arXiv:2602.01603v1 [stat.ML] 2 Feb 2026

Infe通过非线性 GRPO 实现大语言模型的推理感知元对齐

Shokichi Takakura [1] **Akifumi Wachi** [1] **Rei Higuchi** [2 3] **Kohei Miyaguchi** [1] **Taiji Suzuki** [2 3]

摘要

将大语言模型（LLM）与多样的人类偏好对齐具有根本性的挑战，因为不同准则之间往往会相互冲突。推理时对齐方法近来广受欢迎，因为它们能够在推理时通过不同的对齐算法，使大语言模型符合多重准则。然而，推理时对齐的计算开销很大，因为它通常需要对基础模型进行多次前向传播。本文提出推理感知元对齐（IAMA），这是一种新方法，能够在推理时计算预算有限的情况下，使大语言模型符合多重准则。IAMA 训练基础模型，使其能够通过不同的推理时对齐算法有效地对齐多个任务。为解决 IAMA 涉及的非线性优化问题，我们提出非线性 *GRPO*，并证明该方法能够在概率测度空间中收敛到最优解。

1. 引言

通过基于人类反馈的强化学习（RLHF）（Christiano 等人，2017；Ouyang 等人，2022）或可验证奖励（RLVR）（Lambert 等人，2024；Shao 等人，2024；Jaech 等人，2024）对大语言模型（LLM）进行对齐，是确保其生成符合人类偏好或特定标准的响应的关键步骤。然而，在现实应用中，大语言模型响应的要求往往因情境而异，且多种多样，无法通过单一奖励函数来涵盖。已有多项研究探索了同时将大语言模型与多个奖励函数对齐的思路（Agnihotri 等人，2025；Dai 等人，2024），但这从根本上说颇具挑战，甚至可能无法实现，因为不同标准往往彼此冲突（Bai 等人，2022）。

最近，推理时对齐这一范式日益流行，其中基础模型在推理时使用诸如最佳 N 选一（BoN）采样（Nakano 等人，2021）、BoN 的变体（Verdun 等人，2025；Jinnai 等人，2025）以及自一致性（Wang 等人，2023）。由于这些推理时对齐算法可以在推理时根据不同的奖励函数或标准修改基础模型的输出，模型便能够对齐到多样化偏好。然而，推理时对齐的计算开销很大，因为它通常需要在推理时对基础模型进行多次前向传播。在计算预算有限的情况下，如果基础模型未经适当训练，推理时对齐的性能可能会很差。例如，若将多个奖励聚合为单一指标并对模型进行对齐，得到的模型会在所有奖励上仅产生平庸的响应，从而导致多样性丧失。因此，使用 BoN 等推理时对齐方法很难采样出针对特定奖励定制的响应。于是，我们提出如下问题：我们能否训练出一个可通过推理时对齐高效对齐到多样化偏好的模型？

为了解决这一问题，我们提出推理感知元对齐（IAMA），这是一种新颖且通用的框架，用于通过两阶段对齐过程将大语言模型对齐到多个奖励函数：1）元训练基础模型，使其使用推理感知目标函数；以及 2）针对不同奖励函数进行推理时对齐。我们表明，两阶段对齐过程可被视为概率测度空间中的模型无关元学习（MAML）（Finn 等人，2017），且该元学习问题可表述为关于策略分布的非线性优化问题。这与标准的期望奖励最大化问题形成对比，后者相对于策略分布是线性的。由于目标函数的非线性，我们无法直接应用现有的策略优化算法，例如信赖域策略优化（Schulman 等人，2015）、近端策略优化（Schulman 等人，2017）和 GRPO（Shao 等人，2024）来求解该问题。为解决这一问题，我们提出一种新的算法，称为非线性 *GRPO*，它在每次迭代中利用一阶变分将目标函数线性化。尽管大语言模型的优化景观通常高度非凸，但我们在概率测度空间中的目标函数对于某些推理时对齐算法（如 BoN）被证明是凸的。利用这一性质，我们证明了所

[1]LY Corporation [2]东京大学 [3]理化学研究所人工智能项目。通讯请联系：Shokichi Takakura <stakakur@lycorp.co.jp>。

预印本。*2026*年*2*月*3*日。

提算法收敛至最优解。实验表明，所提方法能够有效地同时将大语言模型对齐到多个奖励函数。

我们的贡献可概括如下：

- 我们提出了推理感知元对齐（IAMA），这是一种新颖且通用的框架，通过优化单个基础模型，使其能够通过推理时对齐有效适应多样化偏好。这可以视为在概率测度空间中对著名 MAML 框架的推广。
- 我们证明，IAMA 可以表述为测度空间中的一个非线性优化问题。为求解该问题，我们开发了非线性 *GRPO*，它可视为相对于 KL 散度的镜像下降。
- 理论上，我们证明了 BoN 型推理时对齐方法的目标函数具有凸性，并表明非线性 GRPO 会线性收敛至最优解。我们还通过多个任务和模型的实验，展示了所提方法的有效性。尽管实现过程直接明了，该方法仍显著推动了多目标 RLHF 任务中的帕累托前沿。

1.1. 其他相关工作

推理感知对齐 近期有多项研究探索了将推理时对齐算法纳入训练目标的思路（Chow 等人，2025；Tang 等人，2025；Balashankar 等人，2025）。另有一类研究考察了 pass@k 或 max@k 目标函数，这相当于在推理时应用 BoN 采样（Walder & Karkhanis，2025；Bagirov 等人，2025）。然而，这些研究通常只关注单一的推理时对齐算法和单一奖励函数。因此，如何通过两阶段对齐，同时使大语言模型适配多个奖励函数，仍不明确。此外，这些研究仅考虑 BoN 采样，并未提供能够处理各种推理时对齐算法的通用框架。唯一的例外是 InfAlign 框架（Balashankar 等人，2025），但它只关注单一的推理时对齐算法。此外，尽管其问题表述可应用于一般的推理时对齐算法，但所提出的 *InfAlign-CTRL* 方法仅适用于 BoN 类对齐算法，并将奖励函数限制为相对于参考模型的胜率。因此，即便针对单一推理时对齐算法，如何解决一般的推理感知对齐问题仍不明确。更多讨论请参见附录 B。

元学习 元学习（Thrun 与 Pratt，1998）旨在训练模型，使其能够通过少样本学习或快速适应迅速适应新任务。元学习中最流行的方法之一是 MAML（Finn 等人，2017），它学习模型参数，使模型能够通过少数几个梯度步有效适应多个任务。然而，推理时对齐过程不会更新模型参数，因此无法直接由原始 MAML 框架处理。MetaAlign（Zhang 等人，2025）通过让模型以指定偏好的系统提示为条件，实现动态对齐。该方法在显式指令遵循方面效果良好，但可能难以捕捉无法通过提示轻易表达的复杂或隐式偏好。

概率测度空间中的优化概率测度空间中的优化问题出现在强化学习、变分推断和生成建模等各种机器学习问题中（Chu 等人，2019）。一系列研究探索了基于粒子的方法，例如平均场朗之万动力学和粒子对偶平均（Mei 等人，2018；Nitanda 等人，2021）。然而，这些方法在每次重新采样时都需要重新训练。另一系列研究探讨了概率测度空间中的镜像下降（Aubin-Frankowski 等人，2022；Yao 等人，2024）或对偶平均（Kawata 等人，2025）方法。尽管这些方法在理论上较为直接，但它们侧重于易于处理的设置，在这些设置中，一阶变分可以用闭式形式计算，或者计算过程包含计算开销高昂的步骤，例如归一化流（Yao 等人，2024）。因此，如何求解大语言模型对齐中涉及的非线性优化问题仍不明确。

