]\}$，其中 $s_1 = x$。状态转移 F 是确定性的；也就是说，生成词元 a_h后，它只需通过拼接 $s_{h+1} = (s_h, a_h)$，将新选取的动作词元附加到当前状态，并获取奖励 $r_h := r(s_h, a_h)$。解答轨迹是一个展开 $\tau = (x, a_1, \ldots, a_H)$，位于

MDP 中，其在 τ 上的相应累积奖励为 $R(\tau) = \sum_{h=1}^{H} r(s_h, a_h)$。

遵循Setlur等人（2025年）的做法，我们采用带有步级塑形的二元奖励，以编码对尽早找到解的偏好。令 $h^\star$ 为轨迹找到正确解的第一步。如果 $h^\star \le H$，则设定 $r_h = 1$，适用于 $h \ge h^\star$；并设定 $r_h = 0$，适用于 $h < h^\star$，这样越早找到解，回报就越高，这相当于施加长度惩罚（Alomrani等人，2025年；Arora与Zanette，2025年；Liu等人，2025年）。一个自然的替代方案是，一旦得到正确解就立即终止，并赋予固定的终止奖励。这个目标函数只取决于成功概率，不区分早找到解还是晚找到解；而我们的设定则明确重视在令牌预算内尽早找到解。我们将轨迹 τ 定义为：在 H 个令牌内得到目标答案即为正确。总之，我们寻求一个策略 π，使其最大化

$$J(\pi) := \mathbb{E}_{\rho,\pi}\left[R(\tau)\right] = \mathbb{E}_{\rho,\pi}\left[\sum_{h=1}^{H} r(s_h, a_h)\right], \quad (1)$$

其中 $\mathbb{E}_{\rho,\,\pi}[\cdot]$ 是 $\mathbb{E}_{x\sim\rho}[\mathbb{E}_{\tau\sim\pi(\cdot|x)}[\cdot]]$ 的缩写。我们用 $\tau \sim \pi(\cdot \mid x)$ 表示从 x 开始、由 π 诱导的分布中采样得到的 τ 轨迹。

记号。全文中，我们使用标准渐近记号：$\mathcal{O}(\cdot)$ 表示上界，$\Omega(\cdot)$ 表示下界，而 $\Theta(\cdot)$ 表示两者均成立。我们分别用 $a = \tilde{\mathcal{O}}(b)$ 和 $a = \tilde{\Omega}(b)$ 表示 $a = \mathcal{O}(b \cdot \max(1, \mathrm{polylog}(b)))$ 和 $a = \Omega(b \cdot \max(1, \mathrm{polylog}(b)))$。此外，我们令 $a \lesssim b$ 表示 $a = \mathcal{O}(b)$。对于正整数 $n \in \mathbb{Z}_+$，$[n]$ 表示集合 $\{1, 2, \ldots, n\}$。最后，$[\mathbf{x}]_{\mathbf{y}}^{\mathbf{z}} := \min\{\mathbf{z}, \max\{\mathbf{x}, \mathbf{y}\}\}$ 表示截断。

3.2. 强化学习中的次优差距

Setlur等人(2025年) 从理论上证明，基于验证器的强化学习方法在扩展测试时计算方面优于无验证器基线（例如监督微调）。监督微调受解的异质性限制，而强化学习能够利用基础策略采样出的罕见高奖励轨迹 $\pi_b : \mathcal{S} \to \Delta(\mathcal{A})$。受这一理论优势的启发，我们聚焦于强化学习设置。[1]

遵循 Setlur 等人（2025年）的做法，我们将其与 π_b 周围 χ^2 信赖域内的最佳策略进行比较。对于任意实例 $x \in \mathcal{X}$，定义实例级 χ^2-散度。给定半径 $\kappa > 0$，令 $\Pi_\kappa = \{\pi \mid \mathbb{E}_{x\sim\rho}\left[\mathbb{D}_{\chi^2}\left(\pi(\cdot \mid x) \,\|\, \pi_b(\cdot \mid x)\right)\right] \le \kappa\}$，以及 $\bar{\pi}_\kappa \in \arg\max_{\pi\in\Pi_\kappa} J(\pi)$。尽管 Π_κ 意味着对 $x \sim \rho$ 的均值约束，但该比较策略会诱导出一个实例级散度水平

$$\kappa_x := \mathbb{D}_{\chi^2}\big(\bar{\pi}_\kappa(\cdot \mid x) \,\|\, \pi_b(\cdot \mid x)\big), \qquad \forall x \in \mathcal{X}. \quad (2)$$

[1]我们还将在第 5.3 节讨论与监督微调的联系。

定义 $\sigma_{b,x}^2 := \mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)]$。那么，对于实例 x，半径为 κ_x 的 χ^2 会将预期改进限制为 $\mathcal{O}(\sigma_{b,x}\sqrt{\kappa_x})$。这促使我们测量事件发生的概率，即这些事件偏离均值 $\sqrt{\kappa_x}$ 个标准差。

定义 3.1 (反集中)。令 π_b 为基础策略。对于每个 $x \in \mathcal{X}$ 和标量 $\varepsilon > 0$，定义

$$c_x(\varepsilon) := \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\Big[R(\tau) \ge \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] + \sigma_{b,x}\sqrt{\varepsilon}\Big].$$

那么，给定半径 χ^2，π_b 的反集中系数定义为 $c_0(\kappa) := \min_{x\in\mathcal{X}} c_x(\kappa_x)$。

反集中的重要性可通过以下关于次优差距的界限形式化。

引理 3.2 （强化学习的次优差距，Setlur等人（2025年））。以至少 $1-\delta$ 的概率，策略 $\hat{\pi}_n^{\mathrm{RL}}$ 由*Setlur等人（2025年）*提出的强化学习算法返回，其相对于最佳比较策略 $\bar{\pi}_\kappa$ 的次优差距满足：

$$J(\bar{\pi}_\kappa) - J(\hat{\pi}_n^{\mathrm{RL}}) \lesssim \frac{H\log(|\mathcal{R}|/\delta)}{c_0(\kappa)\, n}. \quad (3)$$

其中，$c_0(\kappa)$ 是基础策略 π_b 在信赖域半径为 $\kappa > 0$ 时的反集中系数，且满足

$$\sqrt{\kappa_x} \cdot \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] \le \sigma_{b,x} \qquad \forall x \in \mathcal{X}. \quad (4)$$

直观而言，反集中系数越大，在采样过程中发现高奖励解答轨迹的概率就越高，因此从理论上讲，强化学习智能体所需的样本复杂度会降低。

4. 相对预算框架

4.1. 正确解所需时间建模

我们对基础策略 π_b下首次得到正确解之前生成的 token 数量进行建模，并将其分布特性与基于验证器的强化学习中的奖励方差和样本复杂度联系起来。给定验证器 $W_x : \mathcal{S} \to \{0,1\}$和解答轨迹 τ，我们定义 $T(\tau) := \min\{h \in \mathbb{Z}_+ : W_x(s_h) = 1\}$，其中，若未解出，则 $T(\tau) = \infty$。直观而言，对于 π_b，问题越难，得到首个正确解之前的预期 token 数量就越多。我们将 $T(\tau)$ 的前两个矩记为

$$\mu_x := \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)] \quad \text{and} \quad v_x := \mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)].$$

问题 x 在 π_b 下的成功概率为

$$q_x := \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\big[T(\tau) \le H\big]. \quad (5)$$

最后，根据双层奖励结构，累积奖励 $R(\tau)$ 仅由 $T(\tau)$ 决定，如下所示：

$$R(\tau) = \begin{cases} H - T(\tau) + 1 & \text{if} \quad T(\tau) \le H, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

4.2. 定义相对预算与假设

为分析计算预算与问题难度之间的相互作用，我们引入相对预算。该参数表示对于基础策略 π_b，预算 H 相对于特定问题难度 x 的充足程度。

定义 4.1（相对预算）。 对于给定问题 x，我们将计算预算 $H \in \mathbb{Z}_+$ 除以其平均求解时间 μ_x 进行归一化，并定义相对预算（即相对于任务难度的计算预算）为

$$\xi_x := \frac{H}{\mu_x} = \frac{H}{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]}. \tag{7}$$

为确保数学上的可处理性，我们将分析限制在可解问题子集 $\mathcal{X}_\sharp \subseteq \mathcal{X}$中，其中 π_b具有有限的预期求解时间（即 $\mu_x < \infty$对所有 $x \in \mathcal{X}_\sharp$均成立）。这一限制有实证依据，因为近期研究（Kim 等人，2025b；Yue 等人，2025年；Wu 等人，2025a）表明，RLVR 主要优化可解任务的效率，而非为根本上不可解的问题赋予新能力。因此，我们将本质上不可解的实例视为研究范围之外。我们假设 $T(\tau)$ 在 $\mathcal{X}_\sharp$ 上的矩有限，如下所示。

假设 4.2 （T（τ）的矩）。 我们假设 $v_x = \Theta(\mu_x^2)$，且二阶矩 $\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2]$ 对所有 $x \in \mathcal{X}_\sharp$均有限。

直观而言，该假设排除了重尾行为，确保求解时间的标准差与其均值成比例。一个具有代表性的充分条件是具有一致的次 Weibull 尾部（Vladimirova 等人，2020）。通过聚焦于 $\mathcal{X}_\sharp$，我们关注的是预算分配的效率，而非任务可行性。

为了分析发现高奖励解轨迹的概率，我们刻画 T（τ）的左尾。

假设 4.3 （T（τ）的左尾）。 存在常数$z_0 \in (0, 1]$、c_-，$c_+ > 0$，以及一个非递减函数f：$[0, z_0] \to [0, 1]$，满足 $f(0) = 0$且 $f(z_1) > 0$，其中某个$z_1 \in (0, z_0]$ 使得对每个问题 $x \in \mathcal{X}_\sharp$ 以及每个满足 z_x 的$t \le \mu_x := t/\mu_x \in [0, z_0]$，

$$c_-\, f(z_x) \;\le\; \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) \le t] \;\le\; c_+\, f(z_x). \tag{8}$$

此外，f 在原点附近满足一个温和的倍增条件。具体而言，存在 $D \ge 1$，使得对所有 $z \in (0, z_0]$，$D^{-1} f(z) \le f(z/2) \le D f(z)$。

这里，$f(z)$刻画了在预期时间的 z 分之一内解决问题的稀有程度。倍增条件确保 f 在原点附近按多项式而非指数速度衰减，从而排除早期成功实际上不可能的分布。

接下来，我们刻画平衡情形，即计算预算与预期求解时间相当的情况。

假设4.4 （ξ平衡时的非退化性）。 固定常数 $0 < \xi_{\min} < \xi_{\max} < \infty$，并定义满足 $\xi_x \in [\xi_{\min}, \xi_{\max}]$ 的实例集合

$$\mathcal{X}_b := \Big\{x \in \mathcal{X} : \xi_x = H/\mu_x \in [\xi_{\min}, \xi_{\max}]\Big\}. \tag{9}$$

对于每个 $x \in \mathcal{X}_b$，我们假设存在普适常数$q_{\min} \in (0, 1/2)$和 $c_v \in (0, 1]$，使得 $q_{\min} \le q_x \le 1 - q_{\min}$，且条件方差满足

$$\mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) \mid T(\tau) \le H] \ge c_v v_x. \tag{10}$$

这一假设认为，在平衡区间内，在 H 处截断不会使推理时间分布的方差骤降；也就是说，即使以成功为条件，难度分布仍然是非退化的。

5. 强化学习在不同相对预算区间下的次优差距

5.1. 依赖区间的反集中界

由于强化学习从成功和失败的解答轨迹中学习，因此累积奖励 $R(\tau)$ 的期望和方差必须针对整个轨迹分布（包括失败情形 $T(\tau) > H$和 $R(\tau) = 0$）来计算。

现在，我们通过对正确解的求解时间建模（如第 4 节所述），分析反集中系数 c_0 随相对预算 $\xi_x = H/\mu_x$ 变化的规律。通过成功概率 q_x分析奖励统计量，我们将 $c_x(\varepsilon)$ 与 $T(\tau)$ 的左尾联系起来。

引理 5.1 (以 $T(\tau)$表示的反集中)。 对于每个 $\varepsilon > 0$以及满足 $q_x > 0$ 的实例 x，存在一个标量系数 $\eta(x, \varepsilon)$ 使得

$$c_x(\varepsilon) = \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau) \le \eta(x,\varepsilon) \cdot \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]\right].$$

此外，对于某个与 $c_- > 0$ 无关的常数 x，有$c_x(\varepsilon) \ge c_- f(\bar\eta(x, \varepsilon))$，其中 $\bar\eta(x, \varepsilon) := [\eta(x, \varepsilon)]_0^{z_0}$。注意，$z_0$ 是定理 *4.3*中引入的常数。

完整证明见第D.2节。在我们的双层设定中，较高的奖励对应较早得到的解。因此，奖励超过其均值σ_b的$x\sqrt{\varepsilon}$事件，–等价于在其平均时间 μ_x的比例$\eta(x, \varepsilon)$ 内求解该实例 μ_x。定理 4.3 进而将该概率下界为 $c_- f(\bar\eta(x, \varepsilon))$。

接下来，我们分析 σ_b、x 和 c_0 在相对预算 ξ_x的三个区间中的缩放情况。对于 $\xi > 0$，我们定义以下 $\mathcal{X}$的子集，记作：

$$\mathcal{X}(\xi) := \{x \mid \xi/2 \le \xi_x \le 2\xi\}, \tag{11}$$

以及相应的反集中系数

$$c_0(\xi; \varepsilon) := \inf_{x\in\mathcal{X}(\xi)} c_x(\varepsilon). \tag{12}$$

据此，我们将以相对预算切片 $\mathcal{X}(\xi)$ 为条件的目标函数定义为：

$$J_\xi(\pi) := \mathbb{E}_{x\sim\rho(\cdot|x\in\mathcal{X}(\xi)),\tau\sim\pi(\cdot|x)}[R(\tau)]. \tag{13}$$

定理 3.2 表明，强化学习的样本效率由反集中系数 c_0 所决定。为理解计算约束的影响，我们现在分析 c_0 对相对预算 ξ 的依赖关系。以下定理刻画了$c_0(\xi)$ 在三个不同区间内的渐近行为，从而确定次优差距的尺度。

定理 *5.2*（区间依赖的反集中及其对强化学习的一个推论）。固定一个相对预算切片 $\mathcal{X}(\xi) := \{x : \xi/2 \le \xi_x \le 2\xi\}$。令 $\hat{\pi}_n^{\mathrm{RL}}$ 为定理 *3.2* 中强化学习过程返回的策略。给定 χ^2 信赖域半径 $\kappa > 0$，定义 $c_0(\xi;\ \kappa) := \inf_{x\in\mathcal{X}(\xi)} c_x(\kappa)$。若信赖域条件（*4*）得到满足，且比较策略 $\bar{\pi}_\kappa$ 对所有 $x \in \mathcal{X}(\xi)$ 均满足 $\kappa_x \le \kappa$，则以下界以至少 $1-\delta$ 的概率成立：

$$J_\xi(\bar{\pi}_\kappa) - J_\xi(\hat{\pi}_n^{\mathrm{RL}}) \lesssim \frac{H\log(|\mathcal{R}|/\delta)}{c_0(\xi;\kappa)\,n}. \tag{14}$$

