arXiv:2602.01485v1 [cs.LG] 1 Feb 2026

通过奖励尾部引导搜索预测并改进测试时缩放定律

Muheng Li† Jian Qian◇ Wenlong Mou†

多伦多大学统计科学系†香港大学人工智能与数据科学系◇

摘要

测试时扩展已成为提升大语言模型（LLM）推理能力的重要途径。尽管直观的“best-of-N”（BoN）策略已显著提升性能，但在 N的选择、预算分配和多阶段决策方面缺乏系统性指导，因此仍有很大的优化空间。尽管许多研究探索了此类优化，严格的理论保证仍然有限。在本研究中，我们提出了通过尾部引导搜索预测并改进扩展特性的新方法。通过估计奖励尾部分布，我们的方法无需穷尽式评估即可预测大语言模型的缩放定律。借助这一预测工具，我们提出了缩放定律引导搜索（SLG），这是一种新的测试时算法，能够动态分配计算资源，以识别并利用预测潜力最高的中间状态。我们从理论上证明，与完全信息预言机相比，SLG 能够实现消失遗憾；此外，若使用 BoN，要获得相同的期望奖励，则需要多项式级别更多的计算预算。我们在不同的大语言模型和奖励模型上对该框架进行了实证验证，证实在相同计算预算下，尾部引导分配持续获得比 Best-of-N 更高的奖励收益。我们的代码见https://github.com/PotatoJnny/Scaling-Law-Guided-search。

1 引言

过去几年，大语言模型（LLM）在能力方面取得了显著进展，尤其是在复杂推理任务中 [OpenAI, 2023年, Bubeck 等人, 2023]。推动这些进步的一个关键因素是采用测试时扩展策略，这些策略利用大语言模型输出的随机性来提升性能，而无需进一步训练模型 [Snell 等人, 2024年, Lightman 等人, 2023年, Brown 等人, 2024]。在这些策略中，“最佳的N”（BoN）方法已成为一种简单而有效的方案。通过生成多个输出，并根据预定义指标选择奖励最高的输出，BoN 已在各种任务中显著提升性能 [Li 等人, 2023年, DeepSeekAI, 2025]。

尽管 BoN 策略在实证中取得了成功，其应用仍面临若干挑战。一方面，尽管已知增加样本数量 N 通常会带来更好的性能，但样本数量 N 与期望奖励之间的确切关系仍未得到充分理解。这一认知缺口妨碍了人们就资源分配作出明智决策，尤其是在计算预算受限的情况下。另一方面，BoN 策略未考虑自适应采样可能带来的益处；在自适应采样中，中间输出可以为后续采样决策提供依据。

应对这些挑战的关键在于能够预测支配 BoN 策略性能的缩放定律，并开发能够有效利用这些预测的算法。例如，在训练时设定中，已经建立了明确的缩放定律，将模型、数据和计算预算与性能指标联系起来 [Kaplan 等人,2020年, Hoffmann 等人, 2022]。这些定律被用于指导资源分配和模型设计，并已在多种设定下得到理论证明 Lin 等人 [2025], Ben-Arous 等人 [2025]。相比之下，针对 BoN 等测试时策略的类似缩放定律在实证和理论上仍未得到充分探索。

在本研究中，我们提出了一种新颖框架，通过对奖励分布进行尾部外推，预测 BoN 的测试时缩放定律，以弥补这些不足。在此预测能力的基础上，我们进一步提出缩放定律引导搜索（SLG），这是一种根据预测奖励动态分配计算资源的自适应算法。我们的贡献包括

- 一种新方法，仅用 $m \ll N$ 个样本，通过对奖励尾部分布建模并外推至更大的 N，来预测 best-of-N 的扩展行为。
- 一种缩放定律引导（SLG）搜索算法利用预测的缩放定律，在两类推理任务中自适应地分配采样预算。我们从理论上证明，与完全信息预言机相比，SLG 能够实现消失遗憾，并达到在 BoN 下需要多项式级别更大计算预算才能实现的期望奖励。
- 我们在各种大语言模型和奖励模型上进行的实证评估表明，在相同计算预算下，我们的尾部引导分配始终优于 BoN。

本文其余部分安排如下。我们首先介绍符号约定并讨论相关工作。第 2 节形式化问题设定。第 3节介绍通过尾部外推预测缩放定律的方法。第 4 节介绍 SLG 搜索算法及其理论分析。最后，第 5 节详细介绍我们的实证评估，第 6 节则以对未来方向的讨论作结。

符号约定。对于任意整数 $n \geqslant 1$，令 $[n] = \{1, \ldots, n\}$。我们分别用 ϕ 和 Φ表示标准正态分布的概率密度函数和累积分布函数，并用 Φ^{-1} 表示其分位数函数。我们将均值为 μ、方差为 σ^2的高斯分布的密度记为 $\phi(\cdot; \mu, \sigma^2)$。

1.1 相关工作

测试时缩放定律。近期研究已确立，推理时计算量是模型性能的一个可扩展维度，有别于预训练参数。实证研究表明，随着测试时预算增加，最终输出的质量会提升，从而使较小的模型在获得充足预算的情况下能够与较大的模型竞争 [Wu 等，2024年；Chen 等，2024年；Arora 和 Zanette，2025年；Liu 等，2025年b；Schaeffer 等， 2025]。关于支配这些定律的因素，Snell 等人 [2024] 通过实证发现，最优缩放策略在很大程度上取决于提示的内在难度。后续理论研究试图解释这些现象，通常通过建模解答覆盖率与预算之间的关系 N [Brown 等，2024年；Levi，2024年；Kazdan 等， 2025] ，或分析不完美奖励模型下的泛化误差 Huang 等 [2025],Di 等 [2025], Halder 和 Pehlevan [2025]。我们的工作通过基于分布尾部特性构建更准确的统计模型，为这一理论线索作出贡献，以解释并预测测试时缩放定律。

测试时搜索算法。现有搜索策略涵盖从迭代改进到复杂结构化探索的多种方法。自我改进等经典技术 [Madaan 等人，2023年，Jiang 等人， 2025] 无需依赖外部信号即可改进推理循环。近期方法则将显式奖励模型融入搜索过程，以引导搜索，并利用不确定性感知验证 [Ye 等人，2025] 或生成式反馈（“会思考的过程奖励模型”） [Khalifa 等人， 2025]来触发修订。在最复杂的一端，思维树（ToT） [Yao 等人， 2023]以及蒙特卡洛树搜索（MCTS）框架（如 RAP [Hao 等人， 2023], 、TreeBoN [Qiu 等人， 2024] 和 LATS[Zhou 等人， 2023] ）利用奖励信号在推理空间中进行搜索。然而，这些方法通常依赖经训练的过程奖励模型（PRM）提供密集监督，或通过前瞻式滚动模拟的经验均值估计状态价值，而这种估计缺乏理论保证。相比之下，我们提出了一种新颖的搜索算法，利用预测得到的缩放定律指导资源分配，并提供具有可证明性能保证的框架。

2 问题设定

我们将测试时对齐问题建模为固定计算预算下的决策任务。

生成过程与状态。我们将生成过程建模为一条轨迹 $x = s_0 \to s_1 \to \cdots \to s_k \to y$，其中 x 是初始提示，s_i 表示一个中间状态（例如推理步骤、部分计划或草稿片段），而 y 是最终终止响应。从任意给定状态 s出发，语言模型遵循随机策略$\pi_{\text{ref}}(\cdot|s)$生成后续状态，或以响应 y终止。我们的目标是在总计算预算的约束下，找出能够产生最高奖励响应 y^* 的生成轨迹。

奖励分布与价值函数。我们假设可以访问奖励函数 $r(y, x)$，该函数用于评估完整响应 y的质量。对于任意中间状态 s （包括提示 x），随机策略会在潜在的未来结果上诱导出一个分布。令 Y_s 表示从状态 s 出发，执行策略 π_{ref} 直至终止所获得的随机最终响应。我们将其奖励的分布记为 $R_s = r(Y_s, x) \sim F_s$，并将 F_s 称为与状态 s相关联的奖励分布。

为了量化状态 s 的扩展潜力，我们将 价值函数 $V_N(s)$ 定义为：为该状态投入 N 的采样预算后所能获得的最大回报的期望值：

$$V_N(s) := \mathbb{E}\Big[\max_{1 \leqslant i \leqslant N} R_s^{(i)}\Big], \quad \text{where } R_s^{(i)} \overset{\text{i.i.d.}}{\sim} F_s. \tag{1}$$

直观地说，$V_N(s)$ 衡量了状态 s 的“扩展潜力”：如果我们投入 N 的预算来探索该状态，它所能取得的最佳结果的期望值。

预测与搜索目标。我们可以利用上述框架，在计算预算下对测试时对齐任务进行形式化。我们考虑以下两个问题。

（*1*）预测问题。给定状态 s 和目标预算 N，我们的首要目标是仅使用少量 m 奖励样本来估计状态 s 的扩展行为。具体而言，我们希望构造一个 $\widehat{\ }$ 估计器 $V_N(s)$，仅利用来自底层分布 F_s 的 $m \ll N$ 个独立同分布奖励样本$\mathcal{D} = \{r_1, r_2, \ldots, r_m\}$，来近似价值函数 $V_N(s)$。

（*2*）搜索问题。给定初始提示 x 和总预算 N，我们的第二个目标是设计一种算法，在预算范围内优化所生成的最大奖励响应的期望奖励。具体而言，在一个预算单位内，算法 $\mathcal{A}$ 可以从任何已生成的状态（包括初始提示）中采样。下一步采样哪个状态，可以根据此前观察到的所有状态及其采样得到的奖励来决定。算法 $\mathcal{A}$ 在预算 N 下的期望奖励定义为

$$V_N(\mathcal{A}; x) := \mathbb{E}\Big[\max_{y \in \mathcal{Y}_{\text{out}}} r(y, x)\Big], \tag{2}$$

其中，$\mathcal{Y}_{\text{out}}$ 是算法 $\mathcal{A}$ 在给定初始提示 x 和预算 N 的情况下生成的最终响应集合。我们的目标是找到一个算法 $\mathcal{A}$，使 $V_N(\mathcal{A}; x)$ 最大化。根据上下文明确时，我们将 $V_N(\mathcal{A}; x)$ 简写为 $V_N(\mathcal{A})$。

在本文余下部分，我们首先在第 3 节讨论预测问题，然后利用由此得到的估计器，在第 4 节设计一种资源高效搜索算法。

3 通过尾部外推预测缩放定律

基于上述设定，我们现在提出一种方法，仅使用 $m \ll N$ 个样本即可预测 $V_N(s)$在较大 N 下的扩展行为。关键洞见在于，最大回报的行为主要受奖励分布的尾部影响。因此，准确建模尾部使我们能够将 $V_N(s)$外推到更大的 N。

本节首先介绍我们方法所依据的尾部假设，并提供实证证据作为支持。随后，我们描述一种根据观测样本估计尾部参数的方法，最后给出缩放定律预测准确性的理论保证。

3.1 尾部假设与实证验证

由于测试时缩放定律的预测问题需要估计超出观测样本范围的奖励值，因此必须对奖励分布的上尾行为进行建模假设。

一项实证观察： 我们首先对大语言模型生成的回答所得到的奖励分布进行实证分析。通过收集不同提示和模型对应的奖励分布，我们观察到其尾部行为呈现出一致的模式。在图1中，我们展示了两个代表性示例的奖励样本直方图（$N = 5000$）。可以观察到，尽管奖励分布的主体通常存在偏斜，但其尾部与截断高斯分布高度吻合。拟合优度分析和Q-Q图也支持这一观察，相关内容见附录A。

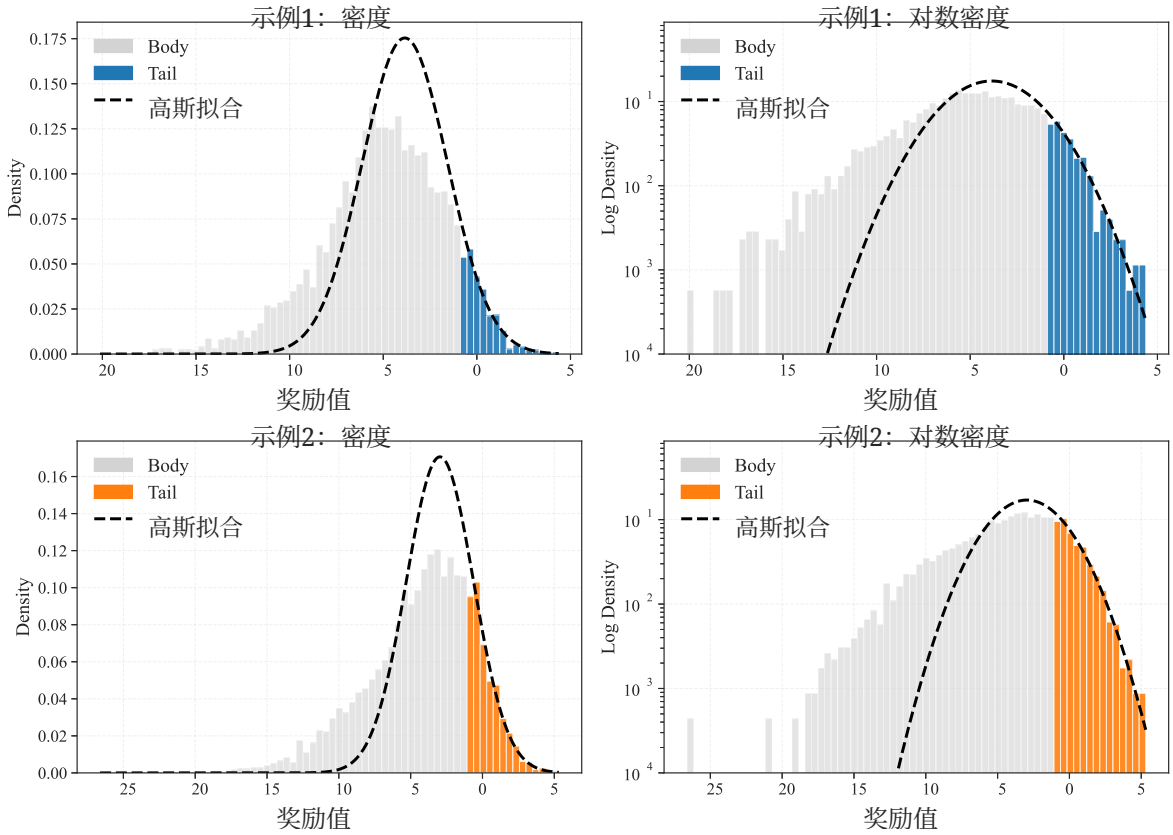


图1。奖励分布的实证分析。 我们对Llama-3.2-1B-Instruct在一道AIME 2024题目上生成的两个不同部分响应的奖励样本（$N = 5000$）进行可视化，这些样本由Skywork-Reward-V2进行评分。**（左）**标准密度直方图显示，分布主体可能包含不规则性，或偏离完美的钟形曲线。**（右）**对数密度图表明，尽管主体存在噪声，但上尾部（以彩色突出显示）严格遵循**抛物线衰减**，证实尾部行为可以用高斯分布进行良好近似。

基于这些发现，我们考虑如下高斯尾部假设。

假设1 (高斯尾部). 设 F_s 为状态 s处奖励分布的累积分布函数，且 $r_\alpha := F_s^{-1}(1-\alpha)$ 为尾部概率 $\alpha \in (0, 1)$的阈值。我们假设对所有 $r \geqslant r_\alpha$，奖励密度与高斯概率密度函数一致：

$$p_s(r) = \phi(r; \mu, \sigma^2). \tag{3}$$

其中，$\phi(\cdot; \mu, \sigma^2)$ 表示均值为 μ、方差为 σ^2 的正态分布的概率密度函数。此外，我们假设二阶矩界 $\mathbb{E}[R_s^2] \leqslant C_R$。

算法 1 通过尾部外推预测缩放定律

要求： 奖励样本 $\mathcal{D}$、尾部比例 α、预算 N。
确保： 预测期望最大值 $\hat{V}_N(s)$。
1：将经验阈值 $\hat{r}_{\alpha/2}$ 计算为 $\mathcal{D}$ 的 $(1-\alpha/2)$ 分位数。2：提取尾部集合 $\mathcal{D}_{\text{tail}} = \{r \in \mathcal{D} \mid r \geqslant \hat{r}_{\alpha/2}\}$。3：计算 $\mathcal{D}_{\text{tail}}$ 的样本均值 $\hat{\mu}_{\text{tail}}$ 和方差 $\hat{\sigma}^2_{\text{tail}}$。
4：通过求逆截断正态矩来估计参数 $(\hat{\mu},\ \hat{\sigma})$：