2. 问题设定

我们将语言模型表示为给定上下文 x 时响应 y 的概率分布（或策略）π。我们分别将所有可能的上下文和响应的空间记为 $\mathcal{X}$ 和 $\mathcal{Y}$。假设上下文从 $\mathcal{X}$ 上的分布 ρ 中采样。我们将所有可能的策略的空间定义为 $\mathcal{P}$，并将参考策略定义为 $\pi_{\text{ref}} \in \mathcal{P}$。在本文中，如果符号不会引起歧义，为简洁起见，我们通常省略其对上下文 x 的依赖。例如，我们写作 $\pi(y)$，而不是 $\pi(y \mid x)$。

训练时对齐 设 $r\colon \mathcal{X}\times\mathcal{Y} \to [0, r_{\max}]$ 为一个奖励函数，用于衡量在给定上下文 x的情况下响应y 的质量。对齐目标是找到一个策略 $\pi(y \mid x)$，使期望奖励最大化，同时与参考策略 π_{ref}保持接近。通常，该目标被表述为 KL散度正则化的期望奖励

最大化问题：

$$\mathcal{R}[\pi] = \mathbb{E}_{y\sim\pi(\cdot|x),x\sim\rho}\left[r(x,y)\right] - \beta\mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}],$$

其中，KL 表示 Kullback-Leibler 散度，定义为 KL $[\pi \mid \pi_{\mathrm{ref}}] = \mathbb{E}_{y\sim\pi(\cdot|x),\ x\sim\rho}\left[\log\ {}^{\pi_{(y|x)}}\!/_{\pi_{\mathrm{ref}}(y|x)}\right]$，而 $\beta \geq 0$ 是一个正则化参数。

推理时对齐 我们在推理时很少直接使用对齐模型。相反，我们通常会采用推理时对齐算法，在推理时修改基础模型 π 的输出。最流行的方法之一是最佳 N 次采样（Nakano 等人，2021）：从基础模型 π 中采样多个响应，并选择奖励最高的响应。Soft BoN（Verdun 等人，2025）是 BoN 的一种软化版本，它从奖励的指数倾斜分布中近似采样响应。这些方法等价于将基础模型 π 转换为一个新策略，并从转换后的策略中采样响应。我们将此类推理时对齐算法表示为一个函数$\mathcal{T}$：$\mathcal{P} \to \mathcal{P}$，它将基础模型 $\pi \in \mathcal{P}$ 映射为对齐模型 $\mathcal{T}[\pi]$。对于含有 N 个样本的 BoN，转换后的策略表示为 $N \cdot C[\pi](y)^{N-1}\pi(y)$（Beirami 等人，2025；Gui 等人，2024；Amini 等人，2025），其中$C[\pi](y) = \int_{r(y')\leq r(y)} \pi(y')\,\mathrm{d}y'$ 是在 π 下奖励的累积分布函数；对于温度参数为 $\tau > 0$ 的 Soft BoN，其表示为$\mathcal{T}[\pi](y) \propto \exp(r(y)/\tau)\pi(y)$（Verdun 等人，2025）。除（Soft）BoN 外，还有多种推理时对齐算法，例如自一致性（Wang 等人，2023）和一种基于 MCMC 的方法（Karan 和 Du，2025）；我们的通用框架也能处理这些算法。

3. 提出框架：推理感知元对齐

如上一节所述，现代对齐流程通常包含两个阶段：1）对基础模型 π 进行训练时对齐；2）通过算法 $\mathcal{T}$ 进行推理时对齐。考虑到存在多种推理时对齐算法 $\{\mathcal{T}_i\}_{i=1}^m$ 及其对应的奖励函数 $\{r_i\}_{i=1}^m$，一个自然的问题是：什么样的模型 π 能够通过推理时对齐算法有效地适应多种奖励函数？为了解决这一问题，我们提出以下训练时对齐目标函数：

$$\mathcal{R}[\pi] = \underbrace{g(R_1[\pi],\ldots,R_m[\pi])}_{R[\pi]} - \beta\mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}], \quad (1)$$

其中 $R_i[\pi] = \mathbb{E}_{y\sim\mathcal{T}_i[\pi],x\sim\rho}\left[r_i(x,y)\right]$, 且 $g: \mathbb{R}^m \to \mathbb{R}$是一个聚合函数。聚合函数 g 可以是加权和 $\sum_{i=1}^m w_i R_i[\pi]$，其中权重为 $w_i \geq 0$,

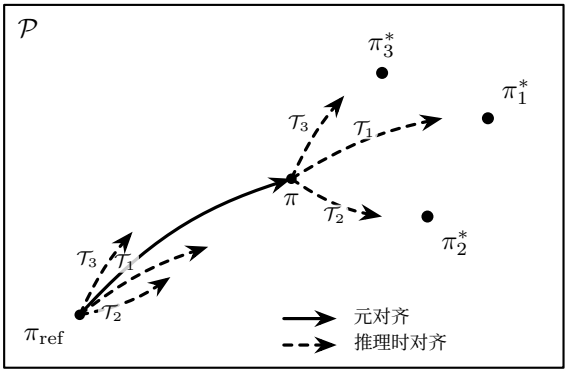


图*1.* 推理感知元对齐（IAMA）示意图。基础模型 π_{ref} 经过元训练，使其能够通过推理时对齐算法 $\{\mathcal{T}_i\}_{i=1}^m$有效适应多个任务最优解 $\{\pi_i^*\}_{i=1}^m$。与直接部署 π_{ref} 相比，IAMA 能够在推理时计算预算有限的情况下，为每项任务获得更好的对齐模型。

最坏情况性能 最小值$_i$ $R_i[\pi]$，或平滑最小值$-\frac{1}{\gamma}\log(\sum_{i=1}^m w_i \exp(-\gamma R_i[\pi]))$，其中 $\gamma > 0$，它在平均性能与最坏情况性能之间进行插值。该目标函数可视为对 InfAlign 框架的推广（Balashankar 等人，2025），可应用于多种推理时对齐算法，以及超越相对于参考模型的胜率的通用奖励函数。

在模型无关元学习（MAML）的背景下也探讨过类似的动机（Finn 等人，2017），其目标是找到能够通过少量梯度步有效适应多个任务的参数。另一方面，在推理时对齐中，适应过程中不会更新模型参数，而是将基础模型 π 转换为概率测度空间中的新策略 $\mathcal{T}_i[\pi]$。我们在图 1 中展示了所提出框架的概念。

虽然这种表述看起来与标准的 KL散度正则化期望奖励最大化相似，但关键区别在于，即使采用加权和，奖励项 $R[\pi]$ 相对于基础模型 π 仍是非线性的，这是由推理时对齐算法 $\{\mathcal{T}_i\}_{i=1}^m$的存在所致。这与期望奖励 $E_\pi[r]$不同，后者相对于 π是线性的。由于这种非线性，我们需要开发新的算法来求解该优化问题。

3.1. 与朴素方法的比较

一种将大语言模型与多个奖励函数对齐的朴素方法，是训练一个基础模型 π，通过加权和最大化期望奖励，而不考虑推理时对齐。也就是说，我们优化以下目标函数：

$$\mathcal{R}_{\mathrm{naive}}[\pi] = \sum_{i=1}^m w_i \mathbb{E}_{y\sim\pi,x\sim\rho}\left[r_i(x,y)\right] - \beta\mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}].$$

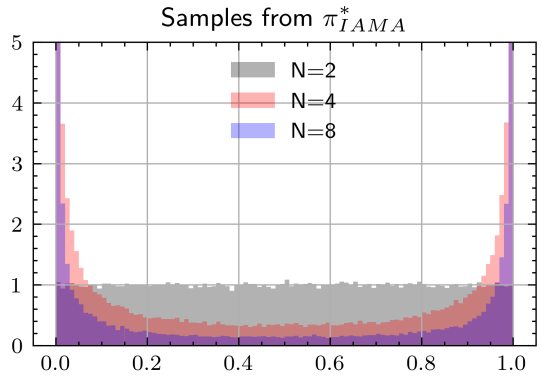


图 2. 采用 BoN 采样的最优 IAMA 策略 π^*_{IAMA}（$N = 2$、4、8）。具有较大 N 的最优策略 π^*_{IAMA} 会产生多样化输出，以覆盖 $y = 0$ 和 $y = 1$ 处的两种模式。