在假设 *4.2*与 *4.3*下，$c_0(\xi;\kappa)$ 的区间依赖行为为：

1.（不足）$c_0(\xi;\ \kappa) = \Theta(f(\xi))$，当 $\xi \to 0$。*2.*（平衡）$c_0(\xi;\kappa) = \Theta(1)$，对 $\xi \in [\xi_{\min}, \xi_{\max}]$，若定理 *4.4* 额外成立，且 κ 的选取使得$\kappa \le (q_{\min}/2)^2$。*3.*（充足）$c_0(\xi;\kappa) = \Theta(1)$，当 $\xi \to \infty$，对所有 $\kappa \le \bar{\kappa}$，对某个常数 $\bar{\kappa} \in (0,1)$。

证明见第D.3节。该定理将强化学习性能与相对预算 ξ 联系起来，为推理效率提供了统一视角。关键在于，在平衡到充足区间中，当计算预算 H 与预期求解时间 μ_x 相当或超过该值时，反集中系数始终严格远离 0；也就是说，$c_0(\xi;\kappa) = \Theta(1)$。这确保样本复杂度保持可处理，且误差随时域 H线性增长。相比之下，在相对预算不足区间中，学习效率会依据解的尾概率 $f(\xi)$ 而下降，从而量化了从稀有成功轨迹中学习的难度。

我们的理论分析还揭示了一个根本局限：由于反集中现象消失，在不足区间（$\xi \ll 1$）应用强化学习在计算上难以承受。从我们的框架来看，初步监督微调或蒸馏等标准做法，可以理解为通过降低有效任务难度 μ_x，将任务转移到平衡至充足区间（$\xi \gtrsim 1$）的机制，从而与增加计算预算 H相辅相成。

5.2. 伽马分布的具体分析

为提供具体估计，我们采用伽马分布来建模 T（τ）。这一选择符合我们的假设，且便于进行数学处理。尽管推理步骤是离散的，但复杂任务中的令牌数量较大，因此采用连续近似是合理的；事实上，当单步成功概率较低时，离散的负二项分布会在极限情况下收敛于伽马分布。

我们根据所需洞见数量 $K \in \mathbb{Z}_+$ 和每一步的成功概率 $p \in [0,1]$来刻画问题。具体而言，我们假设 $T(\tau) \mid x \sim$服从形状参数为 K、率参数为 p 的伽马分布 Gamma(K,p)，从而得到期望时间 $\mathbb{E}[T(\tau)] = K/p$。为便于分析，我们采用代理奖励$R(\tau) := \max\{0, H - T(\tau)\}$；在成功轨迹上，它与原始目标函数的差至多为一个加性常数。该模型推广了 Kim 等人（2025a）的理论框架，并与我们在附录 A 中的实证观察相一致。在此模型下，我们定量刻画奖励方差和反集中性质。

命题 5.3 (伽马模型下的统计量)。令 $\gamma(\cdot,\cdot)$表示正则化下不完全伽马函数。尽管精确表达式涉及不完全伽马函数，并推迟至附录 *D.7*和 *D.8*中给出，我们仍找到了刻画其缩放关系的闭式系数 $C_{\mathrm{RL}}(K,\xi)$ 和 $\psi(K,\xi,\varepsilon)$。具体而言，标准差 σ_b 和反集中系数 c满足：

$$\sigma_b(K,\xi) = C_{\mathrm{RL}}(K,\xi)H\cdot(K\xi)^{-1}, \tag{15}$$

$$c(K,\xi,\varepsilon) = \gamma\big(K, K\xi\cdot[\psi(K,\xi,\varepsilon)]_0^1\big). \tag{16}$$

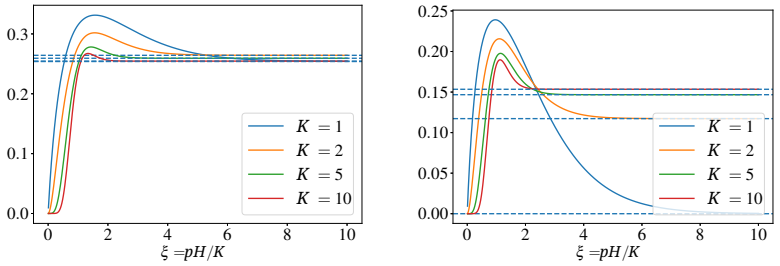
证明见第 D.7 节和 D.8 节。值得注意的是，这些显式形式使我们能够分析渐近行为和最优预算分配。

注 5.4（渐近性质）。当 $\xi \to \infty, C_{\mathrm{RL}}(K,\xi) \to \sqrt{K}$成立时，反集中系数收敛至 $\lim_{\xi\to\infty} c(K,\xi,\varepsilon) = \gamma(K, K-\sqrt{\varepsilon K})$。对于固定的 ε，除非 $\varepsilon \ge K$，该系数不会随着 ξ 增大而趋于零，从而确保在充足预算区间内能够稳定学习。

图2 展示了相对预算 $\xi = pH/K$ 与系数 c_0之间的关系。与定理5.4一致，c_0 在$\xi = \Omega(1)$区间内始终严格为正。此外，对 $\sigma_b(K,\xi)$进行数值分析发现存在唯一的全局最大值点 $\xi^\star(K)$。如图 3所示，对于各种 K，最优相对预算位于范围$1.0 \le \xi^\star(K) \le 1.4$。这意味着，当计算预算$H$ 设为略大于正确解的预期求解时间时，次优差距最小。更多详情请参见第 D.9 节。

5.3. 理论比较：强化学习与监督微调

将我们的发现与无验证器监督微调进行对比，揭示了一项关键权衡。正如 Setluret al. (2025年)所确立的，



(a) $\varepsilon = 0.5$。 *(b)* $\varepsilon = 1.0$。

图 *2.* 不同的 K 和 ε下， ξ 与 c_0之间的关系（K， ξ； ε）。

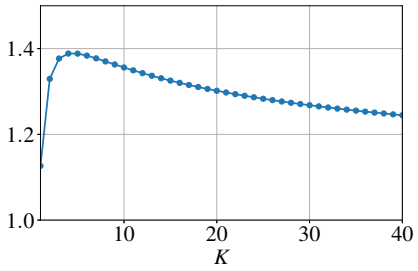


图 *3.* 最优相对预算 $\xi^\star$（K）随 K的变化。

监督微调的次优差距下界由成功条件下的策略异质性决定，记为 $\sigma_b^{\mathrm{SFT}}(\xi)$。形式上，其次优差距按 $\Omega(\sigma_b^{\mathrm{SFT}}(\xi)/\sqrt{\bar{n}})$ 的规模增长，其中 $\bar{n}$ 为成功解轨迹的数量。这意味着，当成功示范具有多样性且缺乏连贯策略时，学习会受到阻碍。

引入相对预算框架后，可以看出平衡区间（$\xi \approx 1$）存在鲜明对比。强化学习凭借稳定的反集中特性实现了最高的样本效率，而监督微调则面临最坏情形。第 F 节的分析表明， $\sigma_b^{\mathrm{SFT}}(\xi)$ 在 $\xi = \Theta(1)$处达到最大值。与解方差较低的不足区间或充足区间不同，平衡区间产生了最为异质的混合分布：既有“幸运”的短解，也有勉强符合预算的解。因此，基于验证器的强化学习的“最佳区间”，同时也是监督微调面临的最严峻理论挑战。

6. 在线强化学习的动态

现在，我们将分析从静态设置扩展到学习动态。我们考虑一种迭代式同策略设置：在每次迭代 i中，智能体使用当前策略 $\pi^{(i)}$ 收集一个新数据集，并将其更新为 $\pi^{(i+1)}$。该协议会丢弃历史样本，使我们能够严格追踪相对预算 ξ_i 的演变，并量化在不同区间实现策略单调改进所需的样本复杂度。

6.1. 单调改进的样本复杂度

令 $m \in \mathbb{Z}_+$ 为迭代总数。我们以基础策略 $\pi^{(0)} := \pi_b$ 初始化该过程，并考察迭代 $i = \{0, 1, \ldots, m-1\}$ 中的学习动态。对于任意实例 x，我们将当前策略 $\pi^{(i)}$ 下的预期求解时间和实例特定相对预算分别记为 $\mu_x^{(i)} := \mathbb{E}_{\tau\sim\pi^{(i)}(\cdot|x)}[T(\tau)]$ 和 $\xi_x^{(i)} := H/\mu_x^{(i)}$。固定目标相对预算 $\xi > 0$，并定义相应的输入切片 $\mathcal{X}(\xi) := \{x : \xi/2 \le \xi_x^{(i)} \le 2\xi\}$，以及以该切片为条件的输入分布 $\rho_\xi(x) := \rho(x \mid x \in \mathcal{X}(\xi))$。在以下分析中，除非另有说明，期望均针对 $m \in \mathbb{Z}_+$ 和 $\pi^{(0)} (= \pi_b)$ 计算。因此，我们将该切片的代表性相对预算定义为

除非另有说明，期望均针对 $x \sim \rho_\xi$ 和 $\tau \sim \pi^{(i)}(\cdot \mid x)$ 计算。因此，我们将该切片的代表性相对预算定义为

$$\xi_i := H/\mathbb{E}_{x\sim\rho_\xi}[\mu_x^{(i)}]. \tag{17}$$

为确保整体分析的有效性，我们针对同一相对预算切片中的问题引入齐性假设。这意味着奖励方差主要由相对预算区间决定，并且在该切片内的不同输入之间近似恒定。**假设 6.1**（相对预算切片内的齐性）。对于每次迭代 i，我们假设$\mathcal{X}(\xi)$内的奖励具有齐性，这意味着对于所有 $x \in \mathcal{X}(\xi)$，条件方差可由切片级统计量近似表示。

据此，我们将当前策略 $\pi^{(i)}$获得的累积奖励的代表性标准差定义为条件标准差的均值： $\sigma(\pi^{(i)}) := \mathbb{E}_{x\sim\rho_\xi}\left[\sqrt{\mathbb{V}_{\tau\sim\pi^{(i)}(\cdot|x)}[R(\tau)]}\right]$.

基于这一形式化框架，我们现在为单次迭代建立有限样本保证。以下定理刻画了利用从 $\pi^{(i)}$抽取的 n_i 条采样轨迹，将 $\pi^{(i)}$更新为 $\hat{}\pi^{(i+1)}$ 所带来的策略改进。

定理 6.2（强化学习带来的一步改进）。 固定一个迭代索引 i。令 $\pi^{(i)}$ 表示当前具有相对预算的基础策略 ξ_i，并令 κ_i 为信赖域半径（参见定理*3.1*）。在假设*6.1*下，令 $\hat{}\pi^{(i+1)}$ 表示采用信赖域半径 κ_i，并使用 n_i 个来自 $\pi^{(i)}$的采样轨迹，根据定理 *3.2*中规定的基于验证器的强化学习更新所得到的策略。则以至少 $1-\delta$的概率，

$$J_{\xi_i}\left(\hat{\pi}^{(i+1)}\right) \ge J_{\xi_i}\left(\pi^{(i)}\right) + \sqrt{\kappa_i}\,\sigma(\pi^{(i)}) - \frac{CH\log(|\mathcal{R}|/\delta)}{c_0(\xi_i;\kappa_i)\,n_i}.$$

此外，如果样本量 n_i 足够大，使得

$$n_i \ge \frac{2CH\log(|\mathcal{R}|/\delta)}{\sqrt{\kappa_i}\,c_0(\xi_i;\kappa_i)\,\sigma(\pi^{(i)})}, \tag{18}$$

然后将下一个基础策略设置为 $\pi^{(i+1)} := \hat{}\pi^{(i+1)}$，得到

$$J_{\xi_i}\left(\pi^{(i+1)}\right) \ge J_{\xi_i}\left(\pi^{(i)}\right) + \frac{1}{2}\sqrt{\kappa_i}\,\sigma(\pi^{(i)}). \tag{19}$$

这里，C 是引理 3.2所蕴含的一个普适常数，而$c_0(\xi_i;\ \kappa_i)$是 $\pi^{(i)}$ 的反集中系数。

该证明依赖于将性能提升分解为两个相互竞争的项：理论提升分量和随机估计误差。该定理通过确定样本量 n_i来建立，该样本量足以使提升信号超过误差，从而保证目标函数单调增加。完整证明详见第E.1节。

6.2. 相对预算在学习动态中的作用

将定理 6.2中的有限样本保证应用于第 5节定义的区间，我们推导出实现单调策略改进所需的最小样本复杂度 n_i。具体而言，代入奖励标准差 σ 和反集中 $c_0(\xi;\ \kappa_i)$ 的标度律后，可得到以下依赖于区间的复杂度界。

定理 6.3（在线强化学习中的三个相对预算区间）。考虑一个满足定理 6.2所述单步改进保证的在线强化学习算法。则可实现的每次迭代改进和 n_i 单调策略改进所需的充分样本复杂度呈现出三个不同的相对预算区间：

1.（不足：$\xi_i \ll 1$）改进幅度被限制为$\mathcal{O}(H\sqrt{\kappa_i} f(\xi_i))$，样本复杂度则趋于无穷：

$$n_i = \widetilde{\Omega}\left(\kappa_i^{-1/2}(f(\xi_i))^{-3/2}\right). \tag{20}$$

2.（平衡：$\xi_i = \Theta(1)$）改进幅度为 $\mathcal{O}(\sqrt{\kappa_i}H)$，充分的样本量为

$$n_i = \widetilde{\Omega}\left(1/\sqrt{\kappa_i}\right). \tag{21}$$

3.（充足：$\xi_i \gg 1$）改进幅度衰减至$\mathcal{O}(\sqrt{\kappa_i}H/\xi_i)$，样本复杂度按如下规律增长

$$n_i = \widetilde{\Omega}\left(\xi_i/\sqrt{\kappa_i}\right). \tag{22}$$

平衡区间内的学习最具样本效率；而相对预算不足或过高都会降低改进效率，原因分别是学习信号趋于消失或样本成本增加。请注意，在充足区间内，为确保改进单调进行，通常需要将信赖域按 $\sqrt{\kappa_i} = \mathcal{O}(\xi_i^{-1})$ 进行缩放，这会导致有效改进率为$\widetilde{\mathcal{O}}(H/\xi_i^2)$，且样本复杂度出现二次退化：$n_i = \widetilde{\Omega}(\xi_i^2)$。

6.3. 伽马模型下的线性预算增长

最后，我们采用与第5.2 节相同的假设，即策略的求解时间服从伽马分布。定理6.4 在该模型下保证，相对预算 ξ_i 随迭代线性增长，而期望回报 $J(\pi^{(i)})$ 渐近收敛至令牌预算 H。[2] 定理6.4 还揭示了充足预算区间中的基本权衡：相对预算线性增长的代价是每次迭代具有二次样本复杂度。

定理 6.4 (Gamma 模型下的线性相对预算增长)。令 $m \in \{1,2,\ldots\}$和 $\delta \in (0,1]$ 为任意值。在伽马分布模型下，通过使用 $\kappa_i := \left(\min\left\{\frac{\sigma_J(\pi^{(i)})}{\pi^{(i)}},\ \frac{H-J(\pi^{(i)})}{2\sigma(\pi^{(i)})}\right\}\right)^2$ 重复定理 6.2 中的迭代，可知以至少 ~~$1-\delta$~~的概率，~~对于~~每个 $i \in [m]$,

$$\xi_i - \xi_0 \geq i\,(2K)^{-1} - \mathcal{O}_K(1), \tag{23}$$

$$J(\pi^{(i)}) = H\left(1 - \xi_i^{-1} + \mathcal{O}_K\left(\xi_i^{K-1}e^{-K\xi_i}\right)\right), \tag{24}$$

$$\sigma^2(\pi^{(i)}) = H^2\left((\xi_i^2 K)^{-1} + \mathcal{O}_K\left(\xi_i^{K-1}e^{-K\xi_i}\right)\right), \tag{25}$$

$$c_0^{(i)} = \Theta_K(1), \tag{26}$$

只要展开次数 n_i 从 $\pi^{(i)}$ 满足

$$n_i \geq \frac{2CH\log\left(m\,|\mathcal{R}|\,/\delta\right)}{\sqrt{\kappa_i}\,c_0^{(i)}\sigma(\pi^{(i)})} = \widetilde{\Theta}\left(K\xi_i^2\right). \tag{27}$$

Here, C is a universal constant implied by Lemma 3.2, and $c_0^{(i)}$ is the anti-concentration coefficient of $\pi^{(i)}$, defined as $c_0^{(i)} := \mathbb{P}_{x\sim\rho,\tau\sim\pi^{(i)}(\cdot|x)}[R(\tau) \geq J(\pi^{(i)}) + \sqrt{\kappa_i}\,\sigma(\pi^{(i)})]$.