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2_{\text{tail}}/\delta(z)}, \quad \hat{\mu} = \hat{\mu}_{\text{tail}} - \hat{\sigma}\lambda(z) \tag{5}$$

其中，$z = \Phi^{-1}(1-\alpha/2)$ 和 $\delta(z) = 1 + z\lambda(z) - \lambda(z)^2$，$\lambda(z)$ 是逆米尔斯比率。5：预测预算 N 下的期望最大值：

$$\hat{V}_N(s) = \hat{\mu} + \hat{\sigma}E(N)$$

其中，$E(N)$ 是 N 个独立同分布标准正态变量的期望最大值。6：**返回** $\hat{V}_N(s)$

在假设1下，完整的奖励密度 $p_s(r)$可视为一个混合分布：

$$p_s(r) = \begin{cases} p_{\text{body}}(r) & \text{if } r < r_\alpha \\ \phi(r;\mu,\sigma^2) & \text{if } r \geqslant r_\alpha \end{cases}$$

我们不对分布主体施加任何约束，因此大多数样本可以服从任意的 p_{body}。

在假设 1下，当 $V_N(s)$ 的取值较大时，其行为完全由尾部参数 N 和 μ决定 σ。我们现在介绍一种估计这些参数的方法，并为估计误差提供理论保证。

3.2 缩放定律的统计估计

我们对尾部统计量采用矩估计法（MoM）。令 ϕ 和 Φ分别表示标准正态分布的概率密度函数和累积分布函数。截断正态分布的标准结果将观测到的尾部矩与生成参数联系起来：

$$\mathbb{E}[R \mid R \geqslant R_\alpha] = \mu + \sigma\lambda(z_\alpha), \tag{4a}$$
$$\text{Var}[R \mid R \geqslant R_\alpha] = \sigma^2\left(1 + z_\alpha\lambda(z_\alpha) - \lambda(z_\alpha)^2\right), \tag{4b}$$

其中 $z_\alpha = \Phi^{-1}(1-\alpha)$ 和 $\lambda(z) = \phi(z)/(1-\Phi(z))$ 是逆米尔斯比率。通过反解这些方程，我们根据经验尾部均值和方差推导出 μ 和 σ 的估计值，进而通过外推预测 $V_N(s)$。完整流程详见算法 1。

请注意，在算法 1 中，我们明确将估计限制在顶部 $\alpha/2$ 分位数，而不是整个α 尾部。该保守阈值减轻了边界处的不规则性，并使我们能够推导出以下关于 $V_N(s)$ 的非渐近误差界。

定理 1。在假设 *1* 下，对于任意 $\delta \in (0,\ 1_2)$，当样本量 $m \geqslant c_1 \log(1/\delta)$ 时，以至少 $1-\delta$ 的概率，

$$\left|\hat{V}_N(s) - V_N(s)\right| \leqslant c_2\left\{\sqrt{\frac{\log N \cdot \log(1/\delta)}{m}} + (1-\frac{\alpha}{2})^N\right\},$$

其中 c_1、c_2 是仅取决于 $(\alpha$、σ、$C_R)$ 的常数。

有关该定理的证明，请参见第 C 节。定理 1 指出，随着样本量 m 增加，估计误差以 $\mathcal{O}(\sqrt{\log N/m})$ 的速率衰减，此外还有一项指数衰减项 $(1-\alpha/2)^N$，当 N较大时，该项可忽略不计。

定理 1 提供了理论保证，使我们能够将 $\hat{} V_N(s)$ 作为可靠的决策信号。这种预测能力带来了若干关键应用。下一小节将介绍一个应用示例。更重要的是，我们将在第 4节利用这一预测工具设计一种自适应搜索算法。

3.2.1 资源分配的应用

下面 ilustrate 我们的缩放定律估计器 $\hat{} V_N(s)$ 在多个提示之间自适应分配计算预算中的一种应用。Raman 等人 [2025] 通过核密度估计（核密度估计）来估计缩放定律，探索了这一方向，并展示了相较于均匀分配的实证改进。然而，众所周知，核密度估计方法在观测样本范围之外的外推效果不佳（例如参见Tsybakov [2008]）。相比之下，我们的基于尾部的方法为外推提供了稳健的理论保证，因此更适合高预算预测。

具体而言，我们考虑将总预算 n_{total} 分配给 K 提示 $\{x_i\}_{i=1}^K$。目标是在给定权重集合 $\{w_i\}_{i=1}^K$的情况下，最大化所有提示的期望最大值奖励的加权和。

$$\max_{\{n_i\}_{i=1}^K} \sum_{i=1}^K w_i V_{n_i}(x_i) \quad \text{s.t.} \sum_{i=1}^K n_i \leqslant n_{\text{total}},\ n_i \in \mathbb{N}_+. \tag{6}$$

使用我们的缩放定律估计器 $\hat{} V_n(x_i)$，我们可以首先分配一小部分试点预算 $n_{\text{pilot}} \ll n_{\text{total}}$，以估计每个提示的缩放定律，每个提示使用 n_{pilot}/K 个样本。然后，我们可以通过将 $V_{n_i}(x_i)$替换为估计值 $\hat{} V_{n_i}(x_i)$，求解上述优化问题（6）。请注意，在我们的 $\hat{}$尾部假设下，$V_{n_i}(x_i)$ 是 n_i的凹函数。通过放宽整数约束，所得问题将成为一个凸优化问题，可以使用标准方法高效求解。最后，我们对解进行取整，以获得整数分配。

4 面向测试时对齐的缩放定律引导搜索

在第 3 节建立了预测单个状态扩展潜力的方法后，我们现在处理第 2 节定义的搜索问题。具体而言，我们聚焦于两阶段生成设定：模型首先从提示 s 中生成一个中间状态 x，然后基于 s生成最终响应 y。如第 2 节所定义，状态 s 可以解释为思维链、部分响应或一个数学步骤。

我们回顾第 2 节中的搜索问题：给定总预算 N，需要在从提示 x 中生成新的中间状态 s 与从这些状态 y 中采样最终响应之间进行分配。目标是在该预算内生成的最终响应中，最大化期望最大回报。为了设计高效的搜索策略，我们利用第 3 节推导出的缩放定律预测。我们首先将局部尾部假设扩展到全局状态空间。

假设 2。 我们假设假设*1* 对所有候选状态 s 均成立，并且具有统一的常数 α和 C_R。此外，我们假设尾部参数满足 $\sigma(s) \in [\sigma_{\min}, \sigma_{\max}]$，几乎必然地对于某些常数 $0 < \sigma_{\min} \leqslant \sigma_{\max} < \infty$ 成立；同时，潜在均值 $\mu(s)$ 具有有限的二阶矩$\mathbb{E}[\mu(s)^2] \leqslant M_\mu$。

由于中间状态 s 是由以提示为条件的随机大语言模型策略生成的x，因此，与之相关的奖励参数 $\mu(s)$ 和 $\sigma(s)$ 本质上是随机变量。假设 2 对这一诱导分布施加了温和的正则性条件：对 $\sigma(s)$ 的界限确保方差非退化（避免噪声消失或爆炸），而对 $\mu(s)$的矩条件则确保状态质量的分布保持良态。

4.1 缩放定律引导的搜索算法

利用第3节推导出的缩放定律，我们提出了一种算法，可动态分配预算 N，用于探索（寻找优质状态）和利用（对最佳状态进行重采样），如算法2所示。

具体而言，该算法将估计预算 m 分配给 K 个候选状态，以收集尾部统计量，从而推断它们在完整预算规模下的潜力。根据这些预测，算法选出最有希望的单个状态，并将剩余预算集中用于该轨迹，以最大化最终奖励。

需要注意的是，SLG算法具有很强的并行性——探索阶段可以同时针对所有 K 候选状态执行，而利用阶段则涉及从所选状态中独立采样。总体而言，SLG算法只需进行两轮顺序交互，因此在实践中效率很高。

算法 2 Scaling-Law Guided（SLG）搜索

1：输入：提示 x、总预算 N、搜索宽度 K、估计样本 m。2：输出：奖励最高的响应 y^*。3：从提示 x生成 K 中间状态 $\{s_i\}_{i=1}^{K}$ 。4：对于每个状态 s_i ，执行以下操作：5：从 s_i 生成 m个响应 $\{y_{i,j}\}_{j=1}^{m}$ ，并观测其奖励 $\{r_{i,j}\}_{j=1}^{m}$。ˆ 6：使用算法 1 和 $\{r_{i,j}\}_{j=1}^{m}$ 估计 $V_N(s_i)$。7：结束循环。ˆ ˆ 8：确定最优状态： $I = \text{argmax}_{i\in[K]} V_N(s_i)$。9：将剩余预算 $N - K \cdot m$ 分配给状态 $s_{\hat{I}}$ ，以生成更多响应。10：返回所有已生成响应中奖励最高的响应 y^* 。

4.2 理论结果

回顾一下，搜索算法的性能 $\mathcal{A}$ 由公式（2）中定义的期望奖励来衡量，为清晰起见，我们在此重述：

$$V_N(\mathcal{A}) := \mathbb{E}\Big[\max_{y\in\mathcal{Y}_{\text{out}}} r(y,x)\Big]$$

其中，集合 $\mathcal{Y}_{\text{out}}$ 由 $\mathcal{A}$ 使用预算 N生成的最终响应组成。
为了将我们的 SLG算法的性能置于具体背景中，我们引入两种参考策略进行比较。

参考策略。首先，我们考虑一种拥有完美知识的理论性能上限：

定义 1 （全信息预言机 $\mathcal{A}^*$）。预言机 $\mathcal{A}^*$ 首先从提示 x中采样一个包含 N 个候选状态$\mathcal{S}_N = \{s_1, \ldots, s_N\}$ 的候选池。然后，它利用完美分布信息确定最优状态 $s^* = \text{argmax}_{s\in\mathcal{S}_N} V_N(s)$ ，并将预算 N 用于从该状态采样。

换言之，预言机 $\mathcal{A}^*$ 基于对其奖励分布的完美了解，从 N 个候选项中选出最佳状态，并充分利用该状态。这是对 BoN 的一种自然增强，额外利用了这类知识。请注意， $\mathcal{A}^*$ 在实践中无法实现，因为它假设尾部参数的评估成本为零。即便不考虑估算成本，盲目复制这一策略也至少需要 $2N$ 的预算（生成N 个候选项，再从最佳候选项中采样 N 次）。其次，我们将标准的 Best-of-N 策略定义为衡量性能提升的基线：

定义 2 (Best-of-N 基线 $\mathcal{A}_{\text{BoN}}$)。 *Best-of-N* 策略直接从提示 x 中采样 N 个响应（等价于从状态的随机混合中采样），不进行任何中间选择。

我们将使用预言机 $\mathcal{A}^*$ 来界定遗憾(第4.2.1节），并使用基线 $\mathcal{A}_{\text{BoN}}$ 来量化优势，并通过一个具体的玩具示例（第4.2.2节）加以说明。

4.2.1 遗憾分析

现在我们来建立 SLG 算法的理论保证。首先，我们推导出在略微降低预算的情况下，SLG 相对于全信息预言机 $\mathcal{A}^*$ 的性能下界。

定理 2. 在假设 2下，考虑配置了估计样本量 m 和搜索宽度 $K = \lfloor N/2m \rfloor$的 *SLG* 算法，其消耗的总预算为 N。那么，对于 $N \geqslant 4(m+1)$ 且$m \geqslant c_0 \log N$，对于某个常数 $c_0 > 0$，有：

$$V_N(\mathcal{A}_{SLG}) \geqslant V_{\lfloor N/2m \rfloor}(\mathcal{A}^*) - c\frac{\log N}{\sqrt{m}}$$

其中，c是一个依赖于模型参数$(\alpha, \sigma_{\min}, \sigma_{\max}, C_R, M_\mu)$的常数。

定理 2 的证明见附录D。

消失遗憾。令 $m \asymp (\log N)^\alpha$，其中 $\alpha > 2$，则 SLG 能以略低的预算 $\mathcal{O}(N/(\log N)^\alpha)$达到与 $\mathcal{A}^*$ 相同的性能，同时估计误差项随 $\mathcal{O}((\log N)^{1-\alpha/2})$ 而消失。在实际设置下，这种预算削减可以忽略不计。对于轻尾奖励，预言机 $\mathcal{A}^*$通常满足如下形式的缩放定律：$V_N(\mathcal{A}^*) \approx C(\log N)^\gamma$，其中 $\gamma \in (0,1)$。[1]因此，因对数级预算削减而导致的性能损失按如下方式缩放：

$$V_N(\mathcal{A}^*) - V_{\frac{N}{\log^\alpha N}}(\mathcal{A}^*) \asymp \frac{\log\log N}{\log^{1-\gamma} N},$$

其会随着 N 的增加而减小。这意味着，在相同预算 N下，SLG 与 $\mathcal{A}^*$ 之间的差距也会消失。具体而言，取 $m \asymp (\log N)^3$,，我们有：

$$V_N(\mathcal{A}^*) - V_N(\mathcal{A}_{\mathrm{SLG}}) \leqslant \mathcal{O}\left(\frac{\log\log N}{\log^{1-\gamma} N}\right). \tag{7}$$

我们在 附录中给出了该结果在温和正则性条件下的形式化表述和证明E .

4.2.2 与Best-of-N基线的比较

为了明确量化 SLG算法带来的性能提升，我们将通用框架应用于一个具体的高斯示例。通过这一实例，我们可以推导出相对于 Best-of-N基线的优势差距（定义 2）的精确非渐近界。

假设所有中间状态的奖励分布均为高斯分布 $p_s(r) = \phi(r; \mu(s), \sigma_1)$，噪声水平固定为 $\sigma_1 > 0$，且潜在均值参数服从高斯分布$\mu(s) \sim \mathcal{N}(\mu_0, \sigma_0^2)$。我们将噪声比记为 $t = \frac{\sigma_1}{\sigma_0}$。

在这种设定下，提示的奖励分布可简化为高斯混合 $\mathcal{N}(\mu_0, \sigma_0^2 + \sigma_1^2)$。我们可以明确量化 SLG 算法相较于标准 Best-of-N基线的优势：

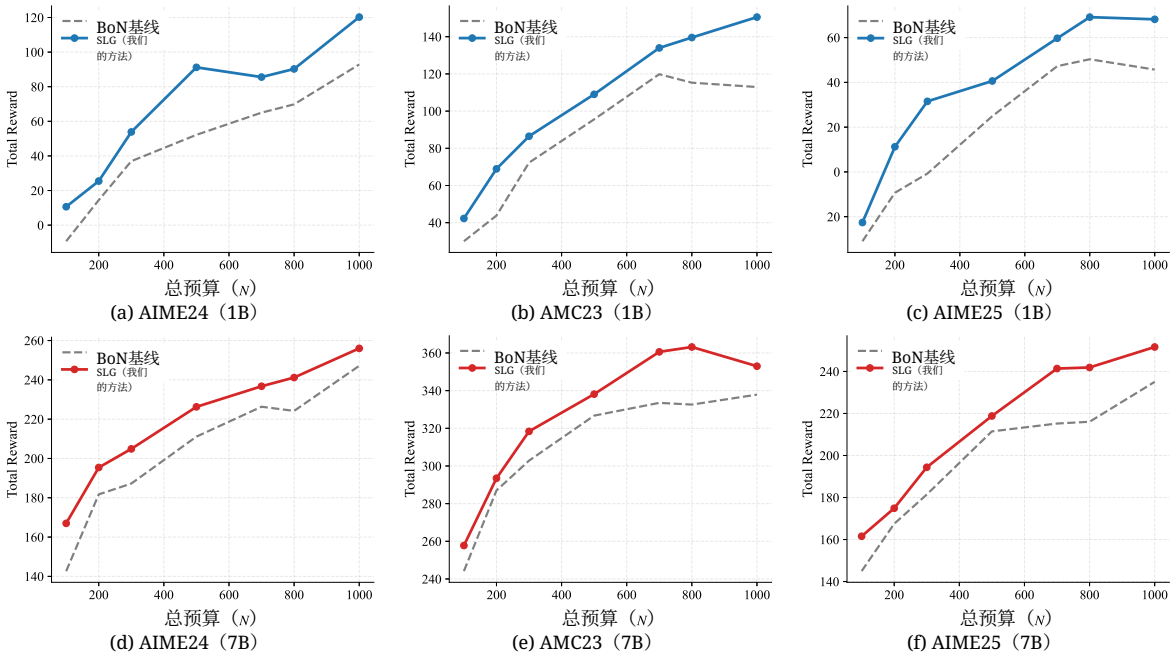
命题1。考虑上述设定。令 $t = \frac{\sigma_1}{\sigma_0}$。我们设估计样本量为$m = t^2 \log N$，候选项数量为 $K = \frac{N}{(1+t)m}$。[2]对于足够大的采样预算 N，*SLG*算法相较于~~*Best-of-N*~~基线的期望奖励提升下界为：

$$V_N(\mathcal{A}_{SLG}) - V_N(\mathcal{A}_{BoN}) \geqslant \frac{\sqrt{2}t}{4(t+1)}\sigma_0\sqrt{\log N}.$$