为了看清朴素方法与 IAMA 之间的差异，我们考虑一个简单的情形，其中 $\mathcal{Y} = [0, 1]$。假设存在两个奖励函数：

$$r_1(y) = 1 - y^2, r_2(y) = 1 - (1 - y)^2.$$

由于奖励 r_1 在 $y = 0$ 处达到最大值，而奖励 r_2 在 $y = 1$ 处达到最大值，因此这两个奖励完全冲突。如果我们采用具有相等权重 $w_1 = w_2 = 0.5$ 的朴素目标函数，并为简单起见忽略 KL 正则化，则最优策略为 $\pi^*_{\text{naive}} = \delta_{0.5}$，它始终输出 $y = 0.5$。其中，δ_a 是以 a 为中心的狄拉克 delta 分布。另一方面，如果我们考虑通过 BoN 采样进行推理时对齐，并令 $N \geq 2$，则可得到以下结果：

命题 3.1。考虑上述奖励函数 r_1和 r_2。设 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 是采用 $N \geq 2$ 进行 *BoN* 采样的策略。则最大化 *IAMA* 目标函数 (1)且权重相等 $w_1 = w_2 = 0.5$和 $\beta = 0$时的最优策略为

$$\pi^*_{\text{IAMA}}(y) = \frac{\alpha y^{\alpha-1}(1-y)^{\alpha-1}}{(y^\alpha + (1-y)^\alpha)^2}$$

其中 $\alpha = \frac{1}{N-1}$。

证明见附录G.1 。我们在图 π^*_{IAMA} 中展示了 $N = 2$、4、8 时的最优策略 2。在 BoN 的情况下，基础策略会生成多样化输出，以覆盖 $y = 0$和 $y = 1$ 两种模式，从而使推理时对齐能够针对两个奖励函数选出高奖励输出。这与朴素的最优策略形成对比；后者π^*_{naive} 只会生成平庸的输出。此外，最优策略会随推理时的计算预算（即样本数量 N）而变化。这个简单的例子清楚地回答了为何应考虑 IAMA。

注释 3.2。另一种可行的方法是将多个模型 $\{\pi_i\}_{i=1}^m$ 分别与每个奖励函数 $\{r_i\}_{i=1}^m$ 对齐并在推理时根据用户偏好选择其中一个模型（Barrett & Narayanan，2008；Rame 等人，2023）。然而，这种方法需要存储多个模型，对于现实应用中使用的大型模型而言可能并不实际。此外，即使在推理时，提供方也可能不知道用户的偏好。在这种情况下，无法采用这种方法；而 IAMA 仍然可用，因为用户可以根据自己的偏好选择最佳响应。

4. 所提方法：非线性 GRPO

由于式（1）中的目标函数相对于策略 π呈非线性，因此我们无法直接应用现有的策略优化算法，例如 TRPO（Schulman 等人，2015）、PPO（Schulman 等人，2017）和 GRPO（Shao 等人，2024）。为解决这一问题，我们提出了一种非线性 GRPO 算法，这是对 GRPO 类算法（Shao 等人，2024；Zheng 等人，2025；Liu 等人，2025）的一种非平凡扩展，这类算法会针对给定的上下文采样多个响应。

从另一个角度来看，Balashankar 等人（2025）提出了 InfAlign-CTRL 算法，通过奖励变换解决式（1）的一个特殊情形，但它仅适用于以下情况：1）BoN 类算法；2）胜率最大化；3）单奖励设定。由于只有在这些限制条件下才能计算这种变换，因此这一限制不可避免。我们的方法消除了这三项限制，因此具有独立的研究价值。

首先，我们重新审视标准的TRPO类算法（包括信赖域策略优化、近端策略优化、GRPO及其变体）。为符号方便起见，我们考虑$\mathcal{L}[\pi]$ 的最小化：$= -\mathcal{R}[\pi]$ ，而非 $\mathcal{R}[\pi]$的最大化。令 π_t为第 t次迭代时的当前策略。于是，TRPO类算法可被解释为概率分布空间中如下镜像下降法的一种近似：

$$\pi_{t+1} = \operatorname*{argmin}_\pi \tilde{\mathcal{L}}[\pi] + \frac{1}{\eta}\mathrm{KL}[\pi \mid \pi_t],$$
$$\tilde{\mathcal{L}}[\pi] := -\mathbb{E}_{y\sim\pi_t}\left[\frac{\pi(y)}{\pi_t(y)} r(y)\right] + \beta\mathrm{KL}[\pi \mid \pi_{\text{ref}}],$$

其中 $\eta > 0$ 是步长参数。在信赖域策略优化中，近端项KL散度 $[\pi \mid \pi_t]$ 被替换为约束KL散度$[\pi \mid \pi_t] \leq \varepsilon$ （对某个 $\varepsilon > 0$），而在近端策略优化和GRPO中，它被替换为如下裁剪代理目标。

$$\pi_{t+1} = \operatorname*{argmin}_\pi \tilde{\mathcal{L}}[\pi],$$
$$\tilde{\mathcal{L}}[\pi] := -\mathbb{E}_{y\sim\pi_t}\left[\min\left(\frac{\pi(y)}{\pi_t(y)} r(y), \mathrm{clip}_\epsilon\left(\frac{\pi(y)}{\pi_t(y)}\right) r(y)\right)\right] + \beta\mathrm{KL}[\pi \mid \pi_{\text{ref}}],$$

其中 $\text{clip}_\epsilon(z) = \max(\min(z, 1+\epsilon)\ 1-\epsilon)$，其中某个 $\epsilon > 0$。

我们将上述 TRPO 类算法扩展为处理非线性奖励项 $R[\pi]$。首先，我们引入一阶变分（泛函导数）的概念。

定义 4.1. 若对所有 $\pi, \pi' \in \mathcal{P}$，均满足以下条件，则称泛函 $\frac{\delta R}{\delta \pi}\colon \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ 为泛函 $R\colon \mathcal{P} \to \mathbb{R}$ 的一阶变分（或泛函导数）：

$$\left.\frac{\mathrm{d}R[\pi + \epsilon(\pi' - \pi)]}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = \int \frac{\delta R}{\delta \pi}[\pi](y) \cdot (\pi' - \pi)(y)\,\mathrm{d}y.$$

请注意，一阶变分在恒定平移下具有自由度。下文假设对所有 $\pi \in \mathcal{P}$ 和 $y \in \mathcal{Y}$，一阶变分 $\frac{\delta R}{\delta \pi}[\pi](y)$ 均存在。为简化符号，我们将 $\frac{\delta R}{\delta \pi}[\pi]$ 记为 $\mathrm{d}R[\pi]$，将 $\int \mathrm{d}R[\pi](y) \cdot g(y)\,\mathrm{d}y$ 记为 $\overline{\langle \mathrm{d}R[\pi], g \rangle}$，其中函数 $g\colon \mathcal{Y} \to \mathbb{R}$。利用一阶变分，我们将当前策略 π_t 附近的目标函数 $R[\pi]$ 近似为

$$R[\pi] \approx R[\pi_t] + \int \frac{\partial R}{\partial \pi}[\pi_t](y) \cdot (\pi(y) - \pi_t(y))\,\mathrm{d}y.$$