第 E.3 节中的证明基于渐近评估，该评估使用了定理 6.2 的逐步保证以及第5.2节中的统计分析。

备注 6.5。在伽马分布模型下，在线强化学习可以改进任意基础策略，使平均求解时间变得任意小。尽管这看起来可能不现实，但每次迭代都需要从当前策略中进行无界次数的（即 $\widetilde{\Omega}_K(i^2)$ 次）采样轨迹。此外，在充足预算区间内，使用伽马分布进行连续近似可能会与负二项分布存在显著差距。

7. 实验

7.1. 相对预算理论的实证验证

为验证定理5.2所预测的依赖于区制的行为，我们在 GSM8K（`Llama-3.2-3B-Instruct`(Grattafiori 等人, 2024年）、`Phi4-mini-instruct`(Abouelenin 等人, 2025年）和`Qwen3-4B-Instruct`(Yang 等人，2025年）以及 MATH-500（Cobbe 等人，2021；Hendrycks 等人，2021；Lightman 等人，2024年上进行了评估。我们测量了不同相对预算下的奖励方差和反集中系数 c_0，其中相对预算 $\xi = H/\mu_{x^0}$。详细设置和更多结果见附录B 。

[2]由于 Gamma 模型假设对任意 $x \in \mathcal{X}$，$T(\tau)\ |\ x \sim$伽马分布(K, p)，我们从整个问题集 $\mathcal{X}$ 中抽取输入 x，而不是从某个输入切片中抽取（从而去掉 J 的下标）。

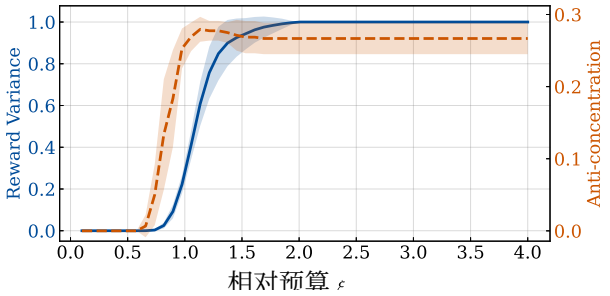


图4。 对于 Llama-3.2-3B-Instruct 在 GSM8K 上的任务，奖励方差和反集中系数随相对预算 ξ 的变化。相变发生在 $\xi \approx 1.0$，在我们的实验中，学习信号在 $\xi \approx 1.5$–1.8 时达到最大值。

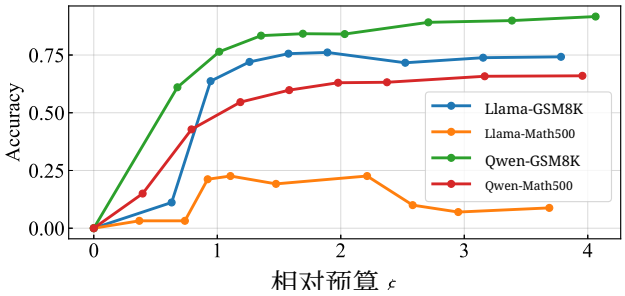


图5。 通过可验证奖励强化学习训练的大语言模型中，相对预算 ξ 与准确率之间的关系。

图4 展示了使用`Llama-3.2-3B-Instruct`在 GSM8K 上得到的代表性结果，与我们的理论框架一致（见图2a 和2b）。在所有评估的模型和数据集上，我们都观察到一致的依赖于区制的行为（完整图表见第B 节）。首先，我们观察到在 $\xi \approx 1.0$附近发生相变。在不足区间（$\xi < 0.8$），两项指标都微乎其微，表明学习信号实际上消失了。尽管反集中系数 c_0 不断上升，并在 $\xi \approx 1.2$附近达到峰值，但奖励方差在更大的相对预算下趋于平稳，在我们的实验中约为 $1.5 \leq \xi \leq 1.8$。这表明，尽管任务进入可解区间（约为 $\xi \approx 1.2$）后便能找到高奖励解，但要最大化学习信号，还需要更大的缓冲区（$H \approx 1.5\mu_x$–$1.8\mu_x$），以适应生成分布的长尾行为。最后，当 $\xi > 2.0$时，反集中系数趋于饱和，并稳定在较高水平，这支持了强化学习在充足预算区制下的稳健性。

7.2. 通过可验证奖励强化学习提升性能

为验证定理 5.2所预测的学习动态，我们使用组相对策略优化（GRPO，`Llama-3.2-3B-Instruct` 和 `Qwen3-4B-Instruct` 对其进行了微调（Shao 等人 (2024年)）。我们测量了模型在两个推理任务上的训练后准确率。首先，我们在 GSM8K 上进行训练和评估。其次，为评估模型在更具挑战性的任务上的表现，我们在 DAPO-Math-17k-Processed 上训练模型（Yu 等人，2025年），并在 MATH-500 测试集上评估模型。我们将令牌预算 H 定义为最大补全文本长度。在固定任务分布的条件下，我们改变令牌预算 H 以遍历不同的相对预算区间 $\xi := H/\mu$(在 μ下估计 π_b）。对于每个 H，我们都训练固定步数的独立模型，并测量测试集准确率。

图5 展示了相对预算 ξ 与各模型及数据集上训练后测试准确率之间的关系。与我们的理论框架一致，该结果呈现出由相对预算支配的相变行为。在不足区间（$\xi \ll 1$）中，准确率仍然可忽略不计，这与我们的分析一致：信息性轨迹极为稀少，模型无法学习有效的推理路径。在平衡区间，随着 ξ 趋近于1，我们观察到急剧的性能提升。这对应于相变点，此时计算预算 H 与固有任务难度 μ_x相当，使高奖励解变得可发现，并实现高效学习。在超出 $\xi \approx 1.5$的充足区间，性能增益趋于饱和：边际收益递减，因为任务已充分处于模型能力范围之内。这些一致的发现支持将相对预算 ξ 作为可验证奖励强化学习的效能的关键因素。

8. 结论与未来工作

我们提出了相对预算理论，将计算约束与任务难度统一起来，以分析基于验证器的强化学习的样本复杂度。我们的分析表明，学习效率取决于求解时间分布的反集中性，并据相对预算 ξ 确定了三个区间。平衡区间（$\xi \approx 1$）会出现相变：这一点为强化学习提供了最强的学习信号，但由于解的异质性达到峰值，对监督微调而言却最具挑战性。我们的实证结果支持了这些动态规律，并确定了与理论相符的最优预算 $\xi \in [1.5, 2.0]$。我们还证明，在线强化学习会推动相对预算线性增长，为测试时扩展提供了理论依据。在实践中，我们的研究结果表明，应将计算资源分配设定在预期难度（$\xi \gtrsim 1$）略高的位置，以确保学习信号可靠，同时避免预算过多造成的低效。

尽管本文为基于验证器的强化学习提供了理论基础，但为了使方差分析易于处理，我们的分析依赖于连续代理奖励。将这一理论扩展到二元奖励设定（即对最终的 1/0 结果进行监督），仍是一个亟待解决的关键问题，这对于描述标准推理流程以及构建更通用的推理理论都至关重要。

致谢

TS 和 RH 的研究部分得到了 JSPS KAKENHI（ 24K02905）和 JST CREST（JPMJCR2115）的资助。本研究由新加坡国家研究基金会、资讯通信媒体发展局通过其可信科技资助计划，以及数字发展与信息部通过人工智能访问学者计划（资助编号 AIVP-2024-004）提供支持。本材料中表达的任何意见、研究结果、结论或建议均属于作者观点，不代表新加坡国家研究基金会、资讯通信媒体发展局及数字发展与信息部的立场。

影响声明

本研究提出了一个框架，用于判断推理强化学习何时可能出现收益递减，从而推动更可持续的人工智能发展。本文为判断推理强化学习何时可能有效、何时可能造成浪费提供指导，使研究和开发团队能够避免不必要的训练，并更合理地分配有限的计算预算。本文的见解有助于降低大规模模型训练带来的财务成本和环境影响。不过，尽管这项贡献属于理论研究，但一种可能的负面结果是，某些组织可能利用这些结论为增加总体计算支出以追求性能提升辩护，从而可能增加排放。

参考文献

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C. 等。Phi-4-mini 技术报告：通过 LoRA 混合构建紧凑而强大的多模态语言模型。*arXiv* 预印本 *arXiv:2503.01743*，2025年。

Albalak, A., Elazar, Y., Xie, S. M., Longpre, S., Lambert, N., Wang, X., Muennighoff, N., Hou, B., Pan, L., Jeong, H., 等。《语言模型数据选择综述》。《机器学习研究汇刊》，2024年。

Alomrani, M. A., Zhang, Y., Li, D., Sun, Q., Pal, S., Zhang, Z., Hu, Y., Ajwani, R. D., Valkanas, A., Karimi, R., 等。《精打细算的推理：大语言模型自适应与可控测试时计算综述》。*arXiv* 预印本 *arXiv:2507.02076*，2025年。

Arora, D. 和 Zanette, A. 训练语言模型以高效推理。*arXiv* 预印本 *arXiv:2502.04463*，2025年。

Bae, S.、Hong, J.、Lee, M. Y.、Kim, H.、Nam, J. 与 Kwak, D. 面向推理的在线难度筛选 re-

强化学习。 *arXiv* 预印本 *arXiv:2504.03380*，2025年。

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R.等。训练验证器以解决数学文字题。*arXiv*预印本 *arXiv:2110.14168*，2021年。

Cui, G., Yuan, L., Wang, Z., Wang, H., Zhang, Y., Chen, J., Li, W., He, B., Fan, Y., Yu, T.等。通过隐式奖励进行过程强化学习。*arXiv*预印本 *arXiv:2502.01456*，2025年。

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., 等。《Llama 3模型群》。*arXiv*预印本 *arXiv:2407.21783*，2024年。Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., 等。DeepSeek-R1通过强化学习激励大语言模型进行推理。《自然》，645(8081):633–638，2025年。Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D. 和 Steinhardt, J. 使用MATH数据集衡量数学问题求解能力。收录于神经信息处理系统数据集与基准赛道，2021年。Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Car- ney, A., 等。OpenAI o1系统卡。*arXiv*预印本 *arXiv:2412.16720*，2024年。Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, C., Gray, S., Radford, A., Wu, J. 和 Amodei, D. 神经语言模型的标度律。*arXiv*预印本 *arXiv:2001.08361*，2020年。Kazemnejad, A., Aghajohari, M., Portelance, E., Sordoni, A., Reddy, S., Courville, A. 和 Roux, N. L. VinePPO：改进大语言模型强化学习训练中的信用分配。收录于国际机器学习会议（*ICML*），2025年。Ke, Z., Jiao, F., Ming, Y., Nguyen, X.-P., Xu, A., Long, D. X., Li, M., Qin, C., Wang, P., Savarese, S., 等。大语言模型推理前沿综述：推理缩放、学会推理与智能体系统。*arXiv*预印本 *arXiv:2504.09037*，2025年。Kim, J., Wu, D., Lee, J. D. 和 Suzuki, T. 思维链推理的亚稳态动力学：搜索、强化学习与蒸馏的可证明优势。收录于国际机器学习会议，2025a。Kim, M., Shrestha, A., Shrestha, S., Nepal, A. 和 Ross, K. 强化学习与蒸馏之比较：理解

大语言模型推理中的准确性与能力。*arXiv*预印本 *arXiv:2505.14216*，2025b。

Kimi 团队、Du, A., Gao, B., Xing, B., Jiang, C., Chen, C., Li, C., Xiao, C., Du, C., Liao, C. 等。Kimi k1.5：使用大语言模型扩展强化学习。*arXiv*预印本 *arXiv:2501.12599*，2025年。

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I. 和 Cobbe, K. 让我们逐步验证。载于国际学习表征会议，2024年。

Liu, W., Zhou, R., Deng, Y., Huang, Y., Liu, J., Deng, Y., Zhang, Y. 和 He, J. 通过自适应长度奖励塑形学习高效推理。*arXiv*预印本 *arXiv:2505.15612*，2025年。

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Cand` es, E.与Hashimoto, T. B. s1：简单测试时扩展。载于*2025*年自然语言处理实证方法会议论文集，2025年。

Setlur, A., Rajaraman, N., Levine, S.与Kumar, A. 在没有验证或强化学习的情况下扩展测试时计算是次优的。载于国际机器学习会议（*ICML*），2025年。

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.等。Deepseekmath：推动开放语言模型中数学推理的极限。*arXiv*预印本*arXiv:2402.03300*，2024年。

Snell, C.、Lee, J.、Xu, K. 和 Kumar, A. 最优扩展大语言模型测试时计算的效果可能优于扩展模型参数。*arXiv* 预印本 *arXiv:2408.03314*，2024年。

Sun, Y.、Shen, J.、Wang, Y.、Chen, T.、Wang, Z.、Zhou, M. 和 Zhang, H. 通过难度导向的在线数据选择和轨迹重放提升大语言模型强化微调的数据效率。载于第三十九届神经信息处理系统年会，2025年。

Vladimirova, M.、Girard, S.、Nguyen, H. 和 Arbel, J. 次 Weibull 分布：将次高斯和次指数性质推广至更重尾分布。*Stat*，9(1)：e318，2020年。

Welleck, S., Bertsch, A., Finlayson, M., Schoelkopf, H., Xie, A., Neubig, G., Kulikov, I. 和 Harchaoui, Z. 从解码到元生成：大语言模型的推理时算法。《机器学习研究汇刊》，2024年。

Wu, F., Xuan, W., Lu, X., Liu, M., Dong, Y., Harchaoui, Z. 和 Choi, Y. 看不见的缰绳：为什么可验证奖励强化学习可能会也可能不会脱离其原点。*arXiv* 预印本 *arXiv:2507.14843*，2025a。

Wu, M.、Zhang, Z.、Dong, Q.、Xi, Z.、Zhao, J.、Jin, S.、Fan, X.、Zhou, Y.、Lv, H.、Zhang, M. 等。推理还是记忆？数据污染导致强化学习结果不可靠。*arXiv*预印本 *arXiv:2507.10532*，2025b。

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., 等。通义千问3技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025年。

Ye, T., Xu, Z., Li, Y., 和 Allen-Zhu, Z. 语言模型的物理学：第 2.2 部分，如何从小学数学题的错误中学习。*arXiv* 预印本 *arXiv:2408.16293*，2024年。

Ye, Y., Huang, Z., Xiao, Y., Chern, E., Xia, S., 和 Liu, P. Limo：少即是多的推理。*arXiv* 预印本 *arXiv:2502.03387*，2025年。

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., 等。DAPO：大规模开源大语言模型强化学习系统。*arXiv* 预印本 *arXiv:2503.14476*，2025年。

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., and Huang, G. 强化学习是否真的能激励大语言模型超越基础模型的推理能力？载于第三十九届神经信息处理系统年会，2025年。