证明见附录F.1中命题1的证明。

命题 1 表明，SLG 严格优于 Best-of-N，其优势幅度随 $\sqrt{\log N}$增长。我们可以将这种加性奖励提升转化为 样本效率 提升，这表明 SLG 能有效放大计算预算。

[1]例如，BoN基线满足此条件，其中 $\gamma = 1/2$，如第3节所示。[2]为便于阐述，我们忽略对 m 和 K 的整数约束。



(a) AIME24（1B） (b) AMC23（1B） (c) AIME25（1B）

(d) AIME24（7B） (e) AMC23（7B） (f) AIME25（7B）

图2. 测试时扩展的主要结果。我们在三个数据集（**列**）和两种模型规模（**行**）上比较 SLG（实线）与 BoN（虚线）。其中，1B 和 7B 分别代表 Llama-3.2-1B-Instruct 和 Qwen2.5-7B-Instruct。SLG 在所有配置下都展现出更高的效率。

推论 1。 在命题 *1*的条件下，令 $t = \frac{\sigma_1}{\sigma_0}$。预算为$N$ 的 *SLG* 性能超过采用多项式级更大预算 N 运行的 *Best-of-* $N^{1+\gamma}$基线：

$$V_N(\mathcal{A}_{SLG}) \geqslant V_{N^{1+\gamma}}(\mathcal{A}_{BoN}),$$

其中我们定义 $\gamma := \frac{\sqrt{2t - ()\sqrt{1+t^2}}}{4\ t+1} > 0$。

证明请参见附录F.2，其中给出了推论 1的证明。

推论 1 表明，SLG 搜索相较于 Best-of-N基线可实现多项式计算放大。放大指数 γ 是方差比 $t = \sigma_1/\sigma_0$的函数。直观而言，如果中间状态生成（$x \to s$）或最终响应生成（$s \to y$）的方差消失，两阶段结构就会有效地坍缩为单个扁平层，导致 SLG 退化为 Best-of-N 基线。

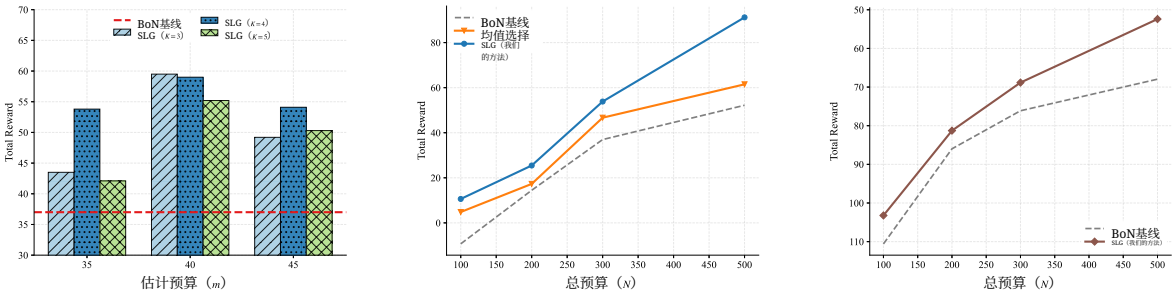
还值得注意的是，尽管状态的扩展潜力主要由其奖励方差决定 $\sigma(s)$ （如第3节所示），命题1和推论1表明，即使所有状态的方差水平都相同，仍然可以实现扩展改进。这说明，基于缩放定律进行自适应状态选择具有内在价值，而不仅仅是识别高方差状态。

5 实验

本节通过实证验证所提出的框架。我们的评估聚焦于两个关键方面：（1）**搜索效率**，评估在计算预算相同的情况下，SLG 是否优于标准的 Best-of-N 策略；以及（2）**鲁棒性验证**，确认 SLG 在各种参数设置和模型选择下的稳定性。

5.1 实验设置

继数学推理测试时扩展的近期研究之后 [Guha 等人，2025年、Agarwal 等人，2025年、Yu 等人，2025年、Wang 等人， 2025], 我们在三个可验证基准上评估我们的方法



(a) 搜索参数（K，m）　(b) 尾部感知与均值　(c) 奖励模型鲁棒性

图3. 消融研究。(a) 搜索宽度 K 和估计预算 m的影响，呈现出明显的凹形趋势，且存在清晰的最优点。(b) 与基于均值的选择基线进行比较，证实了尾部目标优化的必要性。(c) 使用基于通义千问的奖励模型进行验证，展现了对反馈信号的鲁棒性。

旨在实现先进的问题求解能力：**AMC 2023**，这是从2023年美国数学竞赛中选取的40道题目组成的子集，用于评估代数、几何和组合数学方面的非常规技能 [美国数学协会，2023]；**AIME 2024**，由2024年AIME I/II竞赛中的30道题组成 [美国数学协会，2024]；以及**AIME 2025**，即最近2025年赛季中对应的30道题 [美国数学协会，2025]。为提供反馈信号，我们使用`Skywork-Reward-V2-Llama-3.1-8B` [Liu 等人，2025a]，，这是一种最先进的结果奖励模型（ORM），可将候选解映射为标量质量分数$r \in \mathbb{R}$，并作为我们搜索优化的预言机。

基线与指标。我们主要将缩放定律引导搜索（SLG）与标准的 **Best-of-N （BoN）**基线（定义 2）进行比较。BoN 将全部计算预算 N用于生成独立响应，并选择奖励最高的候选项。为评估性能，我们采用**总奖励**作为主要指标，其定义为测试集上最大奖励的累积总和：

$$\mathcal{R}(\mathcal{A}, N) = \sum_{x \in \mathcal{X}} \max_{y \in S_{\mathcal{A}}(x;N)} r(y, x),$$

其中，$S_{\mathcal{A}}(x; N)$表示算法 $\mathcal{A}$ 在预算 N下针对提示 x 生成的候选响应集合，$r(y, x)$是 ORM 提供的奖励信号。该累积指标直接衡量固定计算预算下的整体效用，量化不同测试时策略利用奖励信号的效率。

实现细节。下面我们将说明 SLG算法的实现细节。 [3]

- **状态定义**。我们将状态 s 定义为根据提示 x生成的前 100 个词元的部分响应。

- **超参数**。我们固定尾部比例 $\alpha = 0.2$。根据公式 7，我们随预算 N动态调整参数，将估计预算设置为 $m(N) \approx \frac{1}{5}(\ln N)^3$，并将搜索宽度设置为 $K(N) \approx \sqrt[N]{2m(N)}$。确切的调整计划请参见附录 B。

对于 $K \leqslant m$ 的小预算设置，我们实施贪心试运行阶段，以初始化高质量候选项：首先采样 m 条轨迹（$x \to s \to y$），并根据其初始结果奖励选择排名前 K 的状态 s，用于后续的缩放定律估计。尽管上述配置并非针对每个特定预算都严格最优，但它能带来稳定的增益，表明我们的框架只需进行极少的超参数调优。

[3]我们的代码已公开发布于 https://github.com/PotatoJnny/Scaling-Law-Guided-search。

5.2 主要结果

我们将 SLG 与 Best-of-N （BoN）（图2）进行对比评估。

持续领先。 在所有预算和实验设置下，SLG 的表现均严格优于 BoN。这一优势在不同规模下都很稳健：在 AIME 2024 上，SLG 在 1B 模型上取得了显著领先（ $N = 1000$ 时为 120.2 对 92.9 ），在更强的 7B 模型上也同样如此（256.0 对 247.1），这证实了缩放定律指导即使对能力较强的推理模型也能带来额外增益。

资源放大。 与推论 1一致，SLG 起到了算力倍增器的作用。值得注意的是，在 AIME 2024 上使用通义千问 Qwen-7B 时，SLG 仅用 $N = 1000$，得分 337.9，便达到 BoN 的峰值性能，而所需预算仅为**一半** ($N = 500$，得分 338.1)。同样，对于 Llama-1B，SLG（$N = 500$）在 $N = 700$ (133.9 对 119.8）时也超过了 BoN。这验证了与朴素扩展相比，尾部引导搜索具有更高的样本效率。

5.3 消融研究

我们使用 Llama-3.2-1B-Instruct 在 AIME 2024 数据集上验证核心组件。

搜索参数的影响（K 与 m）。改变搜索宽度 K 和试点预算 m ，同时固定 $N = 300$ （图 3a），表明 SLG 具有稳健性；所有配置的表现均优于基线。在 $K = 3$处， $m = 40$ 我们观察到明显的性能峰值（59.5 对比 36.9），这有效地平衡了估计成本与利用资源。这证实，在广度（探索）与深度（利用）之间找到“最佳平衡点”，对于最大化计算增益至关重要。

尾部感知的重要性。图3b 对比了 SLG 与**均值选择**基线。SLG 始终占优，尤其是在 $N = 500$ ，其得分达到 **91.2** ，而**均值方法为 61.5**。这一实证结果证实，对于要求给出单个最佳响应的推理任务，显式建模奖励尾部比最大化期望值更具优势。

对奖励模型的鲁棒性。将反馈信号切换为`Skywork-Reward-V2-Qwen-8B`（图3c）后，SLG 相较于基线仍保持显著优势。这表明，我们的效率提升源于搜索动态本身，而非特定奖励模型架构带来的假象。

6 讨论

在本研究中，我们建立了一个用于测试时扩展的原则性框架，弥合了统计预测与算法搜索之间的差距。我们首先提出了一种通过尾部外推来预测 Best-of-N 性能缩放定律的方法，使我们无需穷举采样即可估计中间状态的“扩展潜力”。基于这一预测能力，我们设计了缩放定律引导搜索（SLG），这是一种利用预测结果动态分配计算资源的算法。我们的理论分析和实证结果均证实，这种尾部引导方法能够放大资源效用，在相同计算预算下持续获得高于 Best-of-N 基线的奖励。

我们将这项工作视为迈向测试时扩展严谨理论的第一步。未来值得探索的方向包括：（1）将我们的框架推广到蒙特卡洛树搜索等递归场景，以指导剪枝；（2）将这些原则应用于更广泛的智能体工作流，例如代码生成；（3）将不完美奖励模型的误差结构纳入缩放定律预测。

参考文献

Aradhye Agarwal、Ayan Sengupta 和 Tanmoy Chakraborty。先完成搜索：大语言模型中的高效测试时扩展。*arXiv* 预印本 *arXiv:2505.18149*，2025年。Daman Arora 和 Andrea Zanette。训练语言模型以实现高效推理。*arXiv* 预印本 *arXiv:2502.04463*，2025年。G. Ben-Arous、M. A Erdogdu、N. M. Vural 和 D. Wu。高维二次神经网络学习：随机梯度下降动力学与缩放定律。*arXiv* 预印本 *arXiv:2508.03688*，2025年。B. Brown、J. Juravsky、R. Ehrlich、R. Clark、Q. V. Le、C. Ré 和 A. Mirhoseini。大型语言猴子：通过重复采样扩展推理计算。*arXiv* 预印本 *arXiv:2407.21787*，2024年。S. Bubeck、V. Chandrasekaran、R. Eldan、J. Gehrke、E. Horvitz、E. Kamar、P. Lee、Y. T. Lee、Y. Li、S. Lundberg、H. Nori、H. Palangi、M. T. Ribeiro 和 Y. Zhang。通用人工智能的曙光：GPT-4 的早期实验。2023年。Lingjiao Chen、Jared Davis、Boris Hanin、Peter Bailis、Ion Stoica、Matei Zaharia 和 James Zou。更多大语言模型调用就是你所需要的一切吗？探索复合人工智能系统的扩展特性。神经信息处理系统进展，37：45767–45790，2024年。DeepSeekAI。Deepseek-r1：通过强化学习激发大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*，2025年。Qiwei Di、Kaixuan Ji、Xuheng Li、Heyang Zhao 和 Quanquan Gu。多数最优：pass@ k 推理扩展的极小极大最优策略。*arXiv* 预印本 *arXiv:2510.03199*，2025年。Etash Guha、Ryan Marten、Sedrick Keh、Negin Raoof、Georgios Smyrnis、Hritik Bansal、Marianna Nezhurina、Jean Mercat、Trung Vu、Zayne Sprague 等。Openthoughts：推理模型的数据配方。*arXiv* 预印本 *arXiv:2506.04178*，2025年。Indranil Halder 和 Cengiz Pehlevan。揭开大语言模型评判器的神秘面纱：用于推理时扩展的解析可处理模型。*arXiv* 预印本 *arXiv:2512.19905*，2025年。Shibo Hao、Yi Gu、Haodi Ma、Joshua Hong、Zhen Wang、Daisy Wang 和 Zhiting Hu。利用语言模型进行推理，就是借助世界模型进行规划。载于*2023*年自然语言处理实证方法会议论文集，第8154–8173页，2023年。J. Hoffmann、S. Borgeaud、A. Mensch、E. Buchatskaya、T. Cai、E. Rutherford、D. de Las Casas、L. A. Hendricks、J. Welbl、A. Clark 等。训练计算最优大语言模型。*arXiv* 预印本 *arXiv:2203.15556*，2022年。Audrey Huang、Adam Block、Qinghua Liu、Nan Jiang、Akshay Krishnamurthy 和 Dylan J Foster。Best-of-N 是其中最好的选择吗？推理时对齐中的覆盖率、扩展与最优性。*arXiv* 预印本 *arXiv:2503.21878*，2025年。Yuhua Jiang、Yuwen Xiong、Yufeng Yuan、Chao Xin、Wenyuan Xu、Yu Yue、Qianchuan Zhao 和 Lin Yan。Pag：将策略作为生成式验证器的多轮强化学习大语言模型自我纠正。*arXiv* 预印本 *arXiv:2506.10406*，2025年。J. Kaplan、S. McCandlish、T. Henighan、T. B. Brown、B. Chess、R. Child、S. Gray、A. Radford、J. Wu 和 D. Amodei。神经语言模型的缩放定律。*arXiv* 预印本 *arXiv:2001.08361*，2020年。Joshua Kazdan、Rylan Schaeffer、Youssef Allouah、Colin Sullivan、Kyssen Yu、Noam Levi 和 Sanmi Koyejo。高效预测大语言模型的 pass@k 扩展。*arXiv* 预印本 *arXiv:2510.05197*，2025年。Muhammad Khalifa、Rishabh Agarwal、Lajanugen Logeswaran、Jaekyeom Kim、Hao Peng、Moontae Lee、Honglak Lee 和 Lu Wang。能够思考的过程奖励模型。*arXiv* 预印本 *arXiv:2504.16828*，2025年。

Noam Levi. 推理缩放定律的简单模型。*arXiv* 预印本 *arXiv:2410.16377*，202
4. Yujia Li、David Choi 等人。使用 AlphaCode 进行竞赛级代码生成。2023年。H. Lightman、V. Kosaraju、Y. Burda、H. Edwards、B. Baker、T. Lee、J. Leike、J. Schulman、I. Sutskever 和 K. Cobbe。让我们逐步验证。载于第十二届国际学习表征会议，2023年。L. Lin、J. Wu 和 P. L. Bartlett。通过数据复用改进线性回归中的缩放定律。arXiv 预印本 arXiv:2506.08415，2025年。Chris Yuhao Liu、Liang Zeng、Yuzhen Xiao、Jujie He、Jiacai Liu、Chaojie Wang、Rui Yan、Wei Shen、Fuxiang Zhang、Jiacheng Xu 等人。Skywork-Reward-V2：通过人类与人工智能协同扩展偏好数据整理。arXiv 预印本 arXiv:2507.01352，2025a年。Runze Liu、Junqi Gao、Jian Zhao、Kaiyan Zhang、Xiu Li、Biqing Qi、Wanli Ouyang 和 Bowen Zhou。10亿参数大语言模型能否超越4050亿参数大语言模型？重新思考计算最优测试时扩展。arXiv 预印本 arXiv:2502.06703，2025b年。美国数学协会 MAA。AMC2023。 https://artofproblemsolving.com/wiki/ index.php?title=AMC_Problems_and_Solutions，2023年。 [访问日期：03-01-2026]。美国数学协会 MAA。AIME2024。 https://artofproblemsolving.com/wiki/ index.php/2024_AIME_I_Problems，2024年。 [访问日期：03-01-2026]。美国数学协会 MAA。AIME2025。https://artofproblemsolving.com/wiki/ index.php/2025_AIME_I_Problems，2025年。 [访问日期：03-01-2026]。Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang 等人。自我精炼：基于自我反馈的迭代改进。神经信息处理系统进展，36:46534–46594，2023年。OpenAI。GPT-4 技术报告。2023年。Jiahao Qiu、Yifu Lu、Yifan Zeng、Jiacheng Guo、Jiayi Geng、Huazheng Wang、Kaixuan Huang、Yue Wu 和 Mengdi Wang。TreeBoN：利用推测式树搜索和最佳 n 项采样增强推理时对齐。arXiv 预印本 arXiv:2410.16033，2024年。Vinod Raman、Hilal Asi 和 Satyen Kale。Abon：自适应最佳 n 项对齐。arXiv 预印本 arXiv:2505.12050，2025年。Michael R Sampford。关于米尔斯比率及相关函数的若干不等式。《数学统计年刊》，24(1):130–132，1953年。Rylan Schaeffer、Joshua Kazdan、John Hughes、Jordan Juravsky、Sara Price、Aengus Lynch、Erik Jones、Robert Kirk、Azalia Mirhoseini 和 Sanmi Koyejo。大型语言猴子如何获得其幂（律）？ arXiv 预印本 arXiv:2502.17578，2025年。C. Snell、J. Lee、K. Xu 和 A. Kumar。以最优方式扩展大语言模型的测试时计算，可能比扩展模型参数更有效。arXiv 预印本 arXiv:2408.03314，2024年。A. B. Tsybakov。《非参数估计导论》。施普林格科学与商业媒体，2008年。Junxiong Wang、Wen-Ding Li、Daniele Paliotta、Daniel Ritter、Alexander M Rush 和 Tri Dao。M1：利用 Mamba 推理模型实现可扩展的测试时计算。arXiv 预印本 arXiv:2504.10449，2025年。 Yangzhen Wu、Zhiqing Sun、Shanda Li、Sean Welleck 和 Yiming Yang。推理缩放定律：对使用语言模型进行问题求解的计算最优推理的实证分析。arXiv 预印本 arXiv:2408.00724，2024年。