利用这一近似，我们将代理损失定义为

$$\begin{aligned}\tilde{\mathcal{L}}[\pi] &= -\left(R[\pi_t] + \int \frac{\partial R}{\partial \pi}[\pi_t](y) \cdot (\pi(y) - \pi_t(y))\,\mathrm{d}y\right) \\ &\quad + \beta \mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}] \\ &= -\mathbb{E}_{y \sim \pi_t}\left[\frac{\pi}{\pi_t}\frac{\partial R}{\partial \pi}[\pi_t](y)\right] + \beta \mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}] + \text{const.}\end{aligned}$$

然后，我们通过求解以下近端目标函数来更新策略：

$$\pi_{t+1} = \operatorname*{argmin}_{\pi} \tilde{\mathcal{L}}[\pi] + \frac{1}{\eta}\mathrm{KL}[\pi \mid \pi_t].$$

除了 $\mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}]$ 项保持不变这一点外，上述更新可以解释为在概率测度空间中进行的镜像下降，其中 KL 散度是 Bregman 散度。

虽然上述更新在概念上很直接，但在实践中无法直接实现，因为一阶变分 $\frac{\partial R}{\partial \pi}[\pi_t](y)$ 通常以复杂的方式依赖于当前策略 π_t。为了解决这一问题，我们提出使用经验样本对其进行近似。在 GRPO 类算法中，对于给定的上下文 π_t，我们从当前策略中 x 采样多个响应。我们将采样响应记为 $\{y_j\}_{j=1}^M$，并将这些样本的经验分布记为 $\hat{\pi}_t = \frac{1}{M}\sum_{j=1}^M \delta_{y_j}$。请注意，我们也可以使用核密度估计等其他近似方法，以获得对当前策略更平滑的近似（Silverman，2018）。然后，我们将一阶变分近似为 $\frac{\partial R}{\partial \pi}[\hat{\pi}_t](y)$。基于此近似，我们将替代损失定义为

$$\hat{\mathcal{L}}[\pi] = -\mathbb{E}_{y \sim \pi_t}\left[\frac{\pi}{\pi_t}\frac{\partial R}{\partial \pi}[\hat{\pi}_t](y)\right] + \beta \mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}],$$

并通过近似求解以下近端目标来更新策略：

$$\pi_{t+1} \simeq \operatorname*{argmin}_{\pi} \hat{\mathcal{L}}[\pi] + \frac{1}{\eta}\mathrm{KL}[\pi \mid \pi_t].$$

上述更新可以解释为采用了修改后的奖励函数 $\tilde{r}_t(y) = \frac{\partial R}{\partial \pi}[\hat{\pi}_t](y)$ 的 GRPO 更新。也就是说，我们只需引入一个替代奖励函数即可修改原始 GRPO，同时保持算法的其余部分不变。这使得非线性 GRPO 算法易于在实践中实现。请注意，我们也可以使用其他 GRPO 类算法（例如 GSPO（Zheng 等人，2025））作为基础算法，而非使用 GRPO。此外，由于非线性 GRPO 具有较强的通用性，它可以应用于 IAMA 以外的各种问题，并将一些现有算法作为特例涵盖其中。更多讨论请参见附录 C。我们在算法 1 中总结了非线性 GRPO 算法。在实践中，我们还可以进一步近似上述更新：用经验样本替代期望，并像近端策略优化和 GRPO 那样用裁剪替代目标替换近端项。非线性 GRPO 实际实现的完整说明请参见算法 2。

算法 1 非线性 GRPO

输入： 初始策略 π_0、参考策略 π_{ref}、目标函数 R、正则化参数 β、步长 η、迭代次数 迭代次数 T
对于 $t = 0, 1, \ldots, T-1$ **执行**
  采样 M 响应 $\{y_j\}_{j=1}^M$ 来自 π_t
  计算近似的泛函导数：$\tilde{r}_t(y) = \frac{\partial R}{\partial \pi}[\hat{\pi}_t](y)$。
  定义代理损失：$\hat{\mathcal{L}}[\pi] = -E_{y \sim \pi_t}[\frac{\pi(y)}{\pi_t(y)} \cdot \tilde{r}_t(y)] + \beta \mathrm{KL}$散度$[\pi \mid \pi_{\mathrm{ref}}]$
  更新策略：$\pi_{t+1} \simeq \arg\min_\pi \hat{\mathcal{L}}[\pi] + \frac{1}{\eta}\mathrm{KL}[\pi \mid \pi_t]$
结束循环
返回： 最终策略 π_T

4.1. (Soft) BoN 的一阶变分

由于我们的公式相当通用，因此可以适用于各种推理时对齐算法。本节将讨论两个重要情形：1）BoN 和 2）Soft BoN（指数倾斜），并推导实现非线性 GRPO 算法所需的泛函导数的具体形式。根据链式法则，R 的一阶变分可计算为

$$\frac{\partial R}{\partial \pi}[\pi](y) = \sum_{i=1}^m \frac{\partial g}{\partial R_i}(R_1[\pi], \ldots, R_m[\pi]) \cdot \frac{\partial R_i}{\partial \pi}[\pi](y).$$

因此，要计算 R 的一阶变分，只需计算每种推理时

对齐算法 $\mathcal{T}_i$ 和奖励函数 r_i 的一阶变分。下面的命题分别给出了 BoN 和 softBoN 中 R_i 的一阶变分公式。

命题 4.2。令 $R[\pi] = \mathbb{E}_{y\sim \mathrm{BoN}_N[\pi]}[r(y)]$。假设在 π 下，$r(y)$ 的分布具有密度。那么，R 的一阶变分可计算为

$$\frac{\partial R}{\partial \pi}[\pi](y) = -\int_{r(y)}^{r_{\max}} N\cdot (C[\pi](r))^{N-1}\,\mathrm{d}r.$$

命题 4.3。令 $R[\pi] = \mathbb{E}_{y\sim \mathrm{SoftBoN}_\tau[\pi]}[r(y)]$。那么，$R$ 的一阶变分可计算为

$$\frac{\partial R}{\partial \pi}[\pi](y) = \frac{r(y)\exp(r(y)/\tau)}{\int \exp(r(z)/\tau)\cdot\pi(z)\,\mathrm{d}z} - \frac{\exp(r(y)/\tau)\cdot\int r(z)\exp(r(z)/\tau)\cdot\pi(z)\,\mathrm{d}z}{\left(\int \exp(r(z)/\tau)\cdot\pi(z)\,\mathrm{d}z\right)^2}.$$

证明见附录 F 。上述公式涉及积分，但若将 π 替换为经验分布 $\hat{}\pi$，则可化简为对经验样本求和。详情见附录D.1 。因此，我们可以轻松计算非线性 GRPO 算法所需的近似泛函导数。

5. 收敛性分析

本节将给出非线性 GRPO 算法的收敛保证。在参数空间中优化大语言模型通常是非凸的，且难以分析。另一方面，若将优化视为在概率分布（策略）空间中进行，问题则会简单得多。具体而言，对于 BoN 型目标的加权和及（平滑）最小值，$R[\pi]$ 是凹函数，如第 5.3 节所示。我们还将在附录 $R[\pi]$中讨论 C 中其他满足凹性的情形。

首先，我们为一般的光滑凹目标函数给出非线性 GRPO 的收敛保证。需要注意的是，凹性仅为收敛性分析所需，所提出的算法也可应用于非凹目标函数（例如 Soft BoN、Worst-of-N（Balashankar 等人，2025）情形）。这里，我们对目标函数作如下假设：

假设 5.1。奖励泛函 R 是凹的，且相对于 KL 散度是L-光滑的。也就是说，对于所有π，$\pi'\in\mathcal{P}$，我们有

$$R[\pi'] \le R[\pi] + \langle \mathrm{d}R[\pi], \pi'-\pi\rangle,$$
$$R[\pi'] \ge R[\pi] + \langle \mathrm{d}R[\pi], \pi'-\pi\rangle - L\cdot \mathrm{KL}[\pi' \mid \pi].$$