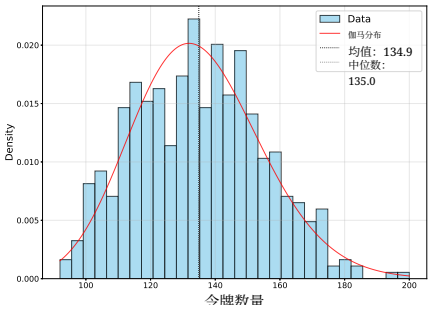
Zhang, C., Neubig, G., and Yue, X. 关于预训练、中期训练与强化学习在推理语言模型中的相互作用。*arXiv*预印本 *arXiv:2512.07783*，2025年。

Zhou, Y.、Liu, H.、Chen, Z.、Tian, Y. 和 Chen, B. GSM-Infinite：当上下文长度和推理复杂度无限增长时，你的大语言模型会如何表现？*arXiv* 预印本 *arXiv:2502.05252*，2025年。

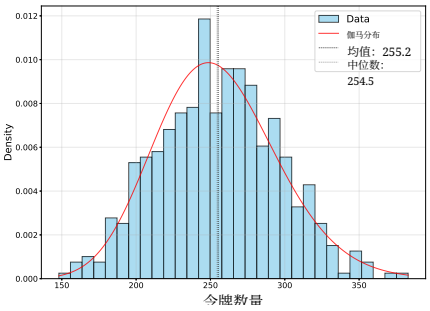
A. 词元分布

为验证理论 第5.2 节和6.3 节中关于推理步骤服从伽马分布的假设，w e
分析了词元数量 由三种流行的大语言模型生成。

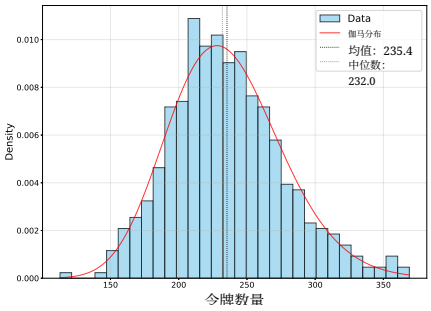
图6 展示了 `Llama-3.2-3B-Instruct` 和 `Phi-4-mini-instruct` 在 GSM8K（Cobbe 等人，2021年）和 MATH-500（Hendrycks 等人，2021年；Lightman 等人，2024年）数据集上的解答长度分布。在每个子图中，蓝色直方图表示生成词元的经验分布，红色曲线则表示拟合的伽马分布。观测数据与拟合曲线高度吻合，这一结果从经验上支持了我们的建模选择，表明伽马分布可作为复杂推理任务中离散词元计数的可靠连续近似。



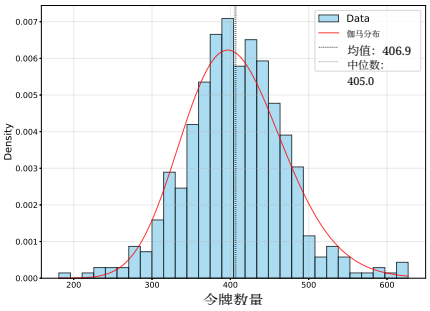
(a) Llama-3.2-3B-Instruct + GSM8K。



(b) Llama-3.2-3B-Instruct + MATH500。



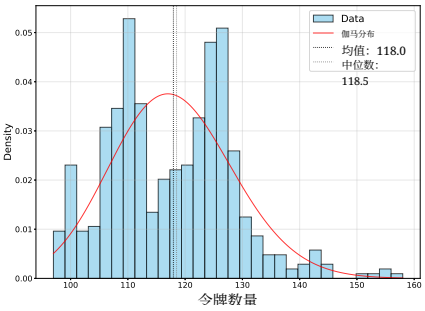
(*c*) Phi-4-mini-instruct + GSM8K。



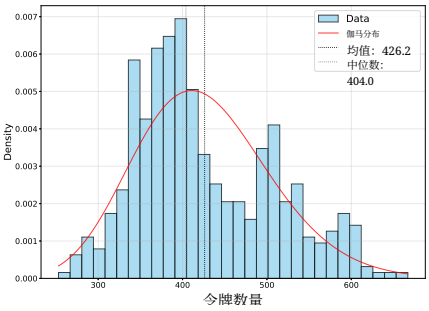
(*d*) Phi-4-mini-instruct + MATH500。

图*6.* 词元分布生成 大语言模型在推理数据中为一个问题实例生成的 数据集及其伽马分布拟合。

然而，像`Qwen3-4B-Instruct` 这样的模型会呈现如下所示的双峰分布。近期研究将这种行为归因于数据污染：模型会在检索记忆解答和进行真正的推理之间摇摆（Wu 等人，2025b）。我们的基于伽马分布的理论刻画的是后一种情况，而这些例外则需要采用混合模型方法。请注意，正如第B节所述，即使对于这类异常模型，我们的核心理论框架依然成立。



(*a*) Qwen3-4B-Instruct + GSM8K。



(b) Qwen3-4B-Instruct + MATH500。

图7。 词元分布 由Qwen3-4B-Instruct生成。实证数据呈现双峰结构，显示出明显的 r
偏离单一伽 mma 分布拟合。

B. 静态分析的实验详情

在第7.1节中，我们进行了实证分析，以验证相对预算理论。详细的实验设置如下。数据收集方面，我们从 GSM8K 和 MATH-500 测试集中随机抽取了 20 道题目。针对每道题目，我们使用 100 的解答轨迹，分别由 `Llama-3.2-3B-Instruct`、`Phi-4-mini-instruct`和`Qwen3-4B-Instruct` 在温度 $T = 0.7$ 和核采样概率 $= 0.9$下生成。为估计真实难度，我们设置了足够大的最大补全文本长度（1024 个词元），并根据正确轨迹计算解答长度的均值 μ_x 。随后，我们模拟了 $\xi \in [0.1, 4.0]$范围内不同的预算约束。塑形奖励计算如下：$R(\tau) = \max\{0, H - T(\tau) + 1\}$，适用于成功轨迹；否则奖励为0。为确保统计结果可靠，我们对各题目的归一化方差曲线进行了汇总。

除了图4 中`Llama-3.2-3B-Instruct` 在 GSM8K 上的结果外，图8 和9 还展示了使用`Llama-3.2-3B-Instruct` 和 `Phi-4-mini-instruct`的其他实验结果。可以看到，以下图表呈现出与图4相似的趋势。

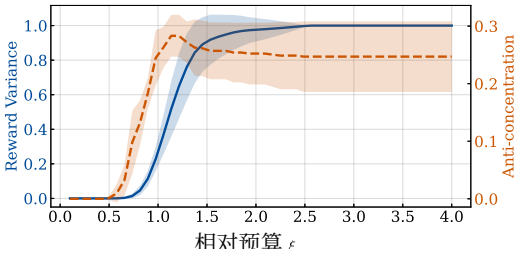
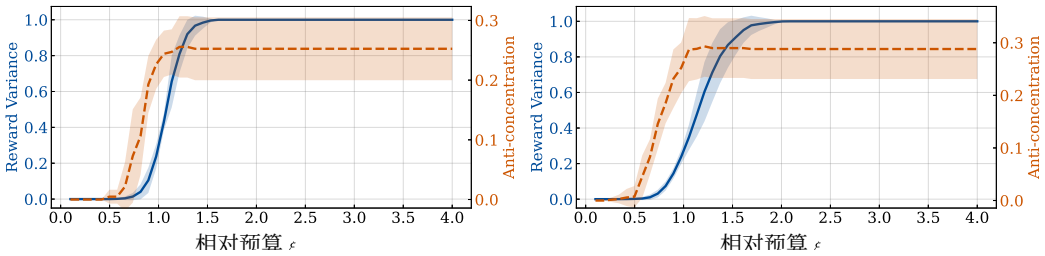


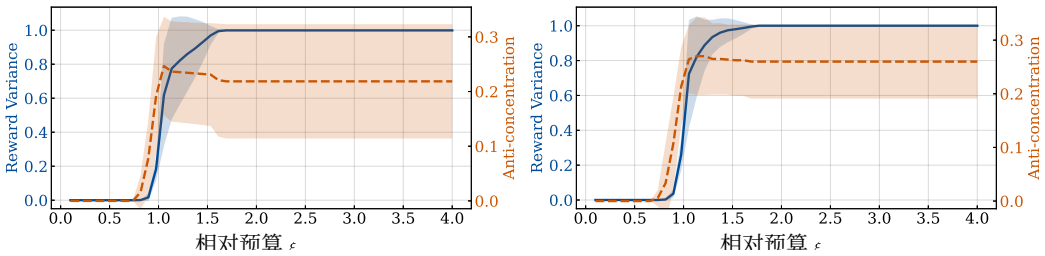
图 *8.* Llama-3.2-3B-Instruct + MATH-500。



*(a)*Phi-4-mini-instruct + GSM8K. (*b*) Phi-4-mini-instruct + MATH-500。

图 *9.* 相对预算区间中的相变。归一化奖励方差（蓝色）和反集中系数（橙色）随相对预算 ξ 的变化。在 $\xi \approx 1.0$ 处发生相变，学习信号在 $\xi \approx 1.5$–2.0 时达到最大值。我们的实验中。

我们还给出了使用`Qwen3-4B-Instruct` 在 GSM8K 和 MATH-500 上进行实验所得结果的以下图示。尽管 `Qwen3-4B-Instruct` 生成的词元分布并不像第 A节所讨论的那样与伽马分布一致，但我们可以看到，反集中现象正如我们的理论所预测的那样。



*(a)*Qwen3-4B-Instruct + GSM8K。 *(b)* Qwen3-4B-Instruct + MATH-500。

图 *10.* Qwen3-4B-Instruct 的实验结果。与其他大语言模型一样，我们观察到相对预算区间中相变的类似趋势。归一化奖励方差（蓝色）和反集中系数（橙色）随相对预算 ξ 的变化。相变发生在 $\xi \approx 1.0$，学习信号在我们的实验中于 $\xi \approx 1.5$ 达到最大。

C. 强化学习微调的实验细节

在第 7.2节中，我们使用 GRPO 算法执行可验证奖励强化学习。本节详细介绍超参数和实现细节。我们使用了 A100 GPU 环境。基础模型为 `Llama-3.2-3B-Instruct` 和`Qwen3-4B-Instruct`。为提高内存效率，我们在加载过程中采用了 4 位量化，但 LoRA 适配器使用 16 位精度进行训练。我们使用了 trl（`https://github.com/huggingface/trl`）库中的 GRPOTrainer。我们通过设置配置来探索令牌预算 H 对学习动态的影响。为了验证主论文中提出的相对预算理论的核心论断，我们通过改变 GRPO 配置中的 $\xi = H/\mu_x$ 最大补全文本长度 参数来调节相对预算。 第 7.2 节结果所使用的具体超参数列于表 1 和 2中。

表 *1*。 Llama-3.2-3B-Instruct 上 GRPO 微调的超参数。

超参数	值
LoRA 秩	64
LoRA Alpha	64
目标模块	q_投影、k_投影、v_投影、o_投影，门控投影、上投影、下投影
优化器	8 位 AdamW
学习率	5×10^{-6}
学习率调度器	余弦（预热比例 0.1）
权重衰减	0.1
梯度累积步数	4
生成数量	4
最大梯度范数	1.0
最大步数	500

表 *2.* Qwen3-4B-Instruct 上 GRPO 微调的超参数。

超参数	值
LoRA 秩	32
LoRA Alpha	64
目标模块	q_投影、k_投影、v_投影、o_投影，门控投影、上投影、下投影
优化器	8 位 AdamW
学习率	5×10^{-6}
学习率调度器	线性
权重衰减	0.01
梯度累积步数	4
生成数量	4
最大梯度范数	1.0
最大步数	500

D. 第5节的证明

D.1. 预备引理

下面的引理用截断求解时间矩来表示奖励矩 s.

引理 D.1 （双层反馈下的奖励矩）。 对于任意固定的问题实例 $x \in \mathcal{X}$，令

$$q_x := \mathbb{P}_{\tau \sim \pi_b(\cdot|x)}[T(\tau) \le H] \tag{28}$$

表示基础策略下的成功概率。令 $\tilde{\mu}_x$ 和 $\tilde{v}_x$ 分别表示在成功条件下求解时间的均值和方差$T(\tau)$，定义如下：

$$\tilde{\mu}_x \coloneqq \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) \mid T(\tau) \le H], \tag{29}$$
$$\tilde{v}_x \coloneqq \mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) \mid T(\tau) \le H]. \tag{30}$$

那么， $R(\tau)$ 的均值和方差满足：

$$\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] = q_x\big(H+1-\tilde{\mu}_x\big), \tag{31}$$
$$\mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] = q_x\tilde{v}_x + q_x(1-q_x)\big(H+1-\tilde{\mu}_x\big)^2. \tag{32}$$

证明。 根据定义，我们有

$$R(\tau) = (H - T(\tau) + 1)\cdot \mathbb{I}\{T(\tau) \le H\}. \tag{33}$$

因此，以下等式链成立：

$$\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] = \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}\big[(H - T(\tau) + 1)\mathbb{I}\{T(\tau) \le H\}\big] \tag{34}$$
$$= q_x\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[H - T(\tau) + 1 \mid T(\tau) \le H] \tag{35}$$
$$= q_x(H+1-\tilde{\mu}_x)。 \tag{36}$$

对于二阶矩，我们有

$$\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)^2] = \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}\big[(H - T(\tau) + 1)^2\cdot \mathbb{I}\{T(\tau) \le H\}\big] \tag{37}$$
$$= q_x\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}\big[(H - T(\tau) + 1)^2 \mid T(\tau) \le H\big] \tag{38}$$
$$= q_x\big(\tilde{v}_x + (H+1-\tilde{\mu}_x)^2\big). \tag{39}$$

因此，以下方程成立：

$$\mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] = \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)^2] - \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)]^2 \tag{40}$$
$$= q_x\tilde{v}_x + q_x(1-q_x)(H+1-\tilde{\mu}_x)^2. \tag{41}$$

□

D.2. 定理5.1 的证明

证明。我们分析奖励 $R(\tau)$ 超过其均值 $\sqrt{\varepsilon}$ 个标准差的事件。根据双层奖励的定义 $R(\tau) = (H - T(\tau) + 1)\cdot\mathbb{I}\{T(\tau) \le H\}$，条件

$$R(\tau) \ge \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] + \sigma_{b,x}\sqrt{\varepsilon} \tag{42}$$

意味着轨迹成功，并可变换如下：

$$H + 1 - T(\tau) \ge \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] + \sigma_{b,x}\sqrt{\varepsilon} \tag{43}$$
$$T(\tau) \le H + 1 - \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] - \sigma_{b,x}\sqrt{\varepsilon}. \tag{44}$$

根据定理 D.1，奖励矩是条件求解时间矩的函数 $\tilde{\mu}_x := \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) \mid T(\tau) \le H]$ 和 $\tilde{v}_x := \mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) \mid T(\tau) \le H]$。特别地，定理 D.1 表明，式（44）右侧可以改写为相对于条件均值的偏差。因此，我们定义归一化阈值

$$\eta(x,\varepsilon) \coloneqq \frac{\min\{H, H+1-\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] - \sigma_{b,x}\sqrt{\varepsilon}\}}{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]}.$$

此外，令

$$\bar{\eta}(x,\varepsilon) \coloneqq [\eta(x,\varepsilon)]_0^{z_0}. \tag{45}$$

因此，以下不等式成立：

$$\mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau)\le \eta(x,\varepsilon)\,\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]\right]\ge \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau)\le \bar{\eta}(x,\varepsilon)\,\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]\right]. \tag{46}$$