Shunyu Yao、Dian Yu、Jeffrey Zhao、Izhak Shafran、Tom Griffiths、Yuan Cao 和 Karthik Narasimhan。思维树：使用大语言模型进行深思熟虑的问题求解。神经信息处理系统进展，36:11809–11822，2023年。

ZihuiwenYe、LuckecianoCarvalho Melo、YounesseKaddar、Phil Blunsom、SamStaton 和 Yarin Gal。《基于生成式奖励模型的不确定性感知逐步验证》。*arXiv* 预印本 *arXiv:2502.11250*，2025年。

Yue Yu、Qiwei Di、Quanquan Gu 和 Dongruo Zhou。《测试时计算的局限：通过序列奖励筛选改进推理》。*arXiv* 预印本 *arXiv:2512.04558*，2025年。

Andy Zhou、Kai Yan、Michal Shlapentokh-Rothman、Haohan Wang 和 Yu-Xiong Wang。《语言模型中的语言代理树搜索统一了推理、行动与规划》。*arXiv* 预印本 *arXiv:2310.04406*，2023年。

奖励分布的高斯尾部检验

为严格验证高斯尾部假设（假设 1），我们提供了全面的 Q-Q（分位数-分位数）图分析，对比全局分布拟合与尾部特定拟合。图 4 展示了第 3 节讨论的两个代表性示例的诊断图，表 1 则汇总了相应的拟合优度指标（R^2 值）。

- **顶行（全局拟合）：** 我们对整个数据集（$N = 5000$）拟合标准高斯分布。虽然拟合总体上较为合理，但在极端分位数处可以观察到偏差（表现为数据点偏离黑色虚线），这表明主体分布并非完全符合高斯分布，这一点也体现在较低的 R^2 值上（例如示例2中的 0.9707）。

- **底行（尾部拟合）：** 我们专门对数据中最高的 20% 拟合截断高斯分布。数据点几乎与理论线完美吻合，因此得到显著更高的 R^2 值。

这一比较证实，尽管全局奖励分布可能较为复杂，但支配缩放定律的渐近尾部行为具有严格良好的性质，并且可以用截断高斯分布准确建模。

表1. 拟合优度比较（R^2）。对于图1所示的奖励分布，我们比较了对整个数据集拟合标准高斯分布与对尾部最高 20% 拟合截断高斯分布的拟合质量。尾部拟合始终优于全局拟合，这表明高斯分布的性质在极端尾部区域最为显著。

	全局拟合 R^2	**尾部拟合 R^2**
示例1	0.9851	**0.9957**
示例2	0.9707	**0.9989**

A.1 更广泛的实证验证

为确保高斯尾部性质并非上述两个选定示例所独有，我们将分析扩展到另外9个状态的奖励分布。这些状态由*Llama-3.2-1B-Instruct* 在从 AIME 2024 数据集中随机采样的问题上生成，并由 Skywork-Reward-V2-Llama-3.1-8B 评分。

图 5 展示了这些分布的奖励密度直方图。在每张图中，我们展示了专门针对顶部 20% 尾部（蓝色条形）拟合的高斯曲线（黑色虚线），其余分布主体则以灰色显示。

如图5所示，分布主体高度偏离高斯分布，呈现出显著的偏度和多模态性。相比之下，尾部区域始终与拟合的高斯分布

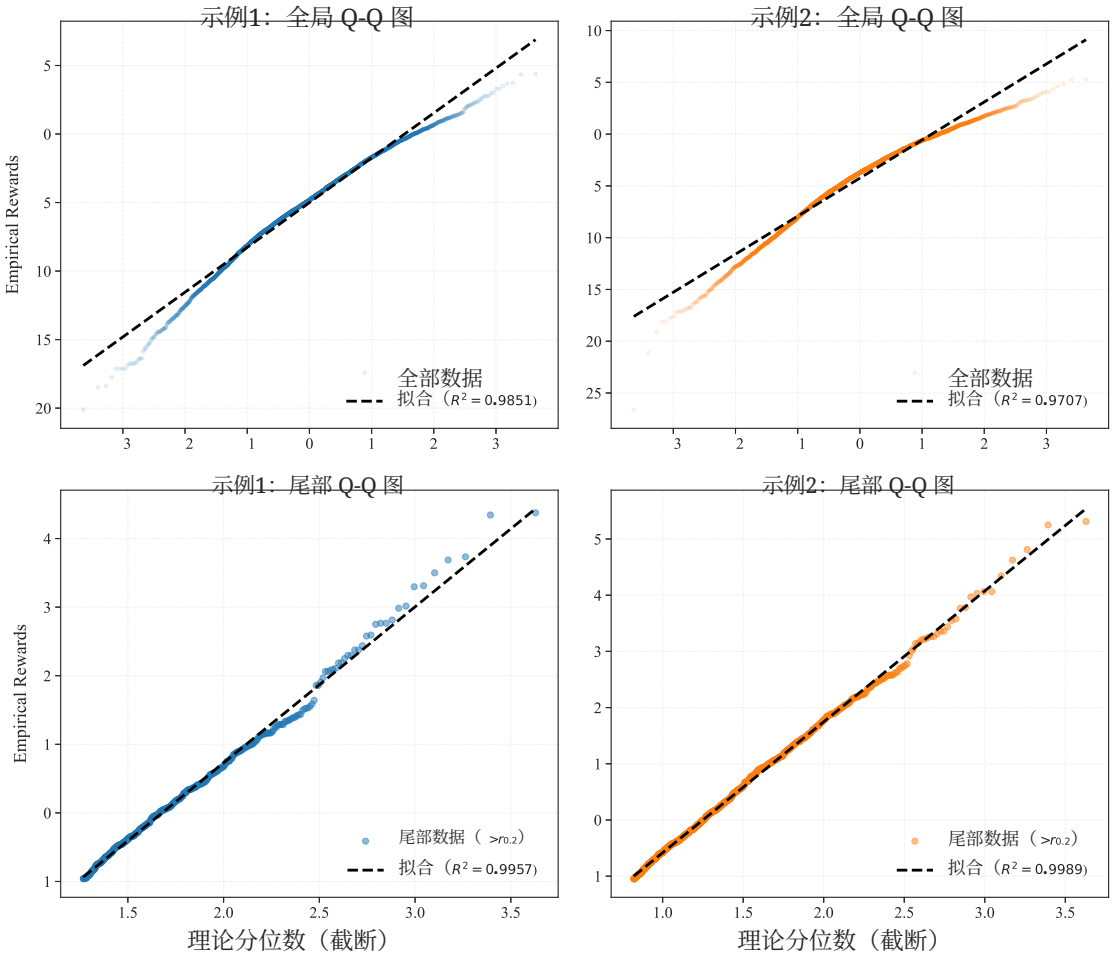


图 4. 详细的 Q-Q 图分析。上排：整个奖励分布相对于标准正态分布的 Q-Q 图。点的密度很高，且两端略有偏离，表明主体存在不规则性。**下排：**尾部区域（$> R_{0.2}$）相对于截断正态分布的 Q-Q 图。严格的线性排列和近乎完美的 R^2 分数证实了高斯尾部假设用于外推的有效性。

包络线吻合，且这一现象在全部 9 个示例中均成立。这证实，尽管奖励景观整体复杂，其关键尾部行为仍具有可预测性，从而严格验证了我们的高斯尾部假设（假设1）。

B 实验细节

我们在表5.2中列出了主要实验（第2节）所用的确切超参数值。搜索宽度 K 和估计预算 m 根据总预算 N动态计算。我们首先通过按 m 缩放并四舍五入到最接近的 5 的倍数来确定估计预算$\ln N$：

$$m(N) = 5 \times \text{round}\left(\frac{(\ln N)^3}{25}\right).$$

给定 $m(N)$，我们计算目标探索宽度 K。为确保总计算成本不超出限制，我们将 K 定义为受最大可行预算约束的目标比率：

$$K(N) = \min\left(\text{round}\left(\frac{N}{2m(N)}\right), \left\lfloor\frac{N}{m(N)+2}\right\rfloor\right).$$

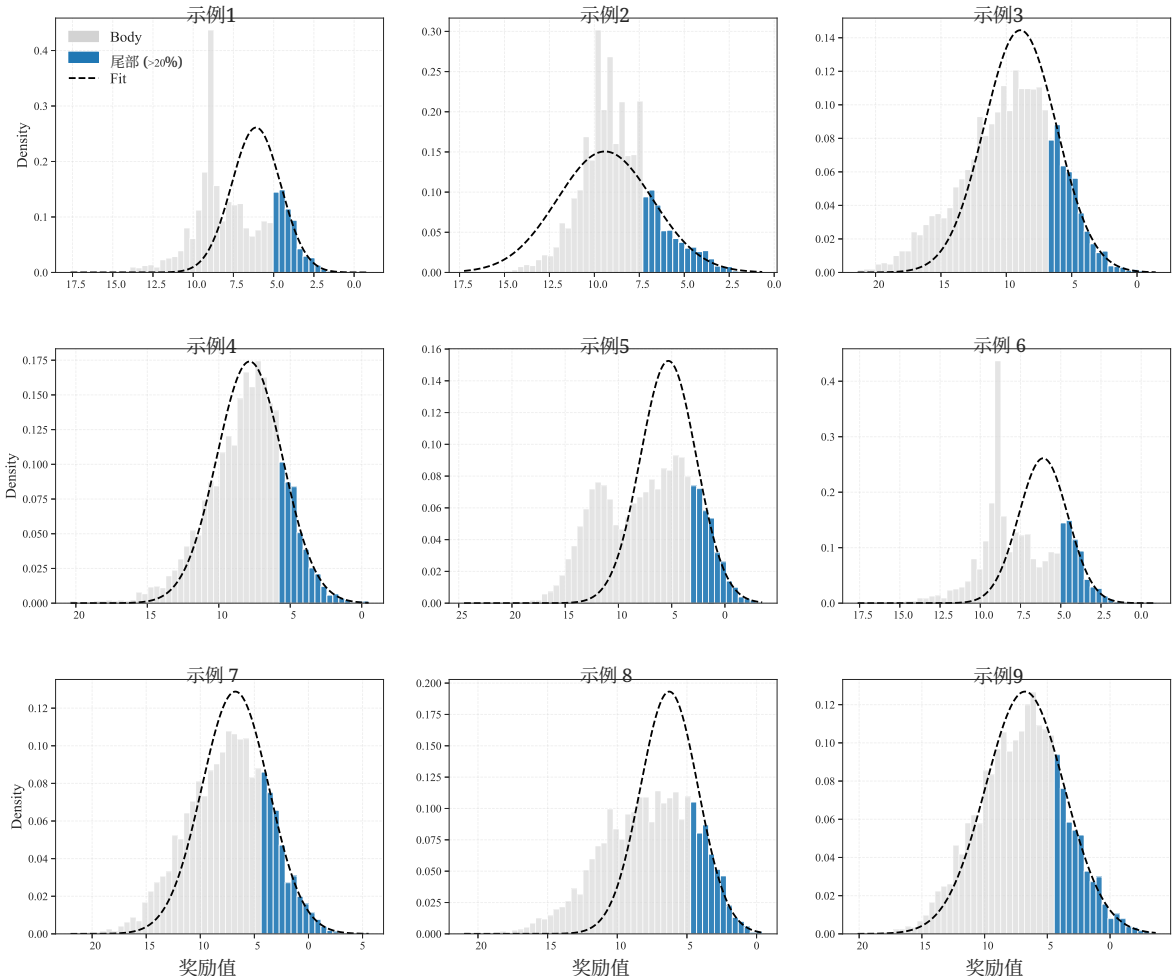


图5。高斯尾部的广泛可视化验证。 我们将展示9个随机抽取的查询实例的奖励密度，这些实例不同于正文中的示例。**灰色条形：** 分布主体（第< 80百分位数）。**蓝色条形：** 尾部区域（第> 80百分位数）。**虚线：** 专门针对尾部统计量拟合的高斯分布。尽管分布主体的形状不规则且多种多样，但尾部区域始终遵循可预测的高斯轮廓。

这一调度方案在准确估计（增大 m）与充分的探索宽度（K）之间取得了平衡，并且符合推论 2 中的理论见解。

C 定理证明1

根据假设 1，我们知道奖励分布 $p(r) = \phi(r;\ \mu,\ \sigma)$ 对于 $r \geqslant r_{\alpha/2}$。我们将 $M(N)$ 定义为从 N 中抽样 $\mathcal{N}(\mu,\ \sigma^2)$ 次时的期望最大值，则有：

$$M(N) = \mu + \sigma E(N), \quad \text{where} \quad E(N) = \mathbb{E}_{Z_i \sim \mathcal{N}(0,1)} \left[\max_{i=1}^{N} Z_i \right] \tag{8}$$

并且 $\hat{}\ V_N(s) = \hat{}\mu + \hat{}\sigma E(N)$ 是 $M(N)$ 的估计器。
我们可以将总误差分解为两部分：

$$\left| V_N(s) - \hat{V}_N(s) \right| \leqslant |V_N(s) - M(N)| + \left| \hat{V}_N(s) - M(N) \right|$$

其中，第一部分是由高斯尾部近似引入的偏差误差，第二部分是高斯最大值的估计误差。

表 2. 超参数调度。 每个总预算 (N) 所使用的确切估计预算 (m) 和探索宽度 (K)。根据以下公式计算：$m \approx \frac{1}{5}(\ln N)^3$ 和 $K \approx \frac{N}{2m}$。

总预算 (N)	**估计预算**(m)	**探索宽度**(K)
100	20	2
200	30	3
300	35	4
500	50	5
700	55	6
800	60	7
1000	65	8

C.1 估计误差

我们注意到，在高斯尾部模型（假设 1）下，可以将估计误差写为：

$$\begin{aligned}\left|\hat{V}_N - M(N)\right| &= |(\hat{\mu}-\mu) + (\hat{\sigma}-\sigma)E(N)| \\ &\leqslant |\hat{\mu}-\mu| + |\hat{\sigma}-\sigma|E(N)\end{aligned}$$

因此，只需根据公式 (4) 对估计误差 $\hat{}\mu - \mu|$ 和 $\hat{}\sigma - \sigma$进行界定，或者等价地，对 $\hat{}\mu_{\text{tail}} - \mathbb{E}[R|R \geqslant r_{\alpha/2}]|$ 和 $\hat{}\sigma_{\text{tail}} - \sqrt{\text{Var}[R|R \geqslant r_{\alpha/2}]}|$ 进行界定，其中 $\hat{}\mu_{\text{tail}}$ 和 $\hat{}\sigma_{\text{tail}}$ 是根据尾部数据集 $D_{\text{tail}} = \{r_i \mid r_i \geqslant \hat{}r_{\alpha/2}, i = 1, 2, \cdots, m_{tail}\}$计算出的经验均值和标准差。我们可以将这些误差分解为两个部分：