对于某个 $L>0$。

由于 $R[\pi]$ 是凹的，目标函数 $\mathcal{L}[\pi] = -R[\pi] + \beta$KL散度$[\pi \mid \pi_{\mathrm{ref}}]$ 关于KL-是强凸的散度，如果 $\beta>0$。因此，该优化问题具有唯一的最优策略 $\pi_*\in\mathcal{P}$。

5.1. 使用精确近端更新的收敛保证

在本节中，我们假设每次迭代时都精确求解近端目标。也就是说，在步骤 t，π_{t+1} 可计算为

$$\pi_{t+1} = \operatorname*{argmin}_{\pi} \tilde{\mathcal{L}}[\pi] + \frac{1}{\eta}\mathrm{KL}[\pi \mid \pi_t].$$

于是，我们得到以下收敛保证：

定理5.2. 设 $\{\pi_t\}_{t=0}^T$ 为由步长 $\eta = 1/L$的（精确）非线性 *GRPO* 算法生成的策略序列。则我们有

$$\mathcal{L}[\pi_T] - \mathcal{L}[\pi_*] \le \frac{\beta\cdot \mathrm{KL}[\pi_* \mid \pi_0]}{((L+\beta)/L)^T - 1}.$$

参见附录H.1 中的证明。证明思路基本上遵循 Aubin-Frankowski 等人（2022），但为处理KL正则化做了一些修改。该定理表明，若 $\beta>0$，则非线性 GRPO 算法以指数速度收敛到最优策略。

5.2. 非精确近端更新下的收敛保证

接下来，我们考虑实际情况：一阶变分使用经验样本进行近似，近端目标也被近似求解。也就是说，在第t步，我们将 π_{t+1} 计算为如下问题的近似解

$$\pi_{t+1} \approx \operatorname*{argmin}_{\pi} \hat{\mathcal{L}}[\pi] + \frac{1}{\eta}\mathrm{KL}[\pi \mid \pi_t].$$

注意，$\frac{\partial R}{\partial \pi}[\hat{}\pi_t]$ 是依赖于采样响应 $\{y_j\}_{j=1}^M$的随机变量。近端目标的一阶最优性条件给出 $\frac{\partial \mathcal{L}}{\partial \pi}[\pi] + \frac{1}{\eta}\log\frac{\pi}{\pi_t} = \text{const}$。在实践中，由于模型的表示能力有限以及非精确优化，存在残差项 $\hat{}r_t := \frac{\delta \mathcal{L}}{\delta \pi}[\pi_{t+1}] + \frac{1}{\eta}\log\frac{\pi_{t+1}}{\pi_t}$. 于是，我们有如下收敛保证：

定理 5.3. 假设近似一阶变分满足 $\mathbb{E}\left[\left\|\frac{\partial R}{\partial \pi}[\hat{}\pi_t](y) - \frac{\partial R}{\partial \pi}[\pi_t](y)\right\|_{\mathrm{sp}}^2\right] \le \varepsilon$（对某个 $\varepsilon \ge 0$），且优化残差满足 $\mathbb{E}\left[\|r_t\|_{\mathrm{sp}}^2\right] \le \delta$（对给定的 π_t，对某个 $\delta \ge 0$）。此处，$\|f\|_{\mathrm{sp}} = (\sup_y f(y) - \inf_y f(y))/2$ 为跨度半范数。设 $\{\pi_t\}_{t=0}^T$ 为由（不精确）非线性 *GRPO* 算法生成的策略序列，其步长为 $\eta = 1/L$，且随机指标 $\hat{} t\in\{1,\dots,T\}$ 服从如下分布

：$\text{Prob}\left(\hat{t}=t\right) \propto \left(\frac{L+\beta/2}{L}\right)^t$。于是，我们有

$$\mathbb{E}\left[\mathcal{L}[\pi_{\hat{t}}] - \mathcal{L}[\pi_*]\right] \leq \frac{(\beta/2)\cdot \mathrm{KL}[\pi_* \mid \pi_0]}{((L+\beta/2)/L)^T - 1} + \frac{2(\varepsilon+\delta)}{\beta}.$$

此处，该期望是针对泛函导数近似、优化ˆ过程以及 t 的选择中的随机性而取的。

参见附录 H.2以获取证明。即使采用非精确近端更新，非线性 GRPO 算法仍以指数速率收敛到最优策略的某个邻域，但由于泛函导数中的近似误差，会出现偏差项。

一般来说，为控制近似误差 ε 所需的样本数 M 可能随响应空间 $\mathcal{Y}$ 的维度呈指数增长。然而，在大多数情况下，测度变换 $\mathcal{T}$ 取决于奖励 $r(y)$ 的分布，而非响应 y 本身。因此，由于 $r(y)$ 是一维的，我们可以预期近似误差 ε 能够用适度数量的样本加以控制。我们将在下一节给出 BoN 类目标的近似误差的与维度无关的界。

5.3. BoN 类目标的性质

本节将证明$R[\pi]$对于 BoN 类对齐算法的凹性与光滑性。随后，我们将推导泛函导数近似误差的与维度无关的界。具体而言，我们考虑以下一般形式:

$$\mathcal{T}[\pi](y) = f(C[\pi](r(y)))\cdot \pi(y),$$

其中 $f:\ \mathbb{R}\to\mathbb{R}$ 单调递增，且 L_f- 利普希茨连续于 $[0,1]$。该形式涵盖了 BoN（$f(z)=N\cdot z^{N-1}$）以及泊松最优（BoP）（Khalaf 等人，2025）（从参数为 N 的泊松分布中抽取样本数 $\lambda>0$）这两种特殊情况，其中 $f(z)=\lambda\cdot e^{\lambda(z-1)}$。

凹性与光滑性 对于 BoN 类算法，我们有以下结果:

引理 5.4。 $R[\pi]=\int r(y)T[\pi](y)\,\mathrm{d}y$ 关于 π 是凹的。

引理 5.5。 $R[\pi]=\int r(y)\mathcal{T}[\pi](y)\mathrm{d}y$ 相对于 *KL* 散度是 $L_f\cdot r_{\max}$- 光滑的。

证明分别见附录 H.3 和 H.4 。这些引理表明，BoN 类目标满足收敛性分析所需的假设。值得注意的是，即使我们考虑多个 BoN 类目标的加权和或（平滑）最小值聚合，凹性仍然得以保留。

近似误差 最后，我们分析 BoN 类目标的泛函导数中的近似误差。下面的引理对使用经验样本时泛函导数中的近似误差进行了界定。

引理 5.6。 设 $\{y_j\}_{j=1}^M$ 为来自 π_t 的独立同分布样本，且ˆπ_t= $\frac{1}{M}\sum_{j=1}^M \delta_{y_j}$ 为经验分布。那么，泛函导数中的近似误差可界定为

$$\mathbb{E}\left[\left\|\frac{\partial R}{\partial \pi}[\hat{\pi}_t](y) - \frac{\partial R}{\partial \pi}[\pi_t](y)\right\|_{\mathrm{sp}}^2\right] \leq \frac{L_f^2\cdot r_{\max}^2}{M}.$$

证明见附录 H.5 。该引理表明，近似误差以 $O(1/M)$ 的速率递减，且与响应空间的维数无关$\mathcal{Y}$。我们还在数值实验中证明，非线性 GRPO 算法使用适量样本 M时也能取得良好效果。