因此，根据定理 4.3，

$$\mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau)\le \bar{\eta}(x,\varepsilon)\,\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]\right]\ge c_-\,f\big(\bar{\eta}(x,\varepsilon)\big). \tag{47}$$

□

D.3. 定理5.2的证明

证明。固定 $\xi>0$，并考虑切片 $\mathcal{X}(\xi)$。定义 $c_0(\xi;\kappa):=\inf_{x\in\mathcal{X}(\xi)}c_x(\kappa)$，其中 κ 是定理 3.1 中的信赖域半径。对于以下切片级界，我们还假设，在切片 $\mathcal{X}(\xi)$ 上，比较策略满足$\kappa_x\le\kappa$，对所有 $x\in\mathcal{X}(\xi)$ 均成立；这排除了这样一种退化解：将较大的散度分配给少数实例，从而满足平均 χ^2 约束。由于 $c_x(\varepsilon)$ 随 ε 增大而单调不增，因此对所有 $x\in\mathcal{X}(\xi)$，都有 $c_x(\kappa_x)\ge c_x(\kappa)\ge c_0(\xi;\kappa)$。因此，基础策略在 $\mathcal{X}(\xi)$ 上是 $c_0(\xi;\kappa)$-反集中（anti-concentrated）的。应用引理 3.2，以至少$1-\delta$ 的概率，式（14）成立。最后，命题 D.2、D.3 和 D.4 给出了 $c_0(\xi;\kappa)$ 所述的依赖区间的行为。

□

D.4. 相对预算不足区间

在 $\xi_x\to 0$ 区间内，计算预算 H 远小于预期求解时间 μ_x。因此，我们有$H\ll\mu_x$，而满足 $T(\tau)\le H$的成功解属于稀有事件。具体而言，根据假设 4.3， $q_x=\Theta\big(f(\xi_x)\big)$当 $\xi_x\to 0$时。以下命题刻画了这一较低成功概率如何影响奖励方差和反集中性质。

命题 D.2 (相对预算不足)。 在 $\xi_x\to 0$区间内，在假设 *4.2* 和 *4.3*下，强化学习奖励方差与反集中系数共同满足

$$\sigma_{b,x}=\mathcal{O}\big(H\sqrt{f(\xi_x)}\big)\quad \textit{and}\quad c_0(\xi)=\Theta\big(f(\xi)\big). \tag{48}$$

证明。 由于在 $\xi\to 0$ 区间内，成功推理（$T(\tau)\le H$）是稀有事件，因此 $R(\tau)=0$以很高的概率成立。因此，奖励方差主要由这些罕见的成功事件决定。利用 $q_x=\Theta\big(f(\xi_x)\big)$,

$$\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)^2]=\sum_{t=1}^{H}(H+1-t)^2\mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)=t]\le\sum_{t=1}^{H}(H+1)^2\mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)=t]=(H+1)^2q_x.$$

由于在此区间内，$\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)]$ 相较于二阶矩可以忽略，因此我们有 $\mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)]\asymp\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)^2]$,，由此可得 $\sigma_{b,x}=\mathcal{O}\big(H\sqrt{f(\xi_x)}\big)$。

至于 反集中系数，引理 5.1 指出，存在一个系数 $\eta(x,\varepsilon)$ 使得

$$c_x(\varepsilon)=\mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\big[T(\tau)\le\eta(x,\varepsilon)\cdot\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]\big]. \tag{49}$$

在相对预算不足区间内，若 $H\ll\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]$,一条轨迹要达到高奖励，就必须在预算 H 内解决问题。因此，引理 5.1 意味着时间阈值 $\eta(x,\varepsilon)\,\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]$ 与H相当。更准确地说，存在与 x无关的常数 $0<c_1\le c_2<\infty$，使得对所有充分小的 ξ_x,

$$c_1H\le\eta(x,\varepsilon)\,\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]\le c_2H. \tag{50}$$

代入 $\xi_x=H/\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]$,，这等价于

$$c_1\xi_x\le\eta(x,\varepsilon)\le c_2\xi_x. \tag{51}$$

现在我们将假设 4.3 应用于 $t=\eta(x,\varepsilon)\,\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]$。由于 $\eta(x,\varepsilon)\le c_2\xi_x$，且我们考虑 $\xi_x\to 0$，对于所有充分小的 ξ_x，都有 $t/\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]\le z_0$，这落在该假设的有效范围内。因此，

$$c_-\,f\big(\eta(x,\varepsilon)\big)\le c_x(\varepsilon)\le c_+\,f\big(\eta(x,\varepsilon)\big). \tag{52}$$

由于 f 是非递减的 ng 且在原点附近满足倍增性质，因此存在常数 $C_1, C_2 > 0$，使得 at

$$C_1\, f(\xi_x) \le f\big(\eta(x,\varepsilon)\big) \le C_2\, f(\xi_x). \tag{53}$$

对于所有充分小的 ξ_x，我们有

$$c_- C_1\, f(\xi_x) \le c_x(\varepsilon) \le c_+ C_2\, f(\xi_x), \tag{54}$$

因此，在 $\xi_x \to 0$的范围内，以下方程成立：

$$c_x(\varepsilon) = \Theta\big(f(\xi_x)\big) \tag{55}$$

最后，固定一个较小的 $\xi > 0$，并考虑切片 $\mathcal{X}(\xi) := \{x : \xi/2 \le \xi_x \le 2\xi\}$。根据 f的单调性和倍增性质，我们有 $f(\xi_x) = \Theta(f(\xi))$，且该关系在 $x \in \mathcal{X}(\xi)$上均匀成立。对 $x \in \mathcal{X}(\xi)$取下确界，得到

$$c_0(\xi) := \inf_{x \in \mathcal{X}(\xi)} c_x(\varepsilon) = \Theta\big(f(\xi)\big). \tag{56}$$

因此，我们得到了所需结果。 □

D.5. 均衡相对预算状态

在 $\xi_x = \Theta(1)$ 的状态中（例如 $H = \Theta(\mu_x)$），成功与失败均以概率 $\Theta(1)$ 发生。因此，在 $T(\tau) \le H$处的截断并不会实质改变 $T(\tau)$ 的矩的尺度。我们现在给出如下命题，用以刻画均衡相对预算状态中的奖励方差与反集中。

命题 D.3（均衡相对预算）。 设假设 *4.2* 与 *4.3*成立，并令 $\varepsilon \le (q_{\min}/2)^2$。固定常数$0 < \xi_{\min} \le \xi_{\max} < \infty$，并定义 $\mathcal{X}_b := \{x : \xi_x \in [\xi_{\min}/2,\, 2\xi_{\max}]\}$。此外，设定理 *4.4* 在$\mathcal{X}_b$ 上一致成立，并固定任意 $\xi \in [\xi_{\min}, \xi_{\max}]$。则对所有实例 $x \in X(\xi)$，强化学习奖励标准差与反集中系数满足

$$\sigma_{b,x} = \Theta(H) \quad \textit{and} \quad c_0(\xi) = \Theta(1). \tag{57}$$

证明。 固定一个实例 $x \in \mathcal{X}_b$（即，$\xi_x \in [\xi_{\min}, \xi_{\max}]$）。根据假设 4.2，

$$\mathbb{V}_{\tau \sim \pi_b(\cdot|x)}[T(\tau)] = \Theta\left(\mathbb{E}_{\tau \sim \pi_b(\cdot|x)}[T(\tau)]^2\right). \tag{58}$$

在平衡 机制下，我们有 $\mu_x = H/\xi_x = \Theta(H)$。因此，无条件方差满足

$$v_x := \mathbb{V}_{\tau \sim \pi_b(\cdot|x)}[T(\tau)] = \Theta(H^2). \tag{59}$$

我们现在考虑成功条件下的方差。根据假设 4.4，在 H 处的截断保持了方差的阶：

$$\tilde{v}_x := \mathbb{V}_{\tau \sim \pi_b(\cdot|x)}[T(\tau) \mid T(\tau) \le H] \ge c_v v_x. \tag{60}$$

根据引理 D.1，奖励的方差 $R(\tau)$ 以条件方差项为下界 :

$$\mathbb{V}_{\tau \sim \pi_b(\cdot|x)}[R(\tau)] = q_x \tilde{v}_x + q_x(1-q_x)(H + 1 - \tilde{\mu}_x)^2 \ge q_x \tilde{v}_x. \tag{61}$$

利用成功 概率 $q_x \ge q_{\min}$ 上界（由假设 4.4 给出）以及式（60）中的方差上界，可得

$$\mathbb{V}_{\tau \sim \pi_b(\cdot|x)}[R(\tau)] = \Omega(H^2). \tag{62}$$

由于奖励有界于 $H+1$，方差显然有上界 $\mathcal{O}(H^2)$。因此，标准差按 $\sigma_{b,\ x} = \Theta(H)$ 的量级变化。

由引理 5.1，对每个 $\varepsilon > 0$，存在 $\eta(x,\varepsilon) \in (0,1]$，使得

$$c_x(\varepsilon) = \mathbb{P}_{\tau \sim \pi_b(\cdot|x)}[T(\tau) \le \eta(x,\varepsilon)\, \mathbb{E}_{\tau \sim \pi_b(\cdot|x)}[T(\tau)]]. \tag{63}$$

只需证明 $\eta(x,\varepsilon)$ 一致地远离 0，在 $x \in \mathcal{X}_b$ 上对固定的 ε（等价地，对 ε 位于某个固定紧区间内）。回顾 $\eta(x,\varepsilon)$ 在引理 5.1 中的定义：

$$\eta(x,\varepsilon) = \frac{\min\{H,\ H+1-\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] - \sigma_{b,x}\sqrt{\varepsilon}\}}{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)]}. \tag{64}$$

利用 $R(\tau) \le H$ 和 $q_x \le 1-q_{\min}$，我们有 $\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] \le (1-q_{\min})H$；因此，

$$H+1-\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] \ge q_{\min}H. \tag{65}$$

此外，$\sigma_{b,\ x} \le H$ 由于 $R(\tau) \in [0,H]$。因此，对于任意固定的 $0 < \varepsilon \le (q_{\min}/2)^2$，

$$H+1-\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] - \sigma_{b,x}\sqrt{\varepsilon} \ge (q_{\min}-\sqrt{\varepsilon})H \ge \frac{q_{\min}}{2}H. \tag{66}$$

由于在 $\mathcal{X}_b$ 上，$\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)] = \mu_x = \Theta(H)$，我们得到一个一致的下界

$$\eta(x,\varepsilon) \ \ge\ \frac{(q_{\min}/2)H}{\mu_x} = \frac{q_{\min}}{2}\,\xi_x \ \ge\ \frac{q_{\min}}{2}\,\xi_{\min} =: \eta_{\min} > 0. \tag{67}$$

令 $\bar{\eta}(x,\varepsilon) := [\eta(x,\varepsilon)]_0^{z_0}$。则假设 4.3 给出

$$c_x(\varepsilon) \ge \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) \le \bar{\eta}(x,\varepsilon)\,\mu_x] \ge c_- f(\bar{\eta}(x,\varepsilon)) \ge c_- f(\min\{\eta_{\min}, z_0\}) =: c_\star > 0. \tag{68}$$

对 $x \in X(\xi)$（其中 ξ 处于平衡范围内）取下确界，得到 $c_0(\xi) \ge c_\star$，而显然有 $c_0(\xi) \le 1$。因此$c_0(\xi) = \Theta(1)$ 在平衡区间内成立，从而证明了（57）。 □

D.6. 充裕相对预算区间

在 $\xi \to \infty$ 区间内，我们有 $H \gg \mu_x$。具有 $T(\tau) > H$的失败事件很少，因此 $q_x \to 1$，且在 H 处的截断可以忽略不计。奖励近似为 $R(\tau) \approx H - T(\tau)$。以下命题刻画了在这一充裕相对预算区间内，奖励方差和反集中行为的特征。

命题 D.4 (充裕相对预算)。 假设 *4.2* 和 *4.3* 成立，并固定 $\varepsilon \le \varepsilon_0$ 为一个足够小的普适常数 $\varepsilon_0 \in (0,1)$。在区间 $\xi_x \to \infty$内，强化学习奖励标准差与反集中系数共同满足

$$\sigma_{b,x} = \mathcal{O}(H/\xi_x) \quad \textit{and} \quad c_0(\xi) = \Theta(1). \tag{69}$$

证明。 在充裕相对预算区间内，我们有 $\xi_x \to \infty$。根据马尔可夫不等式，

$$1-q_x = \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T(\tau) > H] \le \frac{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2]}{H^2}. \tag{70}$$

根据假设 4.2，由于 $v_x = \mathbb{V}[T(\tau)] = \Theta(\mu_x^2)$，我们有

$$\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2] = \mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)] + \mu_x^2 = \Theta(\mu_x^2), \tag{71}$$

因此 $1-q_x = \mathcal{O}(\mu_x^2/H^2) = \mathcal{O}(1/\xi_x^2)$。

令 $R(\tau)$ 表示在 $\pi_b(\cdot \mid x)$ 下的塑形奖励，并回忆 $\sigma_{b,x}^2 := \mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)]$。引理 D.1 给出 s

$$\mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] = q_x\tilde{v}_x + q_x(1-q_x)(H+1-\tilde{\mu}_x)^2. \tag{72}$$

由于 $\tilde{v}_x \le \mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2 \mid T(\tau) \le H]$,

$$\tilde{v}_x \le \frac{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2\,\mathbb{I}\{T(\tau)\le H\}]}{\mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)\le H]} \le \frac{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2]}{q_x} = \mathcal{O}(\mu_x^2), \tag{73}$$

使用 $\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2] = \Theta(\mu_x^2)$ 和 $q_x \to 1$。

接下来，注意到 $H = \xi_x\mu_x$ 和 $\tilde{\mu}_x \le H$ 意味着 $0 \le H+1-\tilde{\mu}_x \le H+1 = \Theta(H)$。因此，(72) 中的第二项满足

$$q_x(1-q_x)\,(H+1-\tilde{\mu}_x)^2 \le (1-q_x)\,(H+1)^2 = \mathcal{O}\left(\frac{\mu_x^2}{H^2}\right)H^2 = \mathcal{O}(\mu_x^2). \tag{74}$$

合并 (72) 中的两个界可得

$$\sigma_{b,x} = \sqrt{\mathbb{V}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)]} = \mathcal{O}(\mu_x). \tag{75}$$

根据 引理 5.1，对于每个 $(x,\ \varepsilon)$，存在一个标量 $\eta(x,\ \varepsilon) > 0$ 使得

$$c_x(\varepsilon) = \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau) \le \eta(x,\varepsilon)\,\mu_x\right]. \tag{76}$$

为明确体现截断效应，注意到

$$\mu_x - \tilde{\mu}_x = \frac{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)\mathbb{I}\{T(\tau)>H\}]}{q_x} \le \frac{\sqrt{\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)^2]\mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T(\tau)>H]}}{q_x}, \tag{77}$$

根据马尔可夫不等式中的 $\mathbb{P}_{\tau\sim\pi_b}(\cdot|x)\ [T(\tau) > H] = O(\mu_x^2/H^2)$ 以及定理 4.2 中的 $\mathbb{E}_{\tau\sim\pi_b}(\cdot|x)\ [T(\tau)^2] = \Theta(\mu_x^2)$，可得 $\mu_x - \tilde{\mu}_x = \mathcal{O}(\mu_x^2/H) = \mathcal{O}(\mu_x/\xi_x)$，从而 $\tilde{\mu}_x/\mu_x \to 1$ 为 $\xi_x \to \infty$。此外，引理 5.1 给出

$$\eta(x,\varepsilon)\mu_x = \min\{H,\ H+1-\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] - \sigma_{b,x}\sqrt{\varepsilon}\}, \tag{78}$$