- **偏差误差：** 该误差源于用经验阈值 $\hat{}r_{\alpha/2}$近似真实阈值$r_{\alpha/2}$ 。记作：

$$\mu(r) = \mathbb{E}[R|R \geqslant r], \quad \sigma^2(r) = \text{Var}[R|R \geqslant r]$$

  表示给定一般阈值 R时的条件均值和方差。偏差误差可表示为：

$$|\mu(\hat{r}_{\alpha/2}) - \mu(r_{\alpha/2})|, \quad |\sigma(\hat{r}_{\alpha/2}) - \sigma(r_{\alpha/2})|$$

- **统计误差：** 该误差源于使用有限数量的尾部样本估计条件矩 $\mu(\hat{}r_{\alpha/2})$ 和 $\sigma^2(\hat{}r_{\alpha/2})$，可表示为：

$$|\hat{\mu}_{\text{tail}} - \mu(\hat{r}_{\alpha/2})|, \quad |\hat{\sigma}_{\text{tail}} - \sigma(\hat{r}_{\alpha/2})|$$

We 首先对经验阈值 $\hat{}r_{\alpha/2}$的误差构造一个高概率界：

引理1。 在假设 *1*下，存在一个常数 $c_1 > 0$ ，使得对于满足 m的所有样本量 $m \geqslant c_1 \log(1/\delta)$，以至少 $1-\delta$的概率：

$$\left|\hat{r}_{\alpha/2} - r_{\alpha/2}\right| \leqslant c_2 \sqrt{\frac{\log(1/\delta)}{m}},$$

其中， c_1、 c_2 是取决于 (α、 σ)的一般正常数。

引理的证明 *1.* 令 $\hat{}$ $F_m(r)$ 表示奖励样本的经验累积分布函数 $\{r_i\}_{i=1}^m$。根据 Dvoretzky–Kiefer–Wolfowitz（DKW）不等式，事件

$$\mathcal{E} = \left\{\sup_{r\in\mathbb{R}} |\hat{F}_m(r) - F(r)| \leqslant \epsilon_m\right\}, \quad \text{where} \quad \epsilon_m = \sqrt{\frac{\log(2/\delta)}{2m}},$$

occ发生的概率至少为 $1-\delta$。在证明的其余部分中，我们以 $\mathcal{E}$ 为条件。

记 $\tau = r_{\alpha/2} - r_\alpha$，其中 $r_\alpha = \Phi^{-1}(1-\alpha;\mu,\sigma)$ 是高斯分布的 α 分位数。我们首先证明 $\hat{r}_{\alpha/2}$ 位于局部邻域 $\mathcal{N} = [r_{\alpha/2} - \tau, r_{\alpha/2} + \tau]$ 内。根据假设1，在该区域内，奖励密度 $p(r)$ 等于高斯密度 $\phi(r;\mu,\sigma)$，因此对所有 $r \in \mathcal{N}$，均有 $p(r) \geqslant \kappa > 0$，其中 $\kappa = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r_{\alpha/2}+\tau-\mu)^2}{2\sigma^2}}$ 是一个依赖于 (α,σ) 的常数。

[4]当 $m \geqslant \frac{4}{\kappa^2\tau^2}\log(1/\delta)$ 时，我们有：

$$\epsilon_m = \sqrt{\frac{\log(2/\delta)}{2m}} \leqslant \frac{1}{2}\kappa\tau.$$

在此条件下，利用密度下界 $F(r_{\alpha/2}+\tau) \geqslant (1-\alpha/2) + \kappa\tau$，我们得到：

$$\hat{F}_m(r_{\alpha/2}+\tau) \geqslant F(r_{\alpha/2}+\tau) - \epsilon_m \geqslant (1-\alpha/2) + \kappa\tau - \frac{1}{2}\kappa\tau > 1 - \alpha/2 + \frac{1}{m}.$$

这意味着 $\hat{r}_{\alpha/2} \leqslant r_{\alpha/2} + \tau$。通过对称论证可得 $\hat{R}_\alpha \geqslant R_\alpha - \tau$。因此，$\hat{r}_{\alpha/2} \in \mathcal{N}$。由于 $\hat{r}_{\alpha/2} \in \mathcal{N}$，我们可以利用密度下界 κ 对累积分布函数之差进行反演。根据累积分布函数 $F(\cdot)$ 的中值定理：

$$|F(\hat{r}_{\alpha/2}) - F(r_{\alpha/2})| = p(\tilde{r})|\hat{r}_{\alpha/2} - r_{\alpha/2}| \geqslant \kappa|\hat{r}_{\alpha/2} - r_{\alpha/2}|.$$

整理后，并通过 DKW 界将真实累积分布函数与经验分位数联系起来：

$$\begin{aligned}
|\hat{r}_{\alpha/2} - r_{\alpha/2}| &\leqslant \frac{1}{\kappa}|F(\hat{r}_{\alpha/2}) - F(r_{\alpha/2})| \\
&\leqslant \frac{1}{\kappa}\left(|F(\hat{r}_{\alpha/2}) - \hat{F}_m(\hat{r}_{\alpha/2})| + |\hat{F}_m(\hat{r}_{\alpha/2}) - (1-\alpha/2)|\right) \\
&\leqslant \frac{1}{\kappa}\left(\epsilon_m + \frac{1}{m}\right).
\end{aligned}$$

代入 $\epsilon_m = \sqrt{\frac{\log(2/\delta)}{2m}}$，可知对于 $m \geqslant c_1\log(1/\delta)$，以下结论以至少 $1-\delta$ 的概率成立：

$$|\hat{r}_{\alpha/2} - r_{\alpha/2}| \leqslant c_2\sqrt{\frac{\log(1/\delta)}{m}},$$

其中某些常数 $c_1, c_2 > 0$ 依赖于 (α,σ)，由此完成证明。 □

那么，对于偏差误差，我们有以下引理：

引理 2. 在假设 *1* 下，存在常数 $\tau > 0$ 和 $c > 0$ (取决于 α,σ)，使得对于所有满足 r 的 $\Delta = |r - r_{\alpha/2}| \leqslant \tau$，都有：

$$\left|\mu(r) - \mu(r_{\alpha/2})\right| \leqslant c\Delta, \quad \textit{and} \quad \left|\sigma(r) - \sigma(r_{\alpha/2})\right| \leqslant c\Delta.$$

引理 *2* 的证明。我们选择 $\tau = r_{\alpha/2} - r_\alpha = \sigma\left(\Phi^{-1}(1-\alpha/2) - \Phi^{-1}(1-\alpha)\right)$，并将 $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$ 作为逆米尔斯比率。
我们分别证明以下两个结论：

1. **均值偏差误差**

   令 $\mu_G(r)$ 表示截断于 r 的高斯分布的均值：

$$\mu_G(r) = \frac{\int_r^\infty x\phi(x;\mu,\sigma^2)dx}{\int_r^\infty \phi(x;\mu,\sigma^2)dx}.$$

[4]由于 $\tau = \sigma(\Phi^{-1}(1-\alpha/2) - \Phi^{-1}(1-\alpha))$ 和 $r_{\alpha/2} - \mu = \sigma\Phi^{-1}(1-\alpha/2)$，常数 κ 与 μ 无关。

对于任意 r 满足 $\Delta = |r - r_{\alpha/2}| \leqslant \tau$的情况，我们知道 r 和 $r_{\alpha/2}$ 都位于奖励密度 $p(r)$ 等于高斯密度 $\phi(r;\, \mu,\ \sigma^2)$ 的区域内，根据假设 1。因此，我们有：

$$\left|\mu(r) - \mu(r_{\alpha/2})\right| = \left|\mu_G(r) - \mu_G(r_{\alpha/2})\right|$$

这意味着，只需对两个截断点处截断高斯分布的均值之差进行界定即可；我们知道，截断高斯分布的均值为$\mu + \sigma\lambda(\frac{r-\mu}{\sigma})$。关于 $\lambda(x)$ 的一个众所周知的性质是 $0 \leqslant \lambda'(x) \leqslant 1$ 对所有 $x \in \mathbb{R}$ [Sampford, 1953],由此可得：

$$\begin{aligned}\left|\mu_G(r) - \mu_G(r_{\alpha/2})\right| &\leqslant \left|\sigma\lambda(\frac{r-\mu}{\sigma}) - \sigma\lambda(\frac{r_{\alpha/2}-\mu}{\sigma})\right| \\ &\leqslant \sup_x |f'(x)| \cdot \sigma \left|\frac{r_{\alpha/2}-\mu}{\sigma} - \frac{r-\mu}{\sigma}\right| \leqslant \Delta.\end{aligned}$$

至此，第一个结论的证明完成。

2. **方差偏差误差**

令 $\sigma_G^2(r)$ 表示在 r 处截断的高斯分布的方差。同样，对于满足 $\Delta = |r - r_{\alpha/2}| \leqslant \tau$ 的任意 r，有：

$$\left|\sigma^2(r) - \sigma^2(r_{\alpha/2})\right| = \left|\sigma_G^2(r) - \sigma_G^2(r_{\alpha/2})\right|.$$

我们知道，截断高斯分布的方差可表示为：

$$\sigma_G^2(r) = \sigma^2\delta(\frac{r-\mu}{\sigma}), \quad \text{where} \quad \delta(z) = 1 + z\lambda(z) - \lambda(z)^2.$$

利用标准截断正态方差 $\delta(z) = 1 + z\lambda(z) - \lambda(z)^2$ 的性质。直接计算可得，对于所有 z，$|\delta'(z)| \leqslant 1$ 有：

$$\left|\sigma_G^2(r) - \sigma_G^2(r_{\alpha/2})\right| = \sigma^2\left|\delta(\frac{r-\mu}{\sigma}) - \delta(\frac{r_{\alpha/2}-\mu}{\sigma})\right| \leqslant \sigma^2\left|\frac{r_{\alpha/2}-\mu}{\sigma} - \frac{r-\mu}{\sigma}\right| \leqslant \sigma\Delta.$$

由于 $\sigma(r) > 0$ 和 $\sigma(r_{\alpha/2})$ 是仅依赖于 σ 和 α 的正常数，因此我们有：

$$\left|\sigma(r) - \sigma(r_{\alpha/2})\right| \leqslant \frac{1}{\sigma(r_{\alpha/2})}\left|\sigma^2(r) - \sigma^2(r_{\alpha/2})\right| \leqslant c\Delta,$$

这就完成了第二个结论的证明。

□

对于统计误差，我们有如下集中性结果：

引理 3。 在假设 *1*下，对于 $\delta \in (0,1)$，当样本量 $m \geqslant c_1 \log(1/\delta)$时，以至少 $1 - \delta$ 的概率，

$$\begin{aligned}\left|\hat{\mu}_{tail} - \mu(\hat{r}_{\alpha/2})\right| &\leqslant c_2\sqrt{\frac{\log(1/\delta)}{m}} \\ \left|\hat{\sigma}_{tail} - \sigma(\hat{r}_{\alpha/2})\right| &\leqslant c_2\sqrt{\frac{\log(1/\delta)}{m}}\end{aligned}$$

其中，c_1、c_2 是仅取决于$(\alpha、\ \sigma)$的常数。

引理的证明3。我们首先定义事件 $E_R = \left\{\left|r_{\alpha/2} - \hat{r}_{\alpha/2}\right| \leqslant r_{\alpha/2} - r_\alpha\right\}$。在此事件下，根据假设 1 $R|R \geqslant \hat{r}_{\alpha/2}$ 是一个截断高斯随机变量，因此是参数为 σ的次高斯变量。根据引理1，当 $m \geqslant c\log(1/\delta)$ 对于某个常数 c_1 （该常数仅取决于 α和 σ）时，我们有$\mathbb{P}(E_R) \geqslant 1 - \delta/3$。

我们进一步定义 以下两个事件，用于经验均值和方差的集中:

$$E_\mu = \left\{\left|\hat{\mu}_{tail} - \mu(\hat{r}_{\alpha/2})\right| \leqslant \sqrt{2}\sigma\sqrt{\frac{\log(6/\delta)}{\lfloor m\alpha/2\rfloor}}\right\},$$
$$E_\sigma = \left\{\left|\sigma^2_{tail} - \sigma^2(\hat{r}_{\alpha/2})\right| \leqslant 2\sqrt{2}\sigma^2\sqrt{\frac{\log(6/\delta)}{\lfloor m\alpha/2\rfloor}}\right\}$$

其中 $\sigma^2_{tail} = \frac{1}{\lfloor m\alpha/2\rfloor}\sum_{r_i\in\mathcal{D}_{tail}}(r_i - \mu(\mathcal{D}_{tail}))^2$ 是尾部数据集 $\mathcal{D}_{tail}$ 的真实方差。根据次高斯和次指数随机变量的集中不等式，我们知道，在事件 E_R 下，当 $m \geqslant c'\log(1/\delta)$ 对某个仅取决于 α 和 σ 的常数 c' 成立时，有 $\mathrm{P}(E_\mu) \geqslant 1 - \delta/3$ 且 $\mathrm{P}(E_\sigma) \geqslant 1 - \delta/3$。

因此，如果我们定义事件 $E = E_\mu \cap E_\sigma \cap E_R$，那么根据并集界，我们知道，当 $m \geqslant c_1\log(1/\delta)$ 对某个仅取决于 α 和 σ 的常数 c_1 成立时，有 $\mathrm{P}(E) \geqslant 1 - \delta$。在事件 E 下，我们可以直接得到:

$$\left|\hat{\mu}_{tail} - \mu(\hat{r}_{\alpha/2})\right| \leqslant \sqrt{2}\sigma\sqrt{\frac{\log(6/\delta)}{\lfloor m\alpha/2\rfloor}} \leqslant c_2\sqrt{\frac{\log(1/\delta)}{m}}$$

对于某个常数 c_2，它仅取决于 α 和 σ。对于方差估计，在事件 E 条件下，我们有:

$$\begin{aligned}
\left|\hat{\sigma}^2_{tail} - \sigma^2(\hat{r}_{\alpha/2})\right| &= \left|\frac{1}{\lfloor m\alpha/2\rfloor}\sum_{r_i\in\mathcal{D}_{tail}}(r_i - \hat{\mu}(\mathcal{D}_{tail}))^2 - \sigma^2(\hat{r}_{\alpha/2})\right| \\
&\leqslant \left|\frac{1}{\lfloor m\alpha/2\rfloor}\sum_{r_i\in\mathcal{D}_{tail}}(r_i - \mu(\hat{r}_{\alpha/2}))^2 - \sigma^2(\hat{r}_{\alpha/2})\right| + \left|\hat{\mu}(\mathcal{D}_{tail}) - \mu(\hat{r}_{\alpha/2})\right|^2 \\
&\leqslant 2\sqrt{2}\sigma^2\sqrt{\frac{\log(6/\delta)}{\lfloor m\alpha/2\rfloor}} + 2\sigma^2\frac{\log(6/\delta)}{\lfloor m\alpha/2\rfloor} \\
&\leqslant c_2'\sqrt{\frac{\log(1/\delta)}{m}}
\end{aligned}$$

并且，由于在事件 E_R 条件下， $\sigma(\hat{r}_{\alpha/2}) \geqslant c_\sigma > 0$，其中某个常数 c_σ 仅取决于 σ 和 α，因此我们有:

$$|\hat{\sigma}_{tail} - \sigma(\hat{r}_\alpha)| \leqslant \frac{1}{\sigma(\hat{r}_{\alpha/2})}\left|\hat{\sigma}^2_{tail} - \sigma^2(\hat{r}_\alpha)\right| \leqslant c_2\sqrt{\frac{\log(1/\delta)}{m}}$$

对于某个常数 c_2，它仅取决于 α 和 σ。证明完毕。 □

综合这些引理，我们得到以下关于 $\hat{V}_N(s)$ 总体估计误差的命题:

命题 2. 在假设 *1* 下，对于任意 $\delta \in (0, 1/2)$，当样本量 $m \geqslant c_1\log(1/\delta)$ 时，以至少 $1 - \delta$ 的概率，

$$\left|\hat{V}_N(s) - M(N)\right| \leqslant c_2\sqrt{\log N}\sqrt{\frac{\log(1/\delta)}{m}}$$

其中， c_1、 c_2 是仅取决于(α、 σ)的常数。

命题2的证明。我们采用引理3证明中的同一事件

$$E = E_R \cap E_\mu \cap E_\sigma$$

其中 E_R, E_μ, E_σ 在引理 3 的证明中定义，并且我们知道当 $m \geqslant c\log(1/\delta$ ），$\mathbf{P}(E) \geqslant 1-\delta$，f 或某个常数 $c > 0$ 取决于 (α, σ)。
并且在事件 E 下，根据引理 2，当 $m \geqslant c'\log(1/\delta)$ 对于某个常数 $c' > 0$ 取决于(α, σ)，我们有：

$$|\mu(\hat{r}_{\alpha/2}) - \mu(r_{\alpha/2})| \leqslant c_3 \left|\hat{r}_{\alpha/2} - r_{\alpha/2}\right| \leqslant c_3' \sqrt{\frac{\log(1/\delta)}{m}}$$
$$|\sigma(\hat{r}_\alpha/2) - \sigma(r_{\alpha/2})| \leqslant c_3 \left|\hat{r}_{\alpha/2} - r_{\alpha/2}\right| \leqslant c_3' \sqrt{\frac{\log(1/\delta)}{m}}$$