6. 数值实验

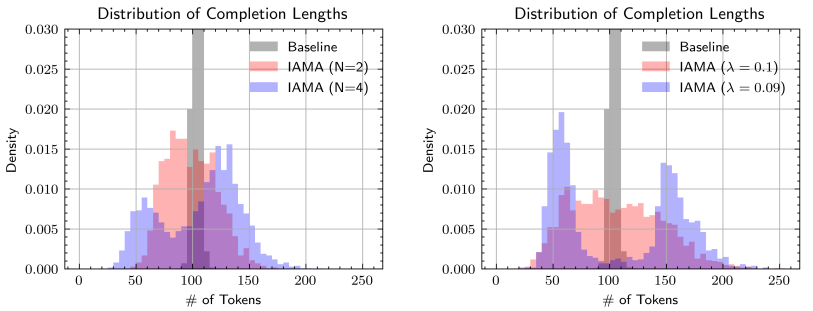
本节通过在多个模型和任务上开展广泛实验，对所提出的框架和算法的有效性进行实证评估。我们使用 TRL（von Werra 等人，2020）库实现（非线性）GRPO 算法。正如前文所述，只需将奖励函数替换为近似的泛函导数，即可实现非线性 GRPO。在所有实验中，我们将采样响应的数量设为 $M=8$，这是 TRL 库中的默认值。详细的实验设置见附录 I，其他实验结果见附录 J，其中包括采用不同超参数和基础模型开展的实验，以及对样本量 M影响的分析。为确保结果可复现，我们还在补充材料中提供了实现代码。

6.1. 长度奖励

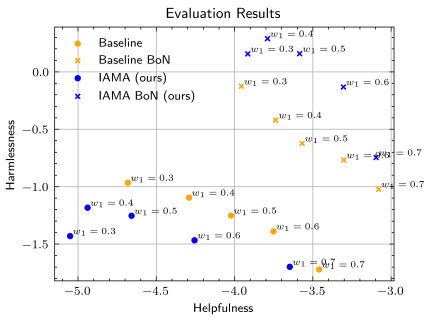
我们首先使用一个简单任务，通过长度奖励来说明对齐策略与元对齐策略之间的差异，并观察所提方法能否有效获得 IAMA 解。具体而言，我们考虑以下两种奖励:

$$r_1(y) = -\left(\frac{|y-L_1|}{L_{\max}}\right)^2, r_2(y) = -\left(\frac{|y-L_2|}{L_{\max}}\right)^2,$$

其中 $L_1=50, L_2=150$ 为目标长度，$L_{\max}=256$ 为最大长度。这些奖励函数分别鼓励模型生成长度接近 L_1 和 L_2的响应。该设置对应于部分用户偏好简短响应、而其他用户偏好冗长且详细响应的情形。这是



(a) 长度奖励任务中不同对齐方法生成的响应长度分布。通过非线性 GRPO 训练的元对齐模型能够借助（软）BoN 采样，有效适应对短响应和长响应的偏好；而通过标准 GRPO 训练的对齐模型只生成中等长度的响应。



(b) 评估集上的平均奖励。通过非线性 GRPO 训练的元对齐模型能够借助 BoN（$N = 4$）采样，有效权衡有用性与无害性。

与命题 3.1 中的设定类似。我们使用 UltraFeedback 数据集（Cui 等人，2023）作为上下文数据，并训练 Mistral 7B Instruct（Jiang 等人，2023）模型，使用非线性 GRPO 算法优化目标函数 $\mathcal{R}[\pi] = 0.5R_1[\pi] + 0.5R_2[\pi] - \beta \mathrm{KL}[\pi \,|\pi_{\mathrm{ref}}]$，其中 $\beta = 10^{-4}$。此处，$R_i[\pi] = \mathbb{E}_{y\sim\mathcal{T}_i[\pi]}[r_i(y)]$，且 $\mathcal{T}_i$ 为 BoN 或软最佳采样。作为基线，我们还使用标准 GRPO 训练基础模型，以最大化平均期望奖励 $\mathbb{E}_\pi[(r_1(y) + r_2(y))/2] - \beta \mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}]$。

图 3a （左）展示了 IAMA 模型针对 BoN 目标（$N = 2,\ 4$）生成的响应长度分布，图 3a （右）则展示了软 BoN 采样（$\tau = 0.09,\ 0.1$）下的分布。注意，这是基础模型的分布，即未进行（软）BoN 采样时的分布。可以看到，元对齐模型生成的响应丰富多样，能够有效涵盖较短和较长的长度，而基线模型只生成中等长度的响应。具体而言，BoN 中较大的 N 和软 BoN 中较小的 τ 会使响应长度更加两极分化。我们还测试了 $N = 8,\ 16$时的 BoN，其值等于或大于训练期间采样的响应数量（$M = 8$），并确认即使在这种情况下，非线性 GRPO 仍能有效发挥作用。由于篇幅限制，结果见附录J.2。

6.2. 有用性与无害性的基于人类反馈的强化学习

接下来，我们使用 hh-rlhf 数据集（Bai 等人，2022）在基于人类反馈的强化学习任务上评估所提方法。我们优化 Alpaca 7B（Taori 等人，2023；Dai 等人，2024）模型的复现版本，以最大化 $w_1 R_{\mathrm{helpful}}[\pi] + (1 - w_1) R_{\mathrm{harmless}}[\pi]$，其中 $R_{\mathrm{helpful}}[\pi]$和 $R_{\mathrm{harmless}}[\pi]$ 分别是针对 $N = 4$ 有用性和无害性奖励的 BoN 目标函数。我们改变权重 w_1，以考察模型在有用性与无害性之间进行权衡的能力。作为奖励模型，我们基于 Bradley–Terry 目标函数（Bradley 和 Terry，1952）训练通义千问 3 4B（Yang 等人，2025）。由于由人工评审者进行评估可能会成本高昂，我们使用基于 Qwen32B 的更大型奖励模型作为黄金奖励，遵循以往工作（Balashankar 等人，2025；Eisenstein等人，2024）来评估对齐模型。

为公平比较不同方法和权重的性能，我们固定目标 KL散度，而不是固定正则化系数 β。随后，我们采用对数空间比例控制器（Ziegler 等人，2019）在训练期间调整 KL正则化系数 β。在此实验中，我们将目标 KL 散度设为0.1。

图3b 展示了推理时使用和不使用 BoN 采样时，有用性与无害性奖励的评估结果。不使用 BoN 采样时，IAMA 模型无法胜过标准对齐模型，因为标准对齐方法可以直接优化期望奖励的加权和。另一方面，使用 BoN 采样时，与标准对齐方法相比，IAMA 模型能够推动帕累托前沿向前。这是因为元对齐模型会生成多样化的响应，涵盖有用性与无害性之间的不同权衡，从而使推理时对齐能够有效地选出高奖励响应。我们还在附录 $N = 8$中给出了J.3这一情况的结果。

7. 结论

在本研究中，我们提出了推理感知元对齐（IAMA），这是一种新颖的框架，通过两阶段对齐过程使大语言模型能够适配多样化的用户偏好。我们将 IAMA 表述为测度空间中的一个非线性优化问题，并开发了非线性 GRPO 算法来求解该问题。我们证明了 BoN 型目标函数的凸性和光滑性，并为所提出的算法提供了收敛保证。尽管实现简单，实验结果仍证明，借助非线性 GRPO，IAMA 能够有效地使大语言模型同时适配多个奖励函数。

致谢

TS 和 RH 的研究部分由 JSPS KAKENHI（24K02905）和 JST CREST（JPMJCR2115）资助。本研究得到新加坡国家研究基金会、资讯通信媒体发展局可信科技资助计划以及数字发展与信息部 AI 访问教授计划（资助编号 AIVP-2024-004）的支持。本材料中表达的任何观点、研究结果、结论或建议均属于作者所有，不代表新加坡国家研究基金会、资讯通信媒体发展局和数字发展与信息部的观点。

影响声明

本研究旨在推进相关方法，使大语言模型能够在推理时计算资源有限的情况下，对齐多样且可能相互冲突的人类偏好。通过提高推理时对齐的效率和灵活性，所提出的方法或可推动对齐语言模型在实际应用中得到更广泛的部署。同时，与其他对齐技术一样，误用或错误设定对齐标准可能导致意料之外的行为，这凸显了谨慎设计和评估对齐目标函数的重要性。总体而言，我们预计本研究不会带来超出大语言模型对齐研究通常涉及范围的独特伦理风险。