并且由于 $\mathbb{E}_{\tau\sim\pi_b(\cdot|x)}[R(\tau)] = q_x(H+1-\tilde{\mu}_x)$ 且 $\sigma_{b,x} = \mathcal{O}(\mu_x)$，对于所有充分大的 ξ_x，我们有 $\eta(x,\varepsilon) \ge \tilde{\mu}_x/\mu_x - C\sqrt{\varepsilon}$，其中某个常数为 C。因此，取足够小的 $\varepsilon_0 > 0$，可得对所有$\varepsilon \le \varepsilon_0$，$\eta(x,\varepsilon) \ge \eta_{\min} > 0$ 均一致成立，这意味着根据假设 4.3，$c_x(\varepsilon) = \Theta(1)$。

令 z_0 为假设 4.3 中的常数，并定义截断值 $\bar{\eta}(x,\varepsilon) := [\eta(x,\varepsilon)]_0^{z_0}$。根据假设 4.3，存在 $z_1 \in (0, z_0]$，使得 $f(z_1) > 0$。选择足够小的 ε，使 $\min\{\eta_{\min}, z_0\} \ge z_1$；等价地，对于所有满足 $\xi_x \ge \xi_0$的 x，$\bar{\eta}(x,\varepsilon) \ge z_1$。根据单调性，

$$c_x(\varepsilon) = \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T \le \eta(x,\varepsilon)\mu_x] \ \ge\ \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}[T \le \bar{\eta}(x,\varepsilon)\mu_x] \ \ge\ c_-f(\bar{\eta}(x,\varepsilon)). \tag{79}$$

由于 f 是非递减的，且满足倍增条件 $f(z_1) > 0$，因此对于所有 $z \in (0, z_0]$，都有 $f(z) > 0$；于是$f(\bar{\eta}(x,\varepsilon)) \ge f(\min\{\eta_{\min}, z_0\}) =: c_f > 0$，因此 $c_x(\varepsilon) \ge c_-c_f = \Omega(1)$。接着，假设 4.3 对这类 x一致成立，

$$c_x(\varepsilon) \ge \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau) \le \bar{\eta}(x,\varepsilon)\,\mu_x\right] \ge c_-\,f(\bar{\eta}(x,\varepsilon)) \ge c_-\,f(z_1) =: c_\star > 0. \tag{80}$$

因此，对于任意满足 $\xi \ge \xi_0$ 的切片 $\mathcal{X}(\xi)$，我们有

$$c_0(\xi) = \inf_{x\in\mathcal{X}(\xi)} c_x(\varepsilon) \ \ge\ c_\star, \tag{81}$$

而显然 $c_0(\xi) \le 1$。因此 $c_0(\xi) = \Theta(1)$。 □

D.7. 伽马分布模型下的标准差

引理 D.5（伽马分布模型下的标准差。）。令 γ 表示正则化下不完全伽马函数。定义

$$\begin{aligned}\big(C_{\mathrm{RL}}(K,\xi)\big)^2 :=\ & (K\xi)^2\gamma(K,K\xi) - 2K^2\xi\,\gamma(K{+}1,K\xi)\\ & + K(K{+}1)\,\gamma(K{+}2,K\xi)\\ & - \left(K\xi\,\gamma(K,K\xi) - K\,\gamma(K{+}1,K\xi)\right)^2.\end{aligned}$$

那么，*Gamma* 模型下累积奖励的标准差满足

$$\sigma_b(K,\xi) = \frac{C_{\mathrm{RL}}(K,\xi)}{p} = \frac{C_{\mathrm{RL}}(K,\xi)}{K\xi}\,H. \tag{82}$$

证明。为简便起见，在 Gamma 模型分析（第 5.3 节和附录 A.6）中，我们将步数 T 视为连续随机变量，并采用双层塑形奖励 $R = \text{最大值}\{0, H - T\}$，它与方程（6）相差一个加性常数（即“+1”），适用于成功轨迹。这种简化不会影响我们区间分析中的渐近尺度关系，因此我们可以得出相同的阶数结论。

定义 $C_{\mathrm{RL}}(K, \xi)$ 为

$$\big(C_{\mathrm{RL}}(K,\xi)\big)^2 := p^2 \mathbb{V}[R(\tau)] \tag{83}$$

要找到该量的表达式，我们必须先计算 $R(\tau)$ 的方差，其计算公式为$\mathbb{V}[R(\tau)] = \mathbb{E}[R(\tau)^2] - (\mathbb{E}[R(\tau)])^2$。接下来，我们计算 $R(\tau)$ 的前两个原始矩。在本分析中，对于 $T(\tau)$ 的离散负二项分布，我们采用连续伽马分布 $T(\tau) \sim$ 伽马分布(K, p) 进行近似；当成功概率 p 较小时，这种近似是有效的。其概率密度函数为

$$f(t; K, p) = \frac{p^K}{\Gamma(K)} t^{K-1} e^{-pt}. \tag{84}$$

首先，我们计算期望奖励 $\mathbb{E}[R(\tau)]$。

$$\begin{aligned}
\mathbb{E}[R(\tau)] &= \mathbb{E}[\max(0, H - T(\tau))] \\
&= \int_0^H (H - t) f(t; K, p)\, dt \\
&= H \int_0^H f(t; K, p)\, dt - \int_0^H t f(t; K, p)\, dt
\end{aligned} \tag{85}$$

第一个积分表示在 H 处取值的伽马分布累积分布函数，即正则化下不完全伽马函数 $\gamma(K, pH)$。第二个积分对应于截断至 H 的伽马分布的一阶矩。利用引理 A.1 的结果，该积分的值为 $\frac{K}{p}\gamma(K+1, pH)$。因此，期望奖励为：

$$\mathbb{E}[R(\tau)] = H \cdot \gamma(K, pH) - \frac{K}{p}\gamma(K+1, pH) \tag{86}$$

接下来，我们计算奖励的二阶原点矩 $\mathbb{E}[R(\tau)^2]$。

$$\begin{aligned}
\mathbb{E}[R(\tau)^2] &= \mathbb{E}[(\max(0, H - T(\tau)))^2] \\
&= \int_0^H (H - t)^2 f(t; K, p)\, dt \\
&= \int_0^H (H^2 - 2Ht + t^2) f(t; K, p)\, dt \\
&= H^2 \int_0^H f(t)\, dt - 2H \int_0^H t f(t)\, dt + \int_0^H t^2 f(t)\, dt
\end{aligned} \tag{87}$$

前两项与为 $\mathbb{E}[R(\tau)]$ 计算的矩直接相关。第三项是截断 Gamma 分布的二阶矩，其计算结果为 $\frac{K(K+1)}{p^2}\gamma(K+2, pH)$。代入这些结果，我们得到：

$$\mathbb{E}[R(\tau)^2] = H^2\gamma(K, pH) - 2H\frac{K}{p}\gamma(K+1, pH) + \frac{K(K+1)}{p^2}\gamma(K+2, pH) \tag{88}$$

现在我们可以计算 $\big(C_{\mathrm{RL}}(K, \xi)\big)^2 = p^2\left(\mathbb{E}[R(\tau)^2] - (\mathbb{E}[R(\tau)])^2\right)$。

$$\begin{aligned}
\big(C_{\mathrm{RL}}(K,\xi)\big)^2 &= p^2\left[H^2\gamma(K, pH) - \frac{2HK}{p}\gamma(K+1, pH) + \frac{K(K+1)}{p^2}\gamma(K+2, pH)\right] \\
&\quad - p^2\left[H\gamma(K, pH) - \frac{K}{p}\gamma(K+1, pH)\right]^2 \\
&= [(pH)^2\gamma(K, pH) - 2pHK\gamma(K+1, pH) + K(K+1)\gamma(K+2, pH)] \\
&\quad - [pH\gamma(K, pH) - K\gamma(K+1, pH)]^2
\end{aligned}$$

最后，我们引入相对预算 $\xi := \frac{pH}{K}$，这意味着 $pH = K\xi$。将其代入上面的表达式，即可得到论文公式（23）中所述的最终结果：

$$\begin{aligned}\big(C_{\mathrm{RL}}(K,\xi)\big)^2 &= [(K\xi)^2\gamma(K,K\xi) - 2K^2\xi\gamma(K+1,K\xi) + K(K+1)\gamma(K+2,K\xi)] \\ &\quad - [K\xi\gamma(K,K\xi) - K\gamma(K+1,K\xi)]^2 \end{aligned} \tag{89}$$

□

图 11 展示了 ξ 与 $C_{\mathrm{RL}}(K,\ \xi)$ 在每个 $K \in \{1,2,5,10\}$ 下的关系。注意到 $C_{\mathrm{RL}}(K,\ \xi)$ 在 $\sqrt{K}$ 时收敛到 $\xi \to \infty$。

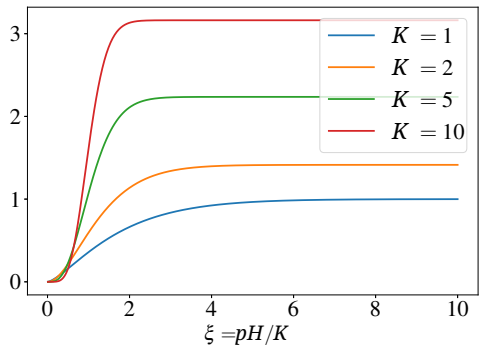


图 *11.* ξ 与 $C_{\mathrm{RL}}(K,\ \xi)$之间的关系。

D.8. Gamma模型下的反集中

引理D.6 (Gamma 模型下的反集中。)。 定义 $\psi(K、\ \xi、\ \varepsilon)$: $= 1-\gamma(K、\ K\xi)+\frac{1}{\xi}\gamma(K+1、\ K\xi)-\frac{C_{\mathrm{RL}}(K,\ \xi)}{K\xi}\ -\ \sqrt{\varepsilon}$。在伽马分布模型下，反集中系数表示为

$$c(K,\xi,\varepsilon) = \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau) \le H\cdot[\psi(K,\xi,\varepsilon)]_0^1\right] \tag{90}$$

$$= \gamma\big(K, K\xi\cdot[\psi(K,\xi,\varepsilon)]_0^1\big). \tag{91}$$

证明。 在伽马分布模型下，对于 $c(K,\ \xi,\ \varepsilon)$，我们有以下一系列方程：

$$c(K,\xi,\varepsilon) = \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau) \le H\left(1 - \frac{\mathbb{E}[R(\tau)]}{H} - \frac{\sigma_b^{\mathrm{RL}}\sqrt{\varepsilon}}{H}\right)\right] \tag{92}$$

$$= \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\left[T(\tau) \le H\left(1 - \big(\gamma(K,K\xi) - \frac{1}{\xi}\gamma(K+1,K\xi)\big) - \frac{C_{\mathrm{RL}}(K,\xi)}{K\xi}\sqrt{\varepsilon}\right)\right]. \tag{93}$$

定义 ψ 使得

$$\psi(K,\xi,\varepsilon) := 1 - \Big(\gamma(K,K\xi) - \frac{1}{\xi}\gamma(K+1,K\xi)\Big) - \frac{C_{\mathrm{RL}}(K,\xi)}{K\xi}\sqrt{\varepsilon}, \tag{94}$$

其中 $[\mathbf{x}]_0^1 := \min\{1,\max\{0,\mathbf{x}\}\}$ 表示截断。由此可简化为：

$$c(K,\xi,\varepsilon) = \mathbb{P}_{\tau\sim\pi_b(\cdot|x)}\mathbb{P}\left[T(\tau) \le H\cdot[\psi(K,\xi,\varepsilon)]_0^1\right] \tag{95}$$

$$= \gamma\big(K, K\xi\cdot[\psi(K,\xi,\varepsilon)]_0^1\big). \tag{96}$$

□

D.9. 最优相对预算

引理 D.7（最优 $\xi^\star$）。对于任意固定的 K，标准差 $\sigma_b(K,\xi) = \frac{C_{\mathrm{RL}}(K,\xi)}{K\xi} H$ 存在唯一的全局最大值点 $\xi^\star(K) \in (0, \infty)$，其特征由一阶条件给出

$$\xi\, C'_{\mathrm{RL}}(K,\xi) = C_{\mathrm{RL}}(K,\xi). \tag{97}$$

证明。定义

$$g(\xi) = \log\left(\frac{C_{\mathrm{RL}}(K,\xi)}{K\xi}\, H\right) = \log C_{\mathrm{RL}}(K,\xi) + \log H - \log(K\xi). \tag{98}$$

求导得

$$g'(\xi) = \frac{C'_{\mathrm{RL}}(K,\xi)}{C_{\mathrm{RL}}(K,\xi)} - \frac{1}{\xi}. \tag{99}$$

驻点满足

$$\xi C'_{\mathrm{RL}}(K,\xi) = C_{\mathrm{RL}}(K,\xi). \tag{100}$$

利用对数方差函数的凸性可知，该驻点是唯一的，并且对应于全局最大值点。 □

E. 第 6 节的证明

E.1. 定理 6.2 的证明

证明。 根据 Setlur等人 (2025年 的引理 A.19，在当前策略 χ^2周围半径为 κ_i 的信赖域内，最佳比较策略 $\pi^{(i)}$ 满足

$$J_{\xi_i}(\bar{\pi}_\kappa) = J_{\xi_i}\left(\pi^{(i)}\right) + \sqrt{\kappa_i}\,\sigma(\pi^{(i)}). \tag{101}$$

使用来自以下策略的 n_i 次推演 $\pi^{(i)}$，以至少 $1-\delta$ 的概率，基于验证器的强化学习估计量满足有限样本界限 d

$$J_{\xi_i}(\bar{\pi}_\kappa) - J_{\xi_i}\left(\hat{\pi}^{(i+1)}\right) \le \frac{CH\log(|\mathcal{R}|/\delta)}{c_0(\xi_i;\kappa_i)\, n_i}, \tag{102}$$

其中 $c_0(\xi_i;\ \kappa_i)$ 是相对于策略 $\pi^{(i)}$ 的反集中系数，且 C 是定理 3.2 所蕴含的普适常数。结合式（101）和式（102）可得

$$J_{\xi_i}\left(\hat{\pi}^{(i+1)}\right) \ge J_{\xi_i}\left(\pi^{(i)}\right) + \sqrt{\kappa_i}\,\sigma(\pi^{(i)}) - \frac{CH\log(|\mathcal{R}|/\delta)}{c_0(\xi_i;\kappa_i)\, n_i}. \tag{103}$$

若 （1 8）也成立，则上述方程中的最后一项至多为 $\frac{1}{2}\sqrt{\kappa_i}\,\sigma$（$\pi^{(i)}$），- 这意味着 es

$$J_{\xi_i}\left(\pi^{(i+1)}\right) = J_{\xi_i}\left(\hat{\pi}^{(i+1)}\right) \ge J_{\xi_i}\left(\pi^{(i)}\right) + \frac{1}{2}\sqrt{\kappa_i}\,\sigma(\pi^{(i)}). \tag{104}$$

□

E.2. 定理6.3的证明

基于定理 6.2中确立的有限样本保证，我们现在分析第 5 节所引入的三个相对预算区间内的具体学习动态。通过代入各区间中特定的奖励标准差和反集中系数标度律，我们得出确保单调策略改进所需的最小样本复杂度 n_i。以下分析说明相对预算 ξ 如何影响在线强化学习的效率；结合以下命题即可得出定理 6.3。