对于某些常数 $c_3, c_3' > 0$ 取决于 (α, σ)。
因此，根据引理 3 和三角不等式，我们知道当 $m \geqslant c_1\log(1/\delta)$ 对于某个常数$c_1 > 0$ 依赖于 (α, σ)，以至少 $1-\delta$ 的概率，

$$|\hat{\mu}_{tail} - \mu(r_{\alpha/2})| \leqslant \left|\hat{\mu}_{tail} - \mu(\hat{r}_{\alpha/2})\right| + \left|\mu(\hat{r}_{\alpha/2}) - \mu(r_{\alpha/2})\right| \leqslant c_\mu \sqrt{\frac{\log(1/\delta)}{m}}$$
$$|\hat{\sigma}_{tail} - \sigma(r_{\alpha/2})| \leqslant \left|\hat{\sigma}_{tail} - \sigma(\hat{r}_{\alpha/2})\right| + \left|\sigma(\hat{r}_{\alpha/2}) - \sigma(r_{\alpha/2})\right| \leqslant c_\sigma \sqrt{\frac{\log(1/\delta)}{m}}$$

并且根据 $\hat{}\mu$、$\hat{}\sigma$ 在算法 1 中的公式，对于某些常数 c_μ'、$c_\sigma' > 0$（其依赖于 (α, σ)），我们有：

$$|\hat{\sigma} - \sigma| = \frac{|\hat{\sigma}(\mathcal{D}_{tail}) - \sigma(r_{\alpha/2})|}{\sqrt{\delta(\frac{r_{\alpha/2}-\mu}{\sigma})}} \leqslant c_\sigma' \sqrt{\frac{\log(1/\delta)}{m}}$$

$$|\hat{\mu} - \mu| = |\hat{\mu}(\mathcal{D}_{tail}) - \mu(r_{\alpha/2})| + \lambda(\frac{r_\alpha - \mu}{\sigma})|\hat{\sigma} - \sigma| \leqslant c_\mu' \sqrt{\frac{\log(1/\delta)}{m}}$$

其中 $\lambda(z)$ 是逆米尔斯比率，且 $\delta(z) = 1 + z\lambda(z) - \lambda(z)^2$。
最后，我们可以界定经验估计误差，其对象为$\hat{}$的概率：V_N

$$\begin{aligned}
|\hat{V}_N(s) - M(N)| &= |\hat{\mu} + \hat{\sigma}E(N) - \mu - \sigma E(N)| \\
&\leqslant |\hat{\mu} - \mu| + |\hat{\sigma} - \sigma| \cdot E(N) \\
&\leqslant c_\mu' \sqrt{\frac{\log(1/\delta)}{m}} + c_\sigma' \sqrt{\frac{\log(1/\delta)}{m}} \sqrt{2\log N} \\
&\leqslant c_2 \sqrt{\log(N)} \sqrt{\frac{\log(1/\delta)}{m}}
\end{aligned}$$

对于某个常数 $c_2 > 0$（其仅依赖于 (α, σ)）。至此证明完成。 □

C.2 偏差误差

我们可以建立一个引理，用于界定近似偏差 $|V_N(s) - M(N)|$。这种偏差源于真实奖励分布 F_s 在分布的“主体”部分与高斯分布不同（$r < r_{\alpha/2}$）。

引理 4. 在假设 *1*下，真实期望最大值 $V_N(s)$ 与高斯期望最大值 $M(N)$满足：

$$|V_N(s) - M(N)| \leqslant c(1-\alpha/2)^N, \tag{9}$$

其中 c 是一个仅依赖于(α, σ, C_R)的有限常数。

引理4的证明。 设 $F(r)$ 为真实奖励分布的累积分布函数，且 $G(r) = \Phi(r; \mu, \sigma^2)$ 为高斯近似的累积分布函数。于是我们有：

$$V_N(s) - M(N) = \int_{-\infty}^{\infty} (G(r)^N - F(r)^N) dr.$$

根据假设 1，当 $r \geqslant r_{\alpha/2}$ 时，累积分布函数相同：

$$F(r) = G(r), \quad \forall r \geqslant r_{\alpha/2},$$

这意味着在 $[r_{\alpha/2}, \infty)$ 上的积分为零。因此，我们只需考虑在 $(-\infty, r_{\alpha/2}]$ 上的积分：

$$|V_N(s) - M(N)| = \left| \int_{-\infty}^{r_{\alpha/2}} (G(r)^N - F(r)^N) dr \right| \leqslant \int_{-\infty}^{r_{\alpha/2}} |G(r)^N - F(r)^N| dr. \tag{10}$$

对于 $r \leqslant r_{\alpha/2}$，两个累积分布函数均以 $1 - \alpha/2$ 为上界。因此，我们可以进一步对积分取上界：

$$\begin{aligned} \int_{-\infty}^{r_{\alpha/2}} |G(r)^N - F(r)^N| dr &\leqslant \int_{-\infty}^{r_{\alpha/2}} (G(r)^N + F(r)^N) dr \\ &\leqslant (1 - \alpha/2)^{N-1} \int_{-\infty}^{r_{\alpha/2}} (G(r) + F(r)) dr. \end{aligned}$$

令 $c = \int_{-\infty}^{r_{\alpha/2}} (G(r) + F(r)) dr$。通过变量替换 $u = r - \mu$，我们注意到该积分仅取决于中心化分布和中心化阈值 $r_{\alpha/2} - \mu$。因此， c 与位置参数无关 μ。最后，假设 1 保证二阶矩有限（$\mathbb{E}[R^2] < C_R$），这意味着 $\mathbb{E}[|R|] < \infty$，并确保该积分收敛到一个有限常数，该常数仅取决于C_R 和 σ。证明完毕。

□

C.3 定理1的证明

定理1的证明。根据三角不等式，我们将总误差分解为偏差误差和估计误差：

$$|\hat{V}_N(s) - V_N(s)| \leqslant |\hat{V}_N(s) - M(N)| + |M(N) - V_N(s)|$$

第一项由命题2界定，第二项由引理4界定。结合这些结果，我们得出，当 $m \geqslant c_1 \log(1/\delta)$时，以至少 $1 - \delta$ 的概率，

$$|\hat{V}_N(s) - V_N(s)| \leqslant c_2 \left(\sqrt{\log N} \sqrt{\frac{\log(1/\delta)}{m}} + (1 - \alpha/2)^N \right)$$

至此，证明完成。

□

D 定理2的证明

我们首先考虑本算法的选择误差。设 K小于 N，且 $\{s_1, \ldots, s_K\}$为具有真实回报水平 $\{V_N(s_i)\}_{i=1}^K$和估计量 $\{\hat{V}_N(s_i)\}_{i=1}^K$的$K$个候选状态。我们将每个候选项的估计误差定义为 $\epsilon_i = \hat{V}_N(s_i) - V_N(s_i)$，对所有 $i \in [K]$成立。

根据假设 2，我们知道，对于任意候选项 s，定理 1 适用于某些常数 c_1，c_2 它们取决于 $\sigma(s)$ α和 C_R。又因为 $\sigma(s) \in [\sigma_{min}, \sigma_{max}]$，存在仅取决于 c_1，c_2、σ_{min}，σ_{max}，α，C_R的统一常数，使得定理 1 对所有候选项均成立。

We 进而可以界定所有候选项中的最大估计误差的期望 s:

引理 5。 在假设 2下，假设样本量 m 满足 $m \geqslant c_0 \log K$，其中常数 $c_0 > 0$ 足够大。则当 N足够大时，有:

$$\mathbb{E}\left[\max_{i\in[K]} |\epsilon_i|\right] \leqslant c_1 \sqrt{\frac{\log N \log K}{m}}$$

其中 c_1 是一个取决于 σ_{min}、σ_{max}、α、M_μ、C_R.的常数

证明。 首先，我们根据定理 1推导出一个尾部概率界。该定理指出，对于任意满足条件 δ的 $m \geqslant c_1 \log(1/\delta)$ （等价于 $\delta \geqslant e^{-m/c_1}$），其成立概率至少为$1-\delta$:

$$|\epsilon_i| \leqslant c_2\big(\sqrt{\log N}\sqrt{\frac{\log(1/\delta)}{m}} + (1-\alpha/2)^N\big).$$

当 N 足够大时，偏差项 $(1-\alpha/2)^N$ 会被第一项支配，这意味着存在一个新的常数 c_2'，使得:

$$|\epsilon_i| \leqslant c_2'\sqrt{\log N}\sqrt{\frac{\log(1/\delta)}{m}}.$$

令 t 为 一个正阈值，则我们知道，当 $t \leqslant t_{max}$ 时，以下尾界成立 :

$$\mathbb{P}(|\epsilon_i| > t) \leqslant \exp\left(-\frac{mt^2}{c_2'^2 \log N}\right)$$

其中 $t_{max} = c_2'\sqrt{\log \frac{N}{c_1}}$。

我们基于这个最大的有效阈值定义良好事件区域。令 E 为这样一个事件：所有候选项的误差均处于定理的有效范围内:

$$E = \left\{\max_{i\in[K]} |\epsilon_i| \leqslant t_{max}\right\}$$

注意，补事件 E^C 对应于误差超出定理中样本量条件所涵盖范围的情况1。单个臂发生这种情况的概率至多为 $\delta_{min} = e^{-m/c_1}$。根据并集界:

$$\mathbb{P}(E^C) \leqslant Ke^{-m/c_1}$$

我们对期望进行分解:

$$\mathbb{E}\left[\max_{i\in[K]} |\epsilon_i|\right] = \mathbb{E}\left[\max_{i\in[K]} |\epsilon_i| \cdot \mathbb{I}(E)\right] + \mathbb{E}\left[\max_{i\in[K]} |\epsilon_i| \cdot \mathbb{I}(E^C)\right]$$

项1：良好事件上的期望。对于第一项，我们使用期望的积分恒等式。令 $M = \max_{i\in[K]} |\epsilon_i|$。

$$\begin{aligned}\mathbb{E}[M \cdot \mathbb{I}(E)] &= \int_0^{t_{max}} \mathbb{P}(M > t) dt \\ &\leqslant \int_0^{t_{max}} \min\left(1, \sum_{i=1}^{K} \mathbb{P}(|\epsilon_i| > t)\right) dt\end{aligned}$$

令 $\lambda = \frac{m}{c_2'^2 \log N}$。于是：

$$\mathbb{E}[M \cdot \mathbb{I}(E)] \leqslant \int_0^{\infty} \min\left(1, Ke^{-\lambda t^2}\right) dt$$

我们在 $t_0 = \sqrt{\log_\lambda K}$ 处对积分进行拆分。对于 $t \leqslant t_0$，被积函数有界于 1。对于 $t > t_0$，它有界于 $Ke^{-\lambda t^2}$。

$$\begin{aligned}\int_0^{\infty} \min\left(1, Ke^{-\lambda t^2}\right) dt &= t_0 + \int_{t_0}^{\infty} Ke^{-\lambda t^2} dt \\ &\leqslant \sqrt{\frac{\log K}{\lambda}} + \frac{K}{\lambda t_0} e^{-\lambda t_0^2} \times \frac{1}{2} \quad (\text{using } \int_x^{\infty} e^{-t^2} dt \leqslant \frac{e^{-x^2}}{2x}) \\ &= \sqrt{\frac{\log K}{\lambda}} + \frac{K}{\lambda\sqrt{\frac{\log K}{\lambda}}} \frac{1}{K} \cdot c \\ &\leqslant c\sqrt{\frac{\log K}{\lambda}} = c'\sqrt{\frac{\log N \log K}{m}}\end{aligned}$$

对于某个常数 c, c'，它取决于 $(\sigma_{min}, \sigma_{max}, \alpha, C_R)$。

项2：坏事件上的期望。对于第二项，我们使用柯西-施瓦茨不等式：

$$\mathbb{E}\left[\max_{i\in[K]} |\epsilon_i| \cdot \mathbb{I}(E^C)\right] \leqslant \sqrt{\mathbb{E}\left[\max_{i\in[K]} \epsilon_i^2\right]} \sqrt{\mathbb{P}(E^C)}$$

我们已经推导出 $\mathbb{P}(E^C) \leqslant K\exp(-m/c_1)$。因此，只需对 $\mathbb{E}\left[\max_{i\in[K]} \epsilon_i^2\right]$ 进行上界估计。并且我们知道：

$$\mathbb{E}\left[\max_{i\in[K]} \epsilon_i^2\right] \leqslant 2\mathbb{E}\left[\max_{i\in[K]} V_N(X_i)^2\right] + 2\mathbb{E}\left[\max_{i\in[K]} \hat{V}_N(X_i)^2\right]$$

对于第一项，我们有：

$$\begin{aligned}\mathbb{E}\left[\max_{i\in[K]} V_N(X_i)^2\right] &= \mathbb{E}\left[\max_{i\in[K]} (\mu_i + \sigma_i E(N))^2\right] \\ &\leqslant 2\mathbb{E}\left[\max_{i\in[K]} \mu_i^2\right] + 2E(N)^2\sigma_{max}^2 \\ &\leqslant 2K\mathbb{E}\left[\mu_i^2\right] + 2\sigma_{max}^2 \log N \\ &\leqslant c(K + \log N)\end{aligned}$$

对于某个常数 c 取决于 (σ_{max}, M_μ)。此处倒数第二个不等式成立，因为 $E(N) \leqslant \sqrt{2\log N}$，且最后一个不等式成立，因为根据假设 2，$\mu_i$ 是从一个分布中采样的，该分布具有有界二阶矩 $\mathbb{E}[\mu_i^2] \leqslant M_\mu$。

而对于第二项，我们有：

$$\begin{aligned}\mathbb{E}\left[\max_{i\in[K]}\hat{V}_N(X_i)^2\right] &\leqslant K\mathbb{E}\left[\hat{V}_N(X_i)^2\right]\\ &\leqslant 2K\mathbb{E}\left[\hat{\mu}_i^2\right]+2KE(N)^2\mathbb{E}\left[\hat{\sigma}_i^2\right]\end{aligned}$$

由于$\hat{}\sigma^2 = C\hat{}\sigma^2(D_{tail})$以及$\hat{}\mu = \hat{}\mu(D_{tail}) - C\hat{}\sigma$对于某个 C取决于$(\alpha, \sigma_{min}, \sigma_{max})$的常数，我们只需界定尾部数据集 D_{tail} 上经验均值和方差的矩。根据假设2，我们知道$\mathbb{E}[R^2] \leqslant C_R$，根据方差和均值的性质，有：

$$\mathbb{E}[\hat{\sigma}^2(D_{tail})] \leqslant C_R, \quad \mathbb{E}[\hat{\mu}(D_{tail})^2] \leqslant C_R$$

因此，存在一个 c' 取决于（$\sigma_{min}, \sigma_{max}, \alpha, C_R$）的常数，使得：

$$\mathbb{E}\left[\max_{i\in[K]}\hat{V}_N(X_i)^2\right] \leqslant c'K\log N$$

考虑到当 N足够大时， $K+\log\ N \leqslant 2K\log\ N$，我们知道存在一个 c取决于（$\sigma_{min}, \sigma_{max}, \alpha, C_R, M_\mu$）的常数，使得：

$$\mathbb{E}\left[\max_{i\in[K]}\epsilon_i^2\right] \leqslant cK\log N$$

因此：

$$\begin{aligned}\mathbb{E}\left[\max_{i\in[K]}|\epsilon_i|\cdot\mathbb{I}(E^C)\right] &\leqslant \sqrt{cK\log N}\cdot\sqrt{K}e^{-m/2c_1}\\ &= CK\sqrt{\log N}\exp(-C'm)\end{aligned}$$

其中 C, C' 取决于（$\sigma_{min}, \sigma_{max}, \alpha, C_R, M_\mu$）。
结合项1和项2可得：

$$\mathbb{E}\left[\max_{i\in[K]}|\epsilon_i|\right] \leqslant c\left(\sqrt{\frac{\log N\log K}{m}}+K\sqrt{\log N}\exp(-C'm)\right)$$