参考文献

Agnihotri, A., Jain, R., Ramachandran, D., 和 Wen, Z. 多目标偏好优化：提升生成模型与人类偏好的对齐。*arXiv* 预印本 *arXiv:2505.10892*，2025。

Amini, A., Vieira, T., Ash, E., and Cotterell, R. 变分最佳 n 选对齐。在国际学习表征会议，2025。

Aubin-Frankowski, P.-C., Korba, A., and L´eger, F. 测度空间中具有相对光滑性的镜像下降及其在 Sinkhorn 和 EM 中的应用。在神经信息处理系统进展，2022。

Bagirov, F., Arkhipov, M., Sycheva, K., Glukhov, E., and Bogomolov, E. N 个世界中的最佳选择：通过 max@ k 优化，利用最佳 N 采样使强化学习对齐。*arXiv* 预印本 *arXiv:2510.23393*，2025。

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., 等. 通过人类反馈强化学习训练有帮助且无害的助手——化学习。*arXiv* 预印本 *arXiv:2204.05862*，2022。Balashankar, A., Sun, Z., Berant, J., Eisenstein, J., Collins, M., Hutter, A., Lee, J., Nagpal, C., Prost, F., Sinha, A., 等。InfAlign：具备推理感知能力的语言模型对齐。发表于国际机器学习会议，2025。Barrett, L. 和 Narayanan, S. 在多重准则下学习所有最优策略。发表于国际机器学习会议，2008。

Beirami, A., Agarwal, A., Berant, J., D'Amour, A., Eisen- stein, J., Nagpal, C., and Suresh, A. T. 最佳 n 选对齐策略的理论保证。在国际机器学习会议，2025。

Bradley, R. A. 和 Terry, M. E. 不完整区组设计的秩分析：I. 配对比较方法。*Biometrika*，39(3/4)：324–345，1952。

Chen, Y.、Wolf, L.、Chakraborty, S.、Pacchiano, A. 和 Paschalidis, I. 通过行列式点过程增强大语言模型的多样性。载于*NeurIPS 2025* 研讨会：第二届强化学习实验研究者与理论研究者对齐研讨会，2025。

Chow, Y.、Tennenholtz, G.、Gur, I.、Zhuang, V.、Dai, B.、Kumar, A.、Agarwal, R.、Thiagarajan, S.、Boutilier, C. 和 Faust, A. 面向大语言模型 Best-of-n 采样的推理感知微调。载于国际学习表征会议，2025。

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S. 和 Amodei, D. 基于人类偏好的深度强化学习。神经信息处理系统进展，2017。

Chu, C., Blanchet, J. 和 Glynn, P. 概率泛函下降：对生成对抗网络、变分推断和强化学习的统一视角。发表于国际机器学习会议，2019。

Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., 等。Ultrafeedback：利用规模化人工智能反馈提升语言模型。*arXiv* 预印本 *arXiv:2310.01377*，2023。Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., 和 Yang, Y。安全人类反馈强化学习：基于人类反馈的安全强化学习。载于国际学习表征会议，2024。Eisenstein, J., Nagpal, C., Agarwal, A., Beirami, A., D'Amour, A. N., Dvijotham, K. D., Fisch, A., Heller, K. A., Pfohl, S. R., Ramachandran, D., Shaw, P., 和

Berant, J. 帮助还是从众？奖励模型集成能够缓解但无法消除奖励黑客行为。载于语言建模会议，2024。

Finn, C.、Abbeel, P. 与 Levine, S. 面向深度网络快速适应的模型无关元学习。载于 国际机器学习会议，2017。

Gui, L.、Gˆarbacea, C. 与 Veitch, V. 面向大语言模型的 BoNBoN 对齐与最佳 N 采样之妙。神经信息处理系统进展，2024。

Jaech, A.、Kalai, A.、Lerer, A.、Richardson, A.、El-Kishky, A.、Low, A.、Helyar, A.、Madry, A.、Beutel, A.、Carney, A. 等。OpenAI o1 系统卡。*arXiv* 预印本 *arXiv:2412.16720*，2024。

Jiang, A. Q.、Sablayrolles, A.、Mensch, A.、Bamford, C.、Chaplot, D. S.、de las Casas, D.、Bressand, F.、Lengyel, G.、Lample, G.、Saulnier, L.、Lavaud, L. R.、Lachaux, M.-A.、Stock, P.、Scao, T. L.、Lavril, T.、Wang, T.、Lacroix, T. 与 Sayed, W. E. Mistral 7B。*arXiv* 预印本 *arXiv:2310.06825*，2023。

Jiang, Y., Huang, J., Yuan, Y., Mao, X., Yue, Y., Zhao, Q., 和 Yan, L. 用风险敏感强化学习缓解大语言模型中的探索困境。*arXiv* 预印本 *arXiv:2509.24261*，2025。

Jinnai, Y., Morimura, T., Ariu, K., 和 Abe, K. 用于语言模型对齐的正则化最佳 n 项采样与最小贝叶斯风险目标。载于 计算语言学协会美洲分会会议 *2025* 年会议论文集：人类语言技术（第 *1* 卷：长篇论文），第 9321–9347 页，2025。

Karan, A. 和 Du, Y. 基于采样的推理：你的基础模型比你想象的更聪明。*arXiv* 预印本 *arXiv:2510.14901*，2025。

Kawata, R., Oko, K., Nitanda, A., 和 Suzuki, T. 用于扩散模型可证明对齐的直接分布优化。载于 国际学习表征会议，2025。

Khalaf, H., Verdun, C. M., Oesterling, A., Lakkaraju, H. 和 Calmon, F. d. P. 大语言模型中的推理时奖励黑客行为。*arXiv* 预印本 *arXiv:2506.19248*，2025。

Lambert, N.、Morrison, J.、Pyatkin, V.、Huang, S.、Ivison, H.、Brahman, F.、Miranda, L. J. V.、Liu, A.、Dziri, N.、Lyu, S. 等。Tulu 3：推动开放语言模型后训练的前沿。*arXiv* 预印本 *arXiv:2411.15124*，2024。

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S ，以及 Lin, M. 理解 R1-Zero 类训练：一种批判性视角。*arXiv* 预印本 *arXiv:2503.20783*，2025。

Mei, S.、Montanari, A. 和 Nguyen, P.-M. 双层神经网络损失景观的平均场视角。美国国家科学院院刊，115(33)：E7665–E7671，2018。Nakano, R.、Hilton, J.、Balaji, S.、Wu, J.、Ouyang, L.、Kim, C.、Hesse, C.、Jain, S.、Kosaraju, V.、Saunders, W. 等。WebGPT：借助浏览器并通过人工反馈进行问答。*arXiv* 预印本 *arXiv:2112.09332*，2021。

Nitanda, A., Wu, D. 和 Suzuki, T. 粒子对偶平均：具有全局收敛速率分析的平均场神经网络优化。神经信息处理系统进展，2021。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. 等。通过人类反馈训练语言模型遵循指令。载于神经信息处理系统进展，2022。

Rame, A., Couairon, G., Dancette, C., Gaya, J.-B., Shukor, M., Soulier, L. 和 Cord, M. 奖励融合：通过插值经多种奖励微调的权重，迈向帕累托最优对齐。神经信息处理系统进展，2023。

Schulman, J.、Levine, S.、Abbeel, P.、Jordan, M. 和 Moritz, P. 信赖域策略优化。在国际机器学习会议，2015 年。

Schulman, J.、Wolski, F.、Dhariwal, P.、Radford, A. 和 Klimov, O. 近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017 年。

Shao, Z.、Wang, P.、Zhu, Q.、Xu, R.、Song, J.、Bi, X.、Zhang, H.、Zhang, M.、Li, Y.、Wu, Y. 等。DeepSeek Math：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。