命题E.1 (不足相对预算)。 当相对预算不足（即 $\xi \to 0$）时，可实现的每次迭代改进受限于 $\sqrt{\kappa}\,\sigma(\pi) = \mathcal{O}(H\sqrt{\kappa f(\xi)})$。所需的充分样本量为：

$$n_i = \widetilde{\Omega}\left(\frac{1}{\sqrt{\kappa_i}\,(f(\xi_i))^{3/2}}\right). \tag{105}$$

在此区间内，策略改进所需的最小样本量会随着 $f(\xi)^{-3/2}$ 快速发散。这一结果将特定模型（例如满足 $f(\xi) \sim \xi^K$、从而得到 $n_i \sim \xi_i^{-3K/2}$ 的伽马分布）中的直觉推广开来。它凸显了相对预算不足区间的根本难点：随着随机成功的概率$f(\xi)$趋于零，无论底层具体分布如何，获取足够数量的信息性轨迹的成本都会高得令人望而却步。

证明。 在相对预算不足区间（满足 $\xi_i \to 0$）中，我们使用一般左尾概率函数 $f(\xi_i)$ 分析其行为，该函数定义于定理 4.3中。根据定理 D.2，成功解轨迹十分罕见，奖励的标准差按以下方式变化：

$$\sigma(\pi^{(i)}) = \mathcal{O}\big(H\sqrt{f(\xi_i)}\big). \tag{106}$$

第 i 次迭代中策略改进所需的最少样本数满足一般界：

$$n_i \gtrsim \frac{H\log(|\mathcal{R}|/\delta)}{c_0(\xi_i;\kappa_i)\,\sqrt{\kappa_i}\,\sigma(\pi^{(i)})}. \tag{107}$$

当 ξ较小时，定理 D.2证明反集中系数的变化规律为

$$c_0(\xi_i;\kappa_i) = \Theta\big(f(\xi_i)\big). \tag{108}$$

结合 $\sigma(\pi^{(i)})$ 的渐近变化规律，样本复杂度下界为：

$$n_i = \widetilde{\Omega}\left(\frac{1}{\sqrt{\kappa_i}\,(f(\xi_i))^{3/2}}\right). \tag{109}$$

□

命题 E.2 (均衡相对预算)。 在均衡相对预算状态下，满足 $\xi = \Theta(1)$，可达到的每次迭代改进为 $\mathcal{O}(\sqrt{\kappa}H)$。此外，样本复杂度为：

$$n_i = \widetilde{\Omega}(1/\sqrt{\kappa_i}). \tag{110}$$

这一结果表明，从样本复杂度的角度看，均衡相对预算状态是学习的最佳区间。在此状态下，奖励方差达到最大值，提供强有力的学习信号；同时，反集中系数保持稳定，确保高奖励轨迹能够被频繁采样。因此，无论时域多长，样本复杂度都保持为常数，从而能够高效地改进策略。

证明。 现在我们分析中间阶段，其定义为相对预算为 $\xi_i = \Theta(1)$ 的状态。在此状态下，计算预算 H与预期求解时间 $\mathbb{E}[T(\tau)]$,相当，这意味着成功事件和失败事件发生的概率均为 $\Theta(1)$。

根据定理 D.3，在均衡相对预算状态下，截断 H 不会改变矩的量级。因此，奖励标准差和反集中系数满足：

$$\sigma(\pi^{(i)}) = \Theta(H) \quad \text{and} \quad c_0(\xi_i;\kappa_i) = \Theta(1). \tag{111}$$

利用这些标度律，我们分析学习动态。在$\sqrt{\kappa_i}\,\sigma$（$\pi^{(i)}$）的情况下，可实现的每次迭代改进与其成正比。由于在这一状态下奖励标准差达到最大值（随 H 线性增长，而不像早期或晚期阶段那样衰减），我们预计策略改进的幅度最大。实现策略改进所需的最小样本数 n_i 由预算与有效学习信号之比决定。代入定理D.3中的标度关系可得：

$$n_i = \widetilde{\Omega}(1/\sqrt{\kappa_i}). \tag{112}$$

□

命题 E.3（充足的相对预算）。 在相对预算充足且 $\xi \to \infty$ 的情况下，可实现的每次迭代改进随 ξ 线性衰减；即，$\sqrt{\kappa}H/\xi$。此外，样本复杂度下界变为：

$$n_i = \widetilde{\Omega}\left(\xi_i/\sqrt{\kappa_i}\right). \tag{113}$$

关键在于，要确保单调改进，通常需要将信赖域按 $\sqrt{\kappa_i} = \Theta(\xi_i^{-1})$进行缩放。这会使有效改进率为 $\mathcal{O}(H/\xi_i^2)$，并使样本复杂度出现二次退化： $n_i = \widetilde{\Omega}(\xi_i^2)$。

证明。我们接着分析该情形 $\xi_i \to \infty$。根据定理D.4，在这一充足相对预算情形下，成功概率趋近于1（$q_x \to 1$），且失败事件变得可忽略。因此，截断效应消失，奖励统计量收敛到无条件分布的相应统计量。

在一般矩 假设（定理4.2）下，标准差与性能差距按如下方式缩放：

$$\sigma(\pi^{(i)}) = \mathcal{O}\left(H/\xi_i\right). \tag{114}$$

因此，方差和次优差距均按 $\mathcal{O}(H/\xi_i)$衰减。

关于反集中系数，定理D.4 指出，在此情形下， $c_0(\xi_i;\kappa_i)$ 不会趋于零，而是始终有一个正的下界：

$$c_0(\xi_i;\kappa_i) = \Theta(1). \tag{115}$$

这与特定的伽马分布情形形成对比：在某些参数选择下，反集中系数可能趋于零；而一般理论则保证，只要妥善管理信赖域，学习信号就会持续存在。

在迭代 i 时，实现策略改进所需的最小样本数由以下一般界确定：

$$n_i \gtrsim \frac{H\log(|\mathcal{R}|/\delta)}{c_0(\xi_i;\kappa_i)\sqrt{\kappa_i}\,\sigma(\pi^{(i)})}. \tag{116}$$

代入一般渐近关系 $\sigma(\pi^{(i)}) = \mathcal{O}(H/\xi_i)$ 和 $c_0(\xi_i;\kappa_i) = \Theta(1)$，可得：

$$n_i = \Omega\left(\frac{H\log(|\mathcal{R}|/\delta)}{1\cdot\sqrt{\kappa_i}\,(H/\xi_i)}\right) = \Omega\left(\frac{\xi_i}{\sqrt{\kappa_i}}\log(|\mathcal{R}|/\delta)\right). \tag{117}$$

因此，所需样本量随 ξ 线性增长。同时，可实现的改进幅度按 $\Theta(H/\xi)$ 缩小，这表明在后期阶段，进展所需的样本复杂度会越来越高；这一结论广泛适用于特定概率模型之外的情形。

□

E.3. 定理6.4 的证明

回顾一下，对于每个 $i \in \{0,1,\ldots,\}$,

$$c_0^{(i)} := \mathbb{P}_{x\sim\rho,\tau\sim\pi^{(i)}(\cdot|x)}\left[R(\tau) \geq J(\pi^{(i)}) + \sqrt{\kappa_i}\,\sigma(\pi^{(i)})\right]. \tag{118}$$

表示策略 $\pi^{(i)}$ 的反集中系数，其信赖域半径为

$$\kappa_i = \left(\min\left\{\frac{\sigma(\pi^{(i)})}{J(\pi^{(i)})}, \frac{H - J(\pi^{(i)})}{2\,\sigma(\pi^{(i)})}\right\}\right)^2, \tag{119}$$

这确保在伽马分布模型下 $c_0^{(i)} > 0$。由于论证与定理 6.2 类似，并根据并集界，以至少 $1-\delta$ 的概率有

$$J(\pi^{(i+1)}) \geq J(\pi^{(i)}) + \frac{1}{2}\sqrt{\kappa_i}\,\sigma(\pi^{(i)}) \qquad \forall i \in \{0,1,\ldots,m-1\}, \tag{120}$$

当从每个 $\pi^{(i)}$ 中使用足够多的轨迹采样，使得

$$n_i \geq \frac{2\,CH\log\left(\delta^{-1}m\,|\mathcal{R}|\right)}{\sqrt{\kappa_i}\,c_0^{(i)}\sigma(\pi^{(i)})}. \tag{121}$$

在伽马模型下，第D.7 节中的矩计算表明

$$J(\pi^{(i)}) = H\left(\gamma\left(K, K\xi_i\right) - \frac{1}{\xi_i}\gamma\left(K+1, K\xi_i\right)\right), \tag{122}$$

$$\begin{aligned}\sigma^2(\pi^{(i)}) = H^2\Bigg(&\gamma\left(K, K\xi_i\right) - \frac{2}{\xi_i}\gamma\left(K+1, K\xi_i\right) + \frac{1+K^{-1}}{\xi_i^2}\gamma\left(K+2, K\xi_i\right)\\ &- \left(\gamma\left(K, K\xi_i\right) - \frac{1}{\xi_i}\gamma\left(K+1, K\xi_i\right)\right)^2\Bigg),\end{aligned} \tag{123}$$

其中，$\gamma(s,\ \lambda)$表示正则化下不完全伽马函数:

$$\gamma(s,\lambda) \coloneqq \frac{\int_0^\lambda x^{s-1}e^{-x}\,\mathrm{d}x}{\int_0^\infty x^{s-1}e^{-x}\,\mathrm{d}x} \qquad\qquad \forall s,\lambda \in \mathbb{R}_{>0}. \tag{124}$$

根据方程（118）和（119），以及 $0 < J\left(\pi^{(0)}\right) < J\left(\pi^{(1)}\right) < \cdots \leq H$这一事实，我们可知极限$_{i\to\infty}\,\sigma^2\left(\pi^{(i)}\right) = 0$，极限$_{i\to\infty}\,J\left(\pi^{(i)}\right) = H$，因此极限$_{i\to\infty}\,\xi_i = \infty$。在极限情况下，有如下渐近展开:

$$J(\pi^{(i)}) = H\left(1 - \frac{1}{\xi_i} + O_K\left(\xi_i^{K-1}e^{-K\xi_i}\right)\right); \tag{125}$$

$$\sigma^2(\pi^{(i)}) = H^2\left(\frac{1}{\xi_i^2 K} + O_K\left(\xi_i^{K-1}e^{-K\xi_i}\right)\right). \tag{126}$$

还存在一个常数 $\tau(K) \in \{0,1,\ldots\}$，它取决于 K，并满足 $\xi_{\tau(K)} \geq 2\left(1+K^{-1}\right)$。当 $i \geq \tau(K)$ 时，有

$$\frac{H - J(\pi^{(i)})}{2\,\sigma(\pi^{(i)})} - \frac{\sigma(\pi^{(i)})}{J(\pi^{(i)})} = \frac{1}{2\xi_i\sigma(\pi^{(i)})J(\pi^{(i)})}\left(\gamma\left(K+1, K\xi_i\right) - \frac{2\left(1+K^{-1}\right)}{\xi_i}\gamma\left(K+2, K\xi_i\right)\right) > 0, \tag{127}$$

这意味着 $\kappa_i = \left(\frac{\sigma(\pi^{(i)})}{J(\pi^{(i)})}\right)^2$。将这些代入方程 (120)，对每个 $i \in \{\tau(K)\ \tau(K)+1,\ \ldots\ldots,\ m-1\}$次 at

$$1 - \frac{1}{\xi_{i+1}} + O_K\left(\xi_{i+1}^{K-1}e^{-K\xi_{i+1}}\right) \geq 1 - \frac{1}{\xi_i} + \frac{1}{2K}\frac{1}{\xi_i^2} + O_K\left(\xi_i^{K-1}e^{-K\xi_i}\right), \tag{128}$$

这意味着，由于 $\xi_i < \xi_{i+1}$,

$$\frac{1}{\xi_{i+1}} \leq \frac{1}{\xi_i} - \frac{1}{2K}\frac{1}{\xi_i^2} + O_K\left(\xi_i^{K-1}e^{-K\xi_i}\right), \tag{129}$$

因此

$$\xi_{i+1} \geq \xi_i\left(1 - \frac{1}{2K}\frac{1}{\xi_i} + O_K\left(\xi_i^K e^{-K\xi_i}\right)\right)^{-1} = \xi_i + \frac{1}{2K} - O_K\left(\xi_i^{K+1}e^{-K\xi_i}\right). \tag{130}$$

由于 $\lim_{i\to\infty}\xi_i = \infty$，对每个 $i \in \{0,1,\ldots,m-1\}$，可得

$$\xi_i - \xi_0 \geq \frac{i}{2K} - O_K(1). \tag{131}$$

最后，只需证明 $c_0^{(i)} = \Omega_K(1)$。再次利用方程（125）和（126），可得

$$H - J(\pi^{(i)}) - \frac{\sigma^2(\pi^{(i)})}{J(\pi^{(i)})} = \frac{H}{\xi_i}\left(1 - O_K\left(\frac{1}{\xi_i}\right)\right). \tag{132}$$

结合伽马函数的累积分布函数，当 $i \geq \tau(K)$ 时，有

$$c_0^{(i)} = \gamma\left(K, \frac{\xi_i K}{H}\left(H - J(\pi^{(i)}) - \frac{\sigma^2(\pi^{(i)})}{J(\pi^{(i)})}\right)\right) = \gamma\left(K, K - O_K\left(\frac{1}{\xi_i}\right)\right), \tag{133}$$

这就完成了定理6.4的证明。

F. 监督微调的理论分析

VF方法（例如监督微调）依赖于模仿专家轨迹。其性能从根本上受解空间异质性的限制。直观而言，如果存在许多长度和风格各异、都能得出正确答案的有效路径，那么在数据有限的情况下，模型难以确定该模仿哪条路径。我们通过策略异质性对其进行量化，该异质性定义为沿轨迹的Q值累积方差。 $Q_\pi(s_h, a_h)$表示给定大语言模型 π在测试问题上的期望累积奖励，定义为

$$Q_\pi(s_h, a_h) \coloneqq \mathbb{E}_{\rho,\pi}\left[\sum_{t=h}^{H} r(s_t, a_t) \mid s_h, a_h\right]. \tag{134}$$

定义 F.1 （策略异质性）。 对于问题 $x \in \mathcal{X}$ 和策略 $\pi \in \Pi$，异质性定义为:

$$\sigma_{\pi,x}^2 \coloneqq \sum_{h=1}^{H} \mathbb{E}_{s_h \sim d_h^\pi}\left[\mathbb{V}_{a\sim\pi(\cdot|s_h)}\left[Q_{\pi_e}(s_h, a_h)\right]\right]. \tag{135}$$

对输入分布 ρ进行汇总可得

$$\sigma_\pi^2 \coloneqq \mathbb{E}_{x\sim\rho}[\sigma_{\pi,x}^2], \quad \tilde{\sigma}_\pi \coloneqq \mathrm{Median}_x(\sigma_{\pi,x}). \tag{136}$$

根据上述定义，我们给出监督微调的次优差距，其尺度如下 s:

引理 F.2 (监督微调的信息论下界，Setlur等人（2025年）中的定理 5.4）。 固定一个分布 ρ、一个奖励类 $\mathcal{R}$、一个基础策略 π_b、一个专家 π_e以及任意整数 $\nu \leq |\mathcal{X}|/4$。存在一个专家族 Π'，满足 $|\Pi'| = 2^\nu$，以及一个奖励子类 $\mathcal{R}' \subset \mathcal{R}$，使得在$\hat{\pi}_{\bar{n}}^{\mathrm{SFT}}$上训练的每个*VF*学习器 $\bar{n}$ 都必须满足

$$\max_{\pi'\in\Pi'} \max_{r'\in\mathcal{R}'} \left[J_{r'}(\pi') - J_{r'}(\hat{\pi}_{\bar{n}}^{\mathrm{SFT}})\right] = \Omega\left(\sigma_b \sqrt{\frac{\log|\Pi'|}{\bar{n}}}\right),$$