当 $m \geqslant c_0\log\ K$ 时，对于依赖于 $(\sigma_{min}, \sigma_{max}, \alpha, C_R, M_\mu)$ 的常数 c_0，我们知道第二项被第一项支配，因此有：

$$\mathbb{E}\left[\max_{i\in[K]}|\epsilon_i|\right] \leqslant c_1\sqrt{\frac{\log N\log K}{m}}$$

其中 c_1 是一个依赖于 $(\sigma_{min}, \sigma_{max}, \alpha, C_R, M_\mu)$ 的常数。证明至此完成。 □

现在我们可以开始证明定理 2。

定理 *2* 的证明。我们使用耦合论证。令 $\xi = (s_1, s_2, \dots)$ 为从参数先验分布中采样得到的状态的无限序列，并以 ξ 的一个固定实现为条件。以 ξ 为条件，我们知道，使用预算 N 的 Scaling-Law Guided（SLG）算法与使用预算 $\mathcal{A}^*$ 的预言机算法 $\lfloor N/2m\rfloor$ 所采样的状态相同，因为两个算法中都采样了 $K = \lfloor N/2m\rfloor$ 个候选项。将采样状态集记为 $S_K = \{s_1, s_2, \dots, s_K\}$，由于 $N \geqslant 4(m+1)$，我们知道 $N - Km \geqslant \lfloor N/2m\rfloor$。

那么，我们可以将预言机算法 $\mathcal{A}^*$ 的性能界定如下：

$$\begin{aligned}\mathbb{E}[V_{\lfloor N/2m\rfloor}(\mathcal{A}^*)\mid\xi] &= \max_{s\in S_K}V_{\lfloor N/2m\rfloor}(s)\\ &\leqslant \max_{s\in S_K}V_{N-Km}(s)\end{aligned}$$

我们将 $s^* = \arg\max_{s \in S_K} V_{N-Km}(s)$ 记为最优状态，将 $\hat{s} = \arg\max_{s \in S_K} \hat{V}_{N-Km}(s)$ 记为 SLG算法基于估计期望最大奖励选出的状态。然后，我们可以将预言机算法 $\mathcal{A}^*$ 与 SLG算法之间在给定ξ 条件下的差异界定如下：

$$\begin{aligned}\mathbb{E}[V_{\lfloor N/2m \rfloor}(\mathcal{A}^*) - V_N(\mathcal{A}_{\mathrm{SLG}}) \mid \xi] &\leqslant V_{N-Km}(s^*) - V_{N-Km}(\hat{s}) \\ &= V_{N-Km}(s^*) - \hat{V}_{N-Km}(s^*) + \hat{V}_{N-Km}(s^*) - \hat{V}_{N-Km}(\hat{s}) + \hat{V}_{N-Km}(\hat{s}) - V_{N-Km}(\hat{s}) \\ &\leqslant 2 \max_{s \in S_K} |\epsilon_s|\end{aligned}$$

其中 $\epsilon_s = \hat{V}_{N-Km}(s) - V_{N-Km}(s)$ 表示每个候选状态 s的估计误差。由于 $K = \lfloor N/2m \rfloor \leqslant N - Km$ 且 $m \geqslant c_0 \log K$，我们可以对两边关于 ξ取期望，并应用引理 5 得到：

$$V_{\lfloor N/2m \rfloor}(\mathcal{A}^*) - V_N(\mathcal{A}_{\mathrm{SLG}}) \leqslant \mathbb{E}\left[2 \max_{s \in S_K} |\epsilon_s|\right] \leqslant c_1 \sqrt{\frac{\log(N - Km) \log K}{m}} \leqslant c_1 \frac{\log N}{\sqrt{m}}$$

对于某个常数 c_1，它取决于（$\sigma_{min}, \sigma_{max}, \alpha, C_R, M_\mu$）。至此，证明完成。

□

E 正则性条件下的形式化遗憾分析

在本节中，我们将第 4.2.1 节中关于 SLG 在相同预算 $\mathcal{A}^*$ 下相对于预言机 N的消失遗憾的直觉形式化。首
e fir先介绍潜在参数分布所需的正则性条件。

假设 3。 对于由提示 x生成的随机状态 s，潜在参数$(\mu(s)\ \sigma(s))$满足

1. $\mu(s)$ 与 $\sigma(s)$ 相互独立。
2. 随机变量 $\sigma(s)$ 在其最大值附近有足够的概率质量： $P(\sigma(s) \geqslant \sigma_{\max} - \delta) \geqslant \delta^\alpha$ 其中某个 $\delta > 0$。
3. $\mu(s)$ 的期望最大值按以下方式缩放：

$$\left|\mathbb{E}[\max_{i \in [N]} \mu_i] - (\mu_0 + C_\mu (\log N)^\gamma)\right| \leqslant \mathcal{O}((\log N)^{-\beta})$$

其中常数为 μ_0、C_μ、$\beta > 0$，尾指数为 $\gamma \in (0,1)$。

我们还有定理 2的以下推论：

推论 2. 在假设 *2* 和假设 *3*下，对于充分大的 N，通过选择参数 $K \asymp N(\log N)^{-3}$和 $m \asymp (\log N)^3$,，我们有

$$V_N(\mathcal{A}^*) - V_N(\mathcal{A}_{SLG}) \leqslant \frac{c \log\log N}{(\log N)^{\min\{\beta, 1-\gamma, 1/2\}}}.$$

对于取决于模型参数的某个常数 $c > 0$ $(\alpha, \sigma_{\min}, \sigma_{\max}, C_R, M_\mu)$.

虽然假设 3 为分析简便起见，假定 μ（s） 与 σ（s） 相互独立，但这一条件可以放宽。正如下一节所给出的证明所示， $\mathcal{A}_{\mathrm{SLG}}$ 会主要选择方差接近 $\sigma_{\max}$ 的状态，随着 N 不断增大。因此，只需假设 μ（s）的尾部缩放性质（条件 3）在方差接近 $\sigma_{\max}$的条件下成立。直观而言，这仅要求在高方差区域内，均值与方差之间不存在强烈的负相关。

条件 (2) 和 (3) 是较宽松的正则性条件，许多标准分布都满足这些条件。例如，如果均值 $\mu(s)$ 服从高斯分布，而方差 $\sigma(s)$ 服从均匀分布，那么正则性条件在 $\alpha = 1, \gamma = 1/2$ 以及任意 $\beta > 0$ 下均成立。我们将在下一节给出推论 2 的证明。

E.1 推论2 的证明

为了帮助我们评估 $\mathcal{A}_{\mathrm{SLG}}$ 在探索宽度为 K 时与全信息预言机 $\mathcal{A}^*$ 之间的差异，我们首先定义一个中间预言机算法 $\mathcal{A}$ 。该算法知道每个候选状态 s 的真实期望最大回报 $V_N(s)$，但只会从预先采样的 K 个候选项中进行选择。

定义 3. 预言机$\hat{}\mathcal{A}$首先从提示 x中抽取一个包含 K个候选状态 $\mathcal{S}_K = \{s_1, \ldots, s_K\}$的池。随后，它利用完美的参数信息确定最优状态 $s^* = \mathrm{argmax}_{s\in\mathcal{S}_K} V_N(s)$ ，并将全部预算 N 用于从该状态进行采样。

$\hat{}$然后，我们可以先比较 $\mathcal{A}_{\mathrm{SLG}}$ 与 $\mathcal{A}$ 的性能，如下述引理所示。

引理 6. 在假设 2下，令 N 表示总采样预算，令 m 表示估计样本量。当我们选择的样本量按 $m = c_0 \left(\frac{N}{K}\right)^{2/3} \log(N)$ 的尺度增长时，其中 $c_0 > 0$为某个常数，$\hat{}$与预言机之间的性能差距 $\mathcal{A}_{SLG}$ 满足：$\mathcal{A}$

$$V_N(\hat{\mathcal{A}}) - V_N(\mathcal{A}_{SLG}) \leqslant C \left(\frac{K}{N}\right)^{1/3} \sqrt{\log N} \tag{11}$$

其中， C 是一个取决于$(\sigma_{min}, \sigma_{max}, \alpha, C_R, M_\mu)$的常数。

引理的证明 6。 我们采用与定理2的证明中相同的耦合论证。设 ξ =(s_1， s_2，.....) 为从参数先验分布中采样得到的无限状态序列，并且我们以 τ的一个固定实现为条件。

在以 ξ为条件的情况下，我们知道由Scaling-Law Guided（SLG）算法与$\hat{}$预言机算法 $\mathcal{A}$ 采样得到的状态是相同的，并将 s^* 记为 K 个候选项中具有最高真实期望最大回报的最优状态，即 s^* = arg $\max_{i\in[K]} V_N(s_i)$，将$\hat{s}$ 记为由$\hat{}$ SLG算法基于估计期望最大奖励选出的状态，即$\hat{s}$ = arg $\max_{i\in[K]} V_N(s_i)$。$\hat{}$于是，我们可以将SLG与预言机算法 $\mathcal{A}$ 之间的性能差距分解如下：

$$\begin{aligned} V_N(\hat{\mathcal{A}}) - V_N(\mathcal{A}_{\mathrm{SLG}}) &\leqslant \mathbb{E}[V_N(s^*) - V_{N-Km}(\hat{s})] \\ &= \mathbb{E}[V_N(s^*) - V_N(\hat{s})] + \mathbb{E}[V_N(\hat{s}) - V_{N-Km}(\hat{s})] \end{aligned}$$

其中该不等式成立，因为在以 τ为条件的情况下，SLG算法的期望奖励以所选状态$\hat{s}$ 在 $N - Km$ 个样本下的期望最大回报为下界。
对于第一项，使用与定理 2的证明相同的技巧，并应用引理 5，我们得到：

$$\mathbb{E}[V_N(s^*) - V_N(\hat{s})] \leqslant C_1 \sqrt{\frac{\log N \log K}{m}}$$

对于某个依赖于（σ_{min}， σ_{max}， α， C_R， M_μ）的常数 C_1 。
对于第二项，我们分析由样本量减少所导致的遗憾。利用缩放律性质，我们得到：

$$\begin{aligned} V_N(\hat{s}) - V_{N-Km}(\hat{s}) &= \sigma(\hat{s})(E(N) - E(N - Km)) \\ &\leqslant \sigma_{max}(E(N) - E(N - Km)) \end{aligned}$$

其中， $E(N)$ 是 N 个标准正态变量的期望最大值。由于上述不等式对任意固定的 τ 都成立，因此我们可以对两边关于 τ 取期望；又由于对于任意 $n > m \geqslant 3$， $E(n)-E(m) \leqslant 2(\sqrt{\log n}-\sqrt{\log m})$，我们有：

$$\begin{aligned} \mathbb{E}\left[V_N(\hat{s}) - V_{N-Km}(\hat{s})\right] &\leqslant 2\sigma_{max}(\sqrt{\log N} - \sqrt{\log(N - Km)}) \\ &\leqslant \sigma_{max} \frac{Km}{(N - Km)\sqrt{\log(N - Km)}} \end{aligned}$$

以及 co 考虑到 $N - Km \geqslant \frac{N}{2}$，存在一个依赖于 σ_{max} 的常数 C_2，使得 at:

$$\mathbb{E}\left[V_N(\hat{s}) - V_{N-Km}(\hat{s})\right] \leqslant C_2 \frac{K}{N\sqrt{\log N}} m$$

ˆ 合并这两部分可得，$\mathcal{A}$ 与 $\mathcal{A}_{\mathrm{SLG}}$ 之间的总性能差距有如下界：

$$V_N(\hat{\mathcal{A}}) - V_N(\mathcal{A}_{\mathrm{SLG}}) \leqslant C_1 \sqrt{\frac{\log N \log K}{m}} + C_2 \frac{Km}{N\sqrt{\log N}}$$

为使上界最小，我们选择样本量 m，使两项的阶相平衡，这表明应设 $m = c_0(N/K)^{2/3} \log N$，其中 $c_0 > 0$ 为某个常数。将此 m 代回不等式可得：

$$V_N(\hat{\mathcal{A}}) - V_N(\mathcal{A}_{\mathrm{SLG}}) \leqslant C \left(\frac{K}{N}\right)^{1/3} (\log N)^{1/2} \tag{12}$$

其中 C 为 一个依赖于问题参数的普适常数。证明完毕。

□

ˆ 接下来，只需将 $\mathcal{A}$ 的性能与全信息预言机 $\mathcal{A}^*$ 的性能进行比较，具体见下述引理。

引理 7. 在假设 2和假设3下，我们可以将全信息预言机 $\mathcal{A}^*$ 与受限预言机 ˆ$\mathcal{A}$ 之间的性能差距界定为 $K \asymp N(\log\ N)^{-3}$：

$$V_N(\mathcal{A}^*) - V_N(\hat{\mathcal{A}}) \leqslant C \frac{\log\log N}{(\log N)^{\min\{\beta, 1-\gamma\}}}.$$

其中，C 是一个取决于$(\sigma_{min}, \sigma_{max}, \alpha, C_R, M_\mu, C_\mu, \mu_0, \beta, \gamma)$的常数。

引理 7的证明。令 $\mathcal{S}_N$ 表示由预言机策略 $\mathcal{A}^*$采样的状态集合。我们将 $\mathcal{A}^*$ 的性能上界如下：

$$\begin{aligned}
V_N(\mathcal{A}^*) &= \mathbb{E}\left[\max_{s \in \mathcal{S}_N} V_N(s)\right] \\
&= \mathbb{E}\left[\max_{s \in \mathcal{S}_N} \left(\mu(s) + \sigma(s) E(N)\right)\right] \\
&\leqslant \mathbb{E}\left[\max_{s \in \mathcal{S}_N} \mu(s)\right] + E(N)\sigma_{\max} \\
&\leqslant \mu_0 + C_\mu (\log N)^\gamma + E(N)\sigma_{\max} + C(\log N)^{-\beta},
\end{aligned}$$

其中 $E(N)$ 是 N 个标准正态变量的期望最大值，最后一个不等式由假设 3得出。

为了建立ˆ$\mathcal{A}$的下界，我们考虑一个较弱的预言机算法ˆ $\mathcal{A}_\delta$，其中 $\delta \in (0,1)$，构造如下：

- $\hat{\mathcal{A}_\delta}$dra ws K 来自先验分布的状态 $\{s_1, \ldots, s_K\}$。它可以访问真实值 $\{V_N(s_i)\}_{i=}^{K}$ $_1$.
- 定义候选集 $\mathcal{S}_\delta = \{s_i \mid \sigma(s_i) \geqslant \sigma_{\text{最大}} - \delta,\ i \in [K]\}$。ˆ $\mathcal{A}_\delta$ 选择 s^* 在 $\mathcal{S}_\delta$ 中取值最大的状态，即 $s^* = \arg\max_{s \in \mathcal{S}_\delta} V_N(s)$。如果 $\mathcal{S}_\delta$ 为空，则任意选择一个状态。
- 该算法从 s^* 中采样 N 次，并输出最佳响应。

由于 $\hat{}\,\mathcal{A}_\delta$ 的搜索空间是 $\hat{}\mathcal{A}$ 所探索空间的子集（并附加了额外约束），因此我们有 $\hat{}\,\hat{}\,\hat{}\ V_N(\mathcal{A}) \geqslant V_N(\mathcal{A}_\delta)$。因此，只需给出 $\mathcal{A}_\delta$性能的下界。

根据假设 3，被采样状态的方差至少为 $\sigma_{\max} - \delta$ 的概率以 δ^α 为下界。这意味着 $\mathbb{E}[|\mathcal{S}_\delta|] \geqslant \delta^\alpha K$。定义事件 $E = \{|\mathcal{S}_\delta| \geqslant \frac{1}{2}\delta^\alpha K\}$。根据标准切尔诺夫界，我们有：

$$\mathbb{P}(E) \geqslant 1 - \exp(-C'\delta^\alpha K),$$

对某个常数 $C' > 0$。

由于在先验分布下 μ 与 σ 相互独立，对仅依赖于 σ 的 E 进行条件化，不会改变 $\mathcal{S}_\delta$ 中各状态的 μ 的分布。我们以有效样本量 $|\mathcal{S}_\delta|$ 调用均值缩放假设：

$$\mathbb{E}\left[\max_{s\in\mathcal{S}_\delta}\mu(s) \mid E\right] \geqslant \mu_0 + C_\mu\left(\log(\delta^\alpha K/2)\right)^\gamma - C\left(\log(\delta^\alpha K/2)\right)^{-\beta}.$$

在以 E 为条件时， $\mathcal{S}_\delta$ 中的每个状态均满足 $\sigma(s) \geqslant \sigma_{\max} - \delta$。我们对期望值给出下界：

$$\begin{aligned} V_N(\hat{\mathcal{A}}_\delta) &= \mathbb{E}[R_N(\hat{\mathcal{A}}_\delta)] \geqslant \mathbb{E}[R_N(\hat{\mathcal{A}}_\delta) \mid E]\mathbb{P}(E) \\ &\geqslant \left(\mu_0 + C_\mu(\log(\delta^\alpha K/2))^\gamma - C(\log(\delta^\alpha K/2))^{-\beta} + E(N)(\sigma_{\max} - \delta)\right)\left(1 - \exp(-C'\delta^\alpha K)\right). \end{aligned}$$