其中 $\sigma_b := \sigma_{\pi_b}$ 是定理*F.1*中定义的基础策略异质性参数。

尽管严格的下界取决于专家的具体异质性 $\tilde{\sigma}_e$，我们仍以基础策略的异质性 σ_b来表示该下界。由于专家策略被限制在基础策略的 χ^2邻域内，因此这种替换是成立的，这意味着 $\sigma_e \approx \sigma_b$ （Setlur等人（2025年）中的引理 5.3）。

F.1. 相对预算下监督微调的次优差距表征

我们使用相对预算来表征监督微调的次优性 ξ。由于监督微调数据集仅包含成功解轨迹，这种异质性概念可简化为给定成功条件下的奖励分布方差。现在，我们为每个问题定义这一异质性参数 x。根据（6），给定 x 时的方差为:

$$\mathbb{V}[R(\tau) \mid T(\tau) \leq H, x] = \mathbb{V}[T(\tau) \mid T(\tau) \leq H, x]. \tag{137}$$

现在，我们引入缩放因子 $C_{\mathrm{SFT}}(\xi_x) \in [0,1]$，以关联问题 x 的截断方差与未截断方差:

$$\mathbb{V}[T(\tau) \mid T(\tau) \leq H, x] = \mathbb{V}[T(\tau) \mid x] \cdot \left(C_{\mathrm{SFT}}(\xi_x)\right)^2. \tag{138}$$

其中 ξ_x 是定理 4.1中的问题相关临界性。该缩放因子 $C_{\mathrm{SFT}}(\xi_x)$ 具有以下定性性质。首先，当 $\xi_x \to 0$ (即 $H \to 0$ 在 μ_x固定时)，成功轨迹的方差 $\mathbb{V}[T \mid T \leq H, x]$ 趋近于

0；因此，$\lim_{\xi_x\to 0} C_{\mathrm{SFT}}(\xi_x) = 0$。此外，当 $\xi_x \to \infty$ (即 $H \to \infty$) 时，对于问题 $T \le H$，截断条件 x不再起作用。根据定理 4.2， $P(T < \infty \mid x) = 1$。截断方差趋近于全方差 $\mathbb{V}[T(\tau) \mid x]$；因此，$\lim_{\xi_x\to\infty} C_{\mathrm{SFT}}(\xi_x) = 1$。

因此，问题 x 的标准差按以下方式缩放

$$\sigma^{\mathrm{SFT}}_{b,x}(\xi_x) \coloneqq \sqrt{\mathbb{V}\big[T(\tau) \mid T(\tau) \le H,\, x\big]} = \Theta\big(\mu_x \cdot C_{\mathrm{SFT}}(\xi_x)\big) = \Theta\left(\frac{C_{\mathrm{SFT}}(\xi_x)}{\xi_x}\, H\right), \tag{139}$$

其中我们使用了 $\xi_x = H/\mu_x$ 定理 4.1中的结果。根据 Setlur等人 (2025年)，监督微调的整体次优性取决于这些逐问题异质性的稳健聚合值；我们采用中位数 $\tilde{\sigma}^{\mathrm{SFT}}_b$: = 中位数$_x\{\sigma^{\mathrm{SFT}}_{b,x}(\xi_x)\}$。

定理F.3 （具有问题相关临界性的监督微调次优差距）。 令 $\bar{n}$ 表示收集到的成功解答轨迹总数。

$$J_r(\bar{\pi}_\kappa) - J_r(\hat{\pi}^{\mathrm{SFT}}_{\bar{n}}) = \Omega\left(\frac{\tilde{\sigma}^{\mathrm{SFT}}_b}{\sqrt{\bar{n}}}\right), \tag{140}$$

其中 $\tilde{\sigma}^{\mathrm{SFT}}_b$ = 中位数$_x\big\{\Theta\big(H\, C_{\mathrm{SFT}}(\xi_x)/\xi_x\big)\big\}$。

该差距仍按 $1/\sqrt{\bar{n}}$缩放，这意味着随着专家数据增多，改进仍然缓慢。对于固定的问题 x，因子$C_{\mathrm{SFT}}(\xi_x)/\xi_x$ 通常在中间的 $\xi_x = \Theta(1)$ 处达到最大值，这表明在成功轨迹异质性较高时，模仿尤其困难。整体（中位数）差距由处于这一中间区域的问题主导。

F.2. 伽马分布的具体分析

与强化学习的分析类似，我们还分析了一种特定情形：临界性在各实例之间保持不变，且正确解的求解时间可用伽马分布建模。截断负二项分布的方差没有简单的闭式表达式。然而，在小-p 情形下，负二项分布可用伽马分布很好地近似，从而可以得到其截断矩的简洁闭式表达式。这里，方差可通过缩放因子 $\big(C_{\mathrm{SFT}}(K,\ \xi)\big)^2$ 与未截断方差 ($\mathbb{V}[T(\tau)] \approx K/p^2$) 相关联，该缩放因子定义如下：

引理F.4。 设 $\Gamma(\mathbf{x})$: $= \int_0^\infty y^{\mathbf{x}-1}e^{-y}dy$ 为伽马函数。利用正则化下不完全伽马函数$\gamma(\mathbf{x}_1, \mathbf{x}_2)$: $= \Gamma^{-1}(\mathbf{x}_1) \int_0^{\mathbf{x}_2} y^{\mathbf{x}_1-1}e^{-y}dy$, ~~定义~~缩放因子 C_{SFT}： $\mathbb{Z}_+ \times \mathbb{R}_+ \to [0,1]$ 如下

$$\big(C_{\mathrm{SFT}}(K,\xi)\big)^2 = (K+1)\frac{\gamma(K+2, K\xi)}{\gamma(K, K\xi)} - K\left(\frac{\gamma(K+1, K\xi)}{\gamma(K, K\xi)}\right)^2. \tag{141}$$

于是，下列方程成立：

$$\mathbb{V}[T(\tau) \mid T(\tau) \le H] = \mathbb{V}[T(\tau)] \cdot \big(C_{\mathrm{SFT}}(K,\xi)\big)^2 \approx \frac{K}{p^2} \cdot \big(C_{\mathrm{SFT}}(K,\xi)\big)^2. \tag{142}$$

证明。 设 X 为服从形状参数 K 、率参数 p的伽马分布的随机变量，记为 $X \sim$ 伽马分布(K,p)。我们旨在求其在事件 $X \le H$条件下的方差，其中 H 为令牌预算。首先，任意随机变量 X 在 H 处上截断后的方差由如下标准公式给出：

$$\mathbb{V}[X|X \le H] = \mathbb{E}[X^2|X \le H] - (\mathbb{E}[X|X \le H])^2 \tag{143}$$

此处，我们的任务归结为求截断 Gamma 分布的一阶矩与二阶矩。

令 $X \sim$ 伽马分布(K,p)。将 X 截断于 H 时，其第 k 阶矩定义为：

$$\mathbb{E}[X^k|X \le H] = \frac{\int_0^H x^k f(x)\,dx}{\int_0^H f(x)\,dx} \tag{144}$$

其中 $f(x)$ 是伽马分布的概率密度函数:

$$f(x;K,p)=\frac{p^K}{\Gamma(K)}x^{K-1}e^{-px} \tag{145}$$

分母 $\int_0^H f(x)\,dx$ 是在 H 处取值的累积分布函数（CDF）。对于伽马分布，这就是正则化下不完全伽马函数 $\gamma(K,pH)$。对于任意矩 k，我们计算分子中的积分:

$$\begin{aligned}\int_0^H x^k f(x)\,dx &= \int_0^H x^k \frac{p^K}{\Gamma(K)}x^{K-1}e^{-px}\,dx\\ &= \frac{p^K}{\Gamma(K)}\int_0^H x^{K+k-1}e^{-px}\,dx\end{aligned}$$

经过变量替换（$t=px$）后，该积分可以用正则化下不完全伽马函数 $\gamma(s,z)$ 表示。根据定义，积分 $\int_0^z t^{s-1}e^{-t}\,dt$ 等于 $\Gamma(s)\gamma(s,z)$，因此我们有:

$$\begin{aligned}\frac{p^K}{\Gamma(K)}\int_0^H x^{K+k-1}e^{-px}\,dx &= \frac{p^K}{\Gamma(K)}\int_0^{pH}\left(\frac{t}{p}\right)^{K+k-1}e^{-t}\frac{dt}{p}\\ &= \frac{1}{p^k\Gamma(K)}\int_0^{pH}t^{K+k-1}e^{-t}\,dt\\ &= \frac{1}{p^k\Gamma(K)}\left[\Gamma(K+k)\gamma(K+k,pH)\right]\\ &= \frac{\Gamma(K+k)}{p^k\Gamma(K)}\gamma(K+k,pH)\end{aligned}$$

X 在 H 处截断后的上截断一阶矩为:

$$\mathbb{E}[X|X\le H]=\frac{\frac{\Gamma(K+1)\gamma(K+1,pH)}{p\Gamma(K)}}{\gamma(K,pH)}=\frac{K}{p}\frac{\gamma(K+1,pH)}{\gamma(K,pH)}. \tag{146}$$

X 在 H 处截断后的上截断二阶矩为:

$$\mathbb{E}[X^2|X\le H]=\frac{\frac{\Gamma(K+2)\gamma(K+2,pH)}{p^2\Gamma(K)}}{\gamma(K,pH)}=\frac{K(K+1)}{p^2}\frac{\gamma(K+2,pH)}{\gamma(K,pH)}. \tag{147}$$

将各阶矩代回方差公式:

$$\mathbb{V}[X|X\le H]=\frac{K(K+1)}{p^2}\frac{\gamma(K+2,pH)}{\gamma(K,pH)}-\left(\frac{K}{p}\frac{\gamma(K+1,pH)}{\gamma(K,pH)}\right)^2. \tag{148}$$

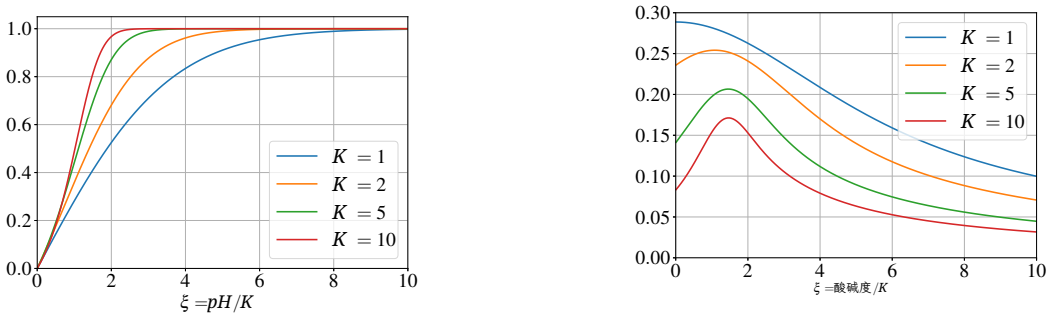
缩放因　　子是截断方差与未截断方差的比值，其中 $\mathbb{V}[X]=K/p^2$。

$$\begin{aligned}\big(C_{\mathrm{SFT}}(K,\xi)\big)^2 &= \frac{\mathbb{V}[X|X\le H]}{\mathbb{V}[X]}\\ &= \frac{\frac{K(K+1)}{p^2}\frac{\gamma(K+2,pH)}{\gamma(K,pH)}-\frac{K^2}{p^2}\left(\frac{\gamma(K+1,pH)}{\gamma(K,pH)}\right)^2}{K/p^2}\\ &= (K+1)\frac{\gamma(K+2,pH)}{\gamma(K,pH)}-K\left(\frac{\gamma(K+1,pH)}{\gamma(K,pH)}\right)^2.\end{aligned}$$

最后，代入临界性参数 $\xi=pH/K$ 可得

$$\big(C_{\mathrm{SFT}}(K,\xi)\big)^2=(K+1)\frac{\gamma(K+2,K\xi)}{\gamma(K,K\xi)}-K\left(\frac{\gamma(K+1,K\xi)}{\gamma(K,K\xi)}\right)^2. \tag{149}$$

□



(a) $C_{\mathrm{SFT}}(K,\ \xi)$ 方程（141）。

(b) $\frac{C_{\text{监督微调}}(K,\xi)\sqrt{K\xi}}{}$ 在方程（153）中。

图 *12.* ξ 与监督微调各项系数之间的关系。

由此可将标准差 σ_b^{SFT}写为:

$$\sigma_b^{\mathrm{SFT}}(K,\xi) = \sqrt{\mathbb{V}[T(\tau) \mid T(\tau) \le H]} \approx \frac{\sqrt{K}}{p} C_{\mathrm{SFT}}(K,\xi). \tag{150}$$

使用临界性参数 $\xi = pH/K$，这也可以表示为:

$$\sigma_b^{\mathrm{SFT}}(K,\xi) \approx \frac{C_{\mathrm{SFT}}(K,\xi)}{\sqrt{K\xi}} H, \tag{151}$$

其中 $C_{\mathrm{SFT}}(K,\xi) \in \mathbb{R}_+$ 是 $[0,1]$ 中的一个正标量，且同时依赖于 K 和 ξ。如图12a所示，对于任意固定的K，其行为类似于几何情形（$K = 1$）；也就是说，$\lim_{\xi\to 0} C_{\mathrm{SFT}}(K,\xi) = 0$ 以及$\lim_{\xi\to\infty} C_{\mathrm{SFT}}(K,\xi) = 1$。该函数刻画了因截断而导致的方差降低。

定理 F.5（监督微调的次优差距）。 设 $\bar{n}$ 表示成功解轨迹的数量。

$$J_r(\bar{\pi}_\kappa) - J_r(\hat{\pi}_{\bar{n}}^{\mathrm{SFT}}) = \Omega\left(\frac{\sigma_b^{\mathrm{SFT}}}{\sqrt{\bar{n}}}\right) = \Omega\left(C_{\mathrm{SFT}}(K,\xi) \cdot \frac{\sqrt{K}}{p\sqrt{\bar{n}}}\right) \tag{152}$$

$$= \Omega\left(\frac{C_{\mathrm{SFT}}(K,\xi)}{\sqrt{K\xi}} \cdot \frac{H}{\sqrt{\bar{n}}}\right). \tag{153}$$

次优差距仍然随 $1/\sqrt{\bar{n}}$缩放，表明即便增加专家数据，改进依然缓慢。此外，随着平均问题难度 $K/p = \mathbb{E}[T(\tau)]$ 增加，监督微调的性能会下降。进一步地，在问题难度固定（固定K、 p）时，增大计算预算 H （从而增大 ξ）会增大末项 $\frac{C_{\mathrm{SFT}}(K,\xi)}{\sqrt{K\xi}}$ H，因为 C_{SFT} 趋近于1，而 ξ随 H线性增长。~~决定~~次优性尺度的项 $\frac{C_{\mathrm{SFT}}(K,\xi)\sqrt{K\xi}}{}$ 在图 12b中被绘制为 ξ的函数。该图~~揭示~~子一个关键洞见：此项在 $\xi = 1$附近达到最大值，表明在该条件下次优差距最大。这一峰值凸显了监督微调的一个关键局限：当计算预算 H与预期求解时间 $\mathbb{E}[T(\tau)]$处于同一量级时，其性能会恶化。该区间使成功轨迹之间的方差最大化，从而使模仿变得尤为困难。这是因为该条件使用于训练的成功解轨迹的异质性最大化。这些成功示例的方差很高，既包括非常快的解答，也包括勉强在预算内完成的解答，因此模型极难通过模仿学到单一的连贯策略。