我们取 $\delta = (\log N)^{-2}$。回顾 $K = C_K N(\log N)^{-3}$，失败概率 $\exp(-C'\delta^\alpha K)$ 随 N 多项式衰减（具体而言，对某个 $c > 0$ 衰减为 $O(N^{-c})$），与多对数项相比可忽略不计。

设 $P = \frac{N}{\delta^\alpha K/2}$。我们有 $\log P = O(\log\log N)$ 和 $\log(\delta^\alpha K/2) = \log N - \log P$。对于足够大的$N$，利用不等式 $(1-x)^\gamma \geqslant 1 - x$ for $x \in [0,1)$ 和 $\gamma < 1$，下界可简化为：

$$\begin{aligned} V_N(\hat{\mathcal{A}}_\delta) &\geqslant \mu_0 + C_\mu(\log N)^\gamma\left(1 - \frac{\log P}{\log N}\right)^\gamma - C''(\log N)^{-\beta} + E(N)\sigma_{\max} - z_N\delta - O(N^{-c}) \\ &\geqslant \mu_0 + C_\mu(\log N)^\gamma\left(1 - \frac{\log P}{\log N}\right) - C''(\log N)^{-\beta} + E(N)\sigma_{\max} - \frac{z_N}{(\log N)^2} - O(N^{-c}). \end{aligned}$$

将其从 $V_N(\mathcal{A}^*)$ 的上界中减去，得到：

$$\begin{aligned} V_N(\mathcal{A}^*) - V_N(\hat{\mathcal{A}}) &\leqslant V_N(\mathcal{A}^*) - V_N(\hat{\mathcal{A}}_\delta) \\ &\leqslant C_\mu(\log N)^{\gamma-1}\log P + \frac{E(N)}{(\log N)^2} + C(\log N)^{-\beta} + O(N^{-c}). \end{aligned}$$

利用 $\log P = O(\log\log N)$ 和 $E(N) \leqslant \sqrt{2}\log N$ 这些事实，性能差距主要由以下项决定：

$$\begin{aligned} V_N(\mathcal{A}^*) - V_N(\hat{\mathcal{A}}) &\leqslant O\left(\frac{\log\log N}{(\log N)^{1-\gamma}}\right) + O\left(\frac{1}{(\log N)^{3/2}}\right) + O\left(\frac{1}{(\log N)^\beta}\right) \\ &\leqslant C\frac{\log\log N}{(\log N)^{\min\{\beta,1-\gamma\}}}, \end{aligned}$$

其中 $C > 0$ 是某个常数。 □

接下来，我们可以证明推论 2。

推论 2 的证明。 我们首先将 $\mathcal{A}^*$ 与 Scaling-Law Guided（SLG）算法之间的性能差距分解如下：

$$V_N(\mathcal{A}^*) - V_N(\mathcal{A}_{\text{SLG}}) = V_N(\mathcal{A}^*) - V_N(\hat{\mathcal{A}}) + V_N(\hat{\mathcal{A}}) - V_N(\mathcal{A}_{\text{SLG}})$$

由于 K， m 满足引理 6 中的条件，因此根据公式（12）可知：

$$\begin{aligned} V_N(\hat{\mathcal{A}}) - V_N(\mathcal{A}_{\text{SLG}}) &\leqslant C\left(\frac{K}{N}\right)^{1/3}(\log N)^{1/2} \\ &\leqslant C_1\frac{1}{(\log N)^{1/2}} \end{aligned}$$

对于某个常数 $C_1 > 0$。根据引理 7，我们有：

$$V_N(\mathcal{A}^*) - V_N(\hat{\mathcal{A}}) \leqslant C_2 \frac{\log\log N}{(\log N)^{\min\{\beta, 1-\gamma\}}}$$

对于某个常数 $C_2 > 0$。综合这两部分，我们得出：

$$\begin{aligned} V_N(\mathcal{A}^*) - V_N(\mathcal{A}_{\mathrm{SLG}}) &\leqslant C_2 \frac{\log\log N}{(\log N)^{\min\{\beta, 1-\gamma\}}} + C_1 \frac{1}{(\log N)^{1/2}} \\ &\leqslant C \frac{\log\log N}{(\log N)^{\min\{\beta, 1-\gamma, 1/2\}}} \end{aligned}$$

for 某个依赖于问题参数的常数 $C > 0$ 。至此完成该证明 f.

□

F 4.2.2节中的证明

F.1 命题1的证明

我们分析同质噪声设定下 SLG 与 Best-of-N 算法之间的性能差距，如命题1所述。

命题*1*的证明。 我们首先在同质噪声设定下分析我们的算法 SLG 与 Best-of-N 算法的性能。

如果我们将 S_{SLG} 表示为 SLG算法 所选择的状态集合（其中我们知道 $|S_{\mathrm{SLG}}| = K$），并将 $\hat{s} = \arg\max_{s \in S_{\mathrm{SLG}}} \hat{V}_N(s)$ 表示为所选状态中估计值最高的候选项，那么由于对于所有 s，方差 $\sigma(s) = \sigma_1$，我们有：

$$\hat{s} = \arg\max_{s \in S_{\mathrm{SLG}}} \hat{V}_N(s) = \arg\max_{s \in S_{\mathrm{SLG}}} \hat{\mu}(s)$$

因此，我们可以如下界定 SLG算法 的期望性能下界：

$$\begin{aligned} V_N(\mathcal{A}_{\mathrm{SLG}}) &\geqslant \mathbb{E}[V_{N-Km}(\hat{s})] \\ &= \mathbb{E}[\mu(\hat{s}) + \sigma_1 E(N - Km)] \\ &= \mathbb{E}[\mu(\hat{s})] + \sigma_1 E(N - Km) \end{aligned}$$

由于我们知道 $\hat{\mu}(s) \mid \mu(s) \sim \mathcal{N}(\mu(s), \sigma_1^2/m)$ 且 $\mu(s) \sim \mathcal{N}(\mu_0, \sigma_0^2)$，根据高斯分布的性质，我们有：

$$\begin{aligned} \mathbb{E}[\mu(\hat{s})] &= \mathbb{E}_{\hat{\mu}}\left[\mathbb{E}[\mu(\hat{s}) \mid \hat{\mu}(\hat{s})]\right] \\ &= \mathbb{E}_{\hat{\mu}}\left[\mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2/m}(\hat{\mu}(\hat{s}) - \mu_0)\right] \\ &= \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2/m}\mathbb{E}_{\hat{\mu}}[\hat{\mu}(\hat{s}) - \mu_0] \\ &= \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2/m}\sqrt{\sigma_0^2 + \sigma_1^2/m}E(K) \end{aligned}$$

其中，最后一个等式来自以下事实：对于所有 s，$\hat{\mu}(s) \sim \mathcal{N}(\mu_0, \sigma_0^2 + \sigma_1^2/m)$，且 $\hat{s}$ 是 K 个独立同分布样本中的最大值项。若记 $t = \frac{\sigma_1}{\sigma_0}$，则可将上述表达式简化为：

$$\mathbb{E}[\mu(\hat{s})] = \mu_0 + \frac{\sigma_0}{\sqrt{1 + t^2/m}}E(K)$$

由此可得 SLG算法 的以下下界：

$$V_N(\mathcal{A}_{\mathrm{SLG}}) \geqslant \mu_0 + \sigma_0 \left[\frac{1}{\sqrt{1+t^2/m}} E(K) + tE(N-Km) \right]$$

对于 Best-of-N 算法，由于提示 x 的奖励分布为 $\mathcal{N}(\mu_0,\ \sigma_0^2+\sigma_1^2)$，我们有：

$$V_N(\mathcal{A}_{\mathrm{BoN}}) = \mathbb{E}[\max_{R_1,\ldots,R_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0,\sigma_0^2+\sigma_1^2)} R_i] = \mu_0 + \sigma_0\sqrt{1+t^2}E(N)$$

因此，两个算法之间的性能差距可以下界估计为：

$$V_N(\mathcal{A}_{\mathrm{SLG}}) - V_N(\mathcal{A}_{\mathrm{BoN}}) \geqslant \sigma_0 \left[\frac{1}{\sqrt{1+t^2/m}} E(K) + tE(N-Km) - \sqrt{1+t^2}E(N) \right]. \tag{13}$$

接下来，我们定义一个替代函数来近似该差距，并推导出该替代函数的下界。令：

$$G(t,N) = \frac{1}{\sqrt{1+t^2/m}}\sqrt{2\log K} + t\sqrt{2\log(N-Km)} - \sqrt{1+t^2}\sqrt{2\log N}.$$

通过选择 $m = t^2 \log N$ 和 $K = \frac{N}{(1+t)m} = \frac{N}{1+t\ t^2\ ()\log N}$，我们有：

$$\begin{aligned}
\frac{1}{\sqrt{1+t^2/m}}\sqrt{2\log K} &= \frac{1}{\sqrt{1+1/\log N}}\sqrt{2\log N - 2\log\log N - 2\log(t^2(1+t))} \\
&\geqslant \frac{1}{\sqrt{1+1/\log N}}\sqrt{2\log N}\left(1 - \frac{\log\log N + \log(t^2(1+t))}{\log N}\right) \\
&\geqslant \frac{1}{\sqrt{1+1/\log N}}\sqrt{2\log N} - C_1\frac{\log\log N}{\sqrt{\log N}}
\end{aligned}$$

对于某个取决于 t 的常数 $C_1 > 0$。类似地，我们有：

$$\begin{aligned}
t\sqrt{2\log(N-Km)} &= t\sqrt{2\log N - 2\log(1+1/t)} \\
&\geqslant t\sqrt{2\log N}\left(1 - \frac{\log(1+1/t)}{\log N}\right) \\
&\geqslant t\sqrt{2\log N} - C_2\frac{1}{\sqrt{\log N}}
\end{aligned}$$

对于某个取决于 t 的常数 $C_2 > 0$。结合这些部分可得：

$$G(t,N) \geqslant \left(\frac{1}{\sqrt{1+1/\log N}} + t - \sqrt{1+t^2}\right)\sqrt{2\log N} - C\frac{\log\log N}{\sqrt{\log N}}$$

对于某个依赖于 t 的常数 $C > 0$。当 $N \geqslant \exp(1+1/t)$ 时，我们观察到：

$$\begin{aligned}
\frac{1}{\sqrt{1+1/\log N}} + t - \sqrt{1+t^2} &\geqslant \frac{1}{\sqrt{1+\frac{t}{t+1}}} + t - \sqrt{1+t^2} \\
&\geqslant \left(1 - \frac{1}{2}\frac{t}{t+1}\right) + t - \sqrt{1+t^2} \\
&\geqslant \frac{t}{t+1} - \frac{1}{2}\frac{t}{t+1} = \frac{1}{2}\frac{t}{t+1},
\end{aligned}$$

其中，最后一个不等式源于以下事实：对于所有 $t > 0$，$t + 1 - \sqrt{1+t^2} \geqslant \frac{t}{t+1}$。因此：

$$G(t,N) \geqslant \frac{1}{2}\frac{t}{t+1}\sqrt{2\log N} - C\frac{\log\log N}{\sqrt{\log N}}$$

对于 $N \geqslant \exp(1+1/t)$。
只需分析原始差距与替代函数之间的近似误差，该误差源于用 $\sqrt{2\log N}$ 近似 $E(N)$。利用已知的界：

$$\left|E(N)-\sqrt{2\log N}\right| \leqslant C_3 \frac{\log\log N}{\sqrt{\log N}}$$

对于某个常数 $C_3>0$，以及所有 $N \geqslant 3$，我们有：

$$\begin{aligned}
&|V_N(\mathcal{A}_{\mathrm{SLG}})-V_N(\mathcal{A}_{\mathrm{BoN}})-\sigma_0 G(t,N)| \\
&\leqslant \sigma_0\left(\frac{1}{\sqrt{1+t^2/m}}\left|E(K)-\sqrt{2\log K}\right|+t\left|E(N-Km)-\sqrt{2\log(N-Km)}\right|+\sqrt{1+t^2}\left|E(N)-\sqrt{2\log N}\right|\right) \\
&\leqslant \sigma_0\left(C_3\frac{\log\log K}{\sqrt{\log K}}+tC_3\frac{\log\log(N-Km)}{\sqrt{\log(N-Km)}}+\sqrt{1+t^2}C_3\frac{\log\log N}{\sqrt{\log N}}\right) \\
&\leqslant C_4\frac{\log\log N}{\sqrt{\log N}}
\end{aligned}$$

对于某个依赖于 t 的常数 $C_4>0$。
结合上述结果，我们发现，下界由一个阶为$\sqrt{\log N}$ 的主导项和一个阶为 $\frac{\log\log N}{\sqrt{\log N}}$ 的负误差项组成。由于 $\lim_{N\to\infty}\frac{\log\log N}{\sqrt{\log N}}=0$，存在一个阈值$N_0$（取决于 t），使得对于所有 $N \geqslant N_0$，误差项至多为主导项的一半。因此，对于 $N \geqslant N_0$：

$$V_N(\mathcal{A}_{\mathrm{SLG}})-V_N(\mathcal{A}_{\mathrm{BoN}}) \geqslant \frac{t}{4(t+1)}\sigma_0\sqrt{2\log N}$$

证明至此完成。 □

F.2 推论的证明1

现在我们证明 推论 1 ： 通过将性能差距转化为等效的预算放大 .

推论 1的证明。 我们旨在找到一个常数 $\gamma>0$，使得对于 $N \geqslant N_0$（如命题 1中所定义），不等式 $V_N(\mathcal{A}_{\mathrm{SLG}}) \geqslant V_{N^{1+\gamma}}(\mathcal{A}_{\mathrm{BoN}})$ 成立。
根据命题 1，SLG算法的性能严格高于标准 Best-of-N 基线，且差距为正：

$$V_N(\mathcal{A}_{\mathrm{SLG}}) \geqslant V_N(\mathcal{A}_{\mathrm{BoN}})+\frac{t}{4(t+1)}\sigma_0\sqrt{2\log N}. \tag{14}$$

现在，考虑使用放大后的预算 $N^{1+\gamma}$运行的 Best-of-N 基线。其期望值为：

$$V_{N^{1+\gamma}}(\mathcal{A}_{\mathrm{BoN}})=\mu_0+\sigma_0\sqrt{1+t^2}E(N^{1+\gamma}).$$

放大基线与标准基线之间的值差为 ：

$$\begin{aligned}
\Delta V_{\mathrm{BoN}} &= V_{N^{1+\gamma}}(\mathcal{A}_{\mathrm{BoN}})-V_N(\mathcal{A}_{\mathrm{BoN}}) \\
&= \sigma_0\sqrt{1+t^2}\left(E(N^{1+\gamma})-E(N)\right).
\end{aligned}$$

由于我们知道，对于任意 $n>m \geqslant 3$，$E(n)-E(m) \leqslant 2(\sqrt{\log n}-\sqrt{\log m})$，因此可以进一步界定该差值：

$$\begin{aligned}
E(N^{1+\gamma})-E(N) &\leqslant 2(\sqrt{\log(N^{1+\gamma})}-\sqrt{\log N}) \\
&= 2(\sqrt{1+\gamma}-1)\sqrt{\log N}.
\end{aligned}$$

将其代回 ΔV_{BoN} 的表达式中，得到：

$$\Delta V_{\mathrm{BoN}} \leqslant 2\sigma_0\sqrt{1+t^2}(\sqrt{1+\gamma}-1)\sqrt{\log N} \leqslant \sigma_0\sqrt{1+t^2}\gamma\sqrt{\log N}.$$

而当我们选择 $\gamma = \frac{\sqrt{2t}}{4}\frac{}{t+1}$ 时，有：

$$\begin{aligned} V_N(\mathcal{A}_{\mathrm{SLG}}) &\geqslant V_{N^{1+\gamma}}(\mathcal{A}_{\mathrm{BoN}}) - \Delta V_{\mathrm{BoN}} + \frac{t}{4(t+1)}\sigma_0\sqrt{2\log N} \\ &\geqslant V_{N^{1+\gamma}}(\mathcal{A}_{\mathrm{BoN}}) - \sigma_0\sqrt{1+t^2}\gamma\sqrt{\log N} + \frac{t}{4(t+1)}\sigma_0\sqrt{2\log N} \\ &= V_{N^{1+\gamma}}(\mathcal{A}_{\mathrm{BoN}}) \end{aligned}$$

证明至此完成。

□