Silverman, B. W. 用于统计学和数据分析的密度估计。劳特里奇，2018 年。

Tang, Y.、Zheng, K.、Synnaeve, G. 和 Munos, R. 使用强化学习优化语言模型的推理时目标。*arXiv* 预印本 *arXiv:2503.19595*，2025 年。

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. 和 Hashimoto, T. B.。斯坦福 Alpaca：一个遵循指令的 LLaMA 模型。https://github.com/tatsu-lab/stanford_alpaca，2023。

Thrun, S. and P拉特, L. Learni学习 l赚取: Introducti在一个d 概述。在学习如何学习，第 3–17 页。施普林格，1998。Verdun, C. M.、Oesterling, A.、Lakkaraju, H. 和 Calmon, F. P. 面向模型对齐的软最佳 n 采样。*arXiv* 预印本 *arXiv:2505.03156*，2025。von Werra, L.、Belkada, Y.、Tunstall, L.、Beeching, E.、Thrush, T.、Lambert, N.、Huang, S.、Rasul, K. 和 Gallou´edec, Q. TRL：Transformers 强化学习，2020。网址：https://github.com/huggingface/trl。Walder, C. 和 Karkhanis, D. Pass@ K 策略优化：解决更难的强化学习问题。*arXiv* 预印本 *arXiv:2505.15201*，2025。Wang, X.、Wei, J.、Schuurmans, D.、Le, Q. V.、Chi, E. H.、Narang, S.、Chowdhery, A. 和 Zhou, D. 自一致性可提升语言模型中的思维链推理。在国际学习表征会议，2023。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。Yao, R.、Huang, L. 和 Yang, Y. 通过 KL 散度梯度流最小化概率测度空间上的凸泛函。在国际人工智能与统计学会议，第 2530–2538 页，2024。Ying, C.、Zhou, X.、Su, H.、Yan, D.、Chen, N. 和 Zhu, J. 通过约束条件风险价值实现安全强化学习。在国际人工智能联合会议，2022。Zhang, M.、Wang, P.、Tan, C.、Huang, M.、Zhang, D.、Zhou, Y. 和 Qiu, X. MetaAlign：在推理时根据多样化偏好对齐大语言模型。在计算语言学协会 *NAACL 2025* 论文集，第 5827–5845 页。计算语言学协会，2025。Zheng, C.、Liu, S.、Li, M.、Chen, X.-H.、Yu, B.、Gao, C.、Dang, K.、Liu, Y.、Men, R.、Yang, A.、Zhou, J. 和 Lin, J. 群组序列策略优化。*arXiv* 预印本 *arXiv:2507.18071*，2025。Ziegler, D. M.、Stiennon, N.、Wu, J.、Brown, T. B.、Radford, A.、Amodei, D.、Christiano, P. 和 Irving, G. 基于人类偏好的语言模型微调。*arXiv* 预印本 *arXiv:1909.08593*，2019。

A. 符号说明

为方便读者，我们在表1中汇总了本文使用的符号说明。

表 *1.* 符号说明汇总

符号说明	描述
$\mathcal{P}$	响应空间上的概率分布集合
π, π_{ref}	$\mathcal{P}$中的策略和参考策略
$r(y)$	将响应 y 映射为标量奖励的奖励函数
$\text{BoN}_N[\pi]$	基于基础策略 π 并使用 N 个样本的最佳 N 选一（BoN）对齐模型
$\text{SoftBoN}_\tau[\pi]$	基于基础策略 π、温度参数为 τ 的 Soft BoN 对齐模型
$\frac{\partial R}{\partial \pi}[\pi](y)$	在响应 y处，对策略 π 上的 R 求泛函导数
$\mathrm{d}R[\pi]$	$\frac{\partial R}{\partial \pi}[\pi]$ 作为关于 $\mathcal{P}$ 的线性泛函
$\langle \mathrm{d}R[\pi],\ \pi' \rangle$	线性泛函 $\mathrm{d}R[\pi]$ 在 π' 上的作用定义为 $\int \frac{\partial R}{\partial \pi}[\pi](y)\, \pi'(y)\, \mathrm{d}y$
$\text{KL散度}[\pi \mid \pi_{\text{ref}}]$	从 π_{ref} 到 π的 Kullback–Leibler 散度
$D_\phi[\pi \mid \mu]$	由凸泛函 ϕ诱导的 Bregman 散度，介于 μ 和 π之间
π_r	当 $y \sim \pi$时，奖励分布 $r(y)$

B. 相关工作的详细讨论

本节将更详细地讨论推理时对齐相关工作，并将我们提出的方法与以往研究进行对比。近期，Balashankar 等人 (2025) 提出了 InfAlign-CTRL，该方法对奖励函数进行非线性变换。然而，他们的研究聚焦于：1）单一奖励函数；2）BoN 类对齐算法；以及 3）胜率最大化目标。这些假设对于推导该方法至关重要，因为非线性变换取决于奖励分布；在单一奖励且以胜率最大化为目标的情况下，奖励分布均匀这一事实可以确定该分布。需要注意的是，当存在多个奖励时，即使考虑胜率最大化，由于相关性的存在，奖励分布也不再均匀。

一些研究（Walder & Karkhanis，2025；Bagirov 等人，2025）提出了用于 Pass@k（BoN）目标的无偏梯度估计量，以通过策略梯度方法对大语言模型进行对齐。然而，这些方法无法处理 BoN 以外的推理时对齐算法，并且每个梯度估计步骤都需要 $M \geq N$ 个样本。因此，它们不适用于大型 N 场景或泊松最优（BoP），因为这两者可能需要无上限数量的样本。

另一方面，我们提出的非线性 GRPO 算法能够处理范围广泛的推理时对齐算法和多个奖励函数。此外，即便将研究范围限定在 BoN 设置下，这些先前的研究也没有探讨该优化问题的理论性质。我们分析了该优化问题的损失景观，并证明了其凸性。基于这一分析，我们给出了收敛保证。这构成了一项额外且独立的贡献。

C. 与大语言模型对齐相关的其他非线性优化问题

由于我们提出的非线性 GRPO 算法具有很强的通用性，因此可将其应用于与大语言模型对齐相关的其他非线性优化问题。此外，非线性 GRPO 也可视为针对特定非线性优化问题的现有算法的推广。值得注意的是，现有研究并未明确讨论其与概率测度空间中的非线性优化之间的联系，而这一联系使我们能够提出一种通用算法并对其进行理论分析。

风险感知优化 条件风险价值（CVaR）是金融和强化学习领域广泛使用的一种风险度量，其定义为

$$\text{CVaR}_\alpha[\pi] := \mathbb{E}_{y\sim\pi}\left[r(y) \mid r(y) \leq q_\alpha[\pi]\right],$$

其中 $q_\alpha[\pi]$ 是 α时奖励分布的分位数 $y \sim \pi$。由于分位数函数 π 的存在，该目标函数关于 $q_\alpha[\pi]$是非线性的。Ying 等人提出的 CPPO (2022) 是一种针对 CVaR 目标函数的策略优化算法，

这可以视为我们非线性 GRPO 算法的一个特例。 Jiang 等人 (2025) 还提出了一种面向大语言模型的风险敏感型强化学习算法，使用定义如下的熵风险度量（ERM）

$$R_{\mathrm{ERM}}[\pi] := \frac{1}{\beta} \log \mathbb{E}_{y\sim\pi}\left[e^{\beta r(y)}\right].$$

如果 $\beta > 0$，该目标函数会鼓励风险偏好行为，并且关于 π 是凹函数，因为 log 是凹函数，而 $\mathbb{E}_{y\sim\pi}\left[e^{\beta r(y)}\right]$ 在 π 中是线性的。

多样性促进优化 多样性促进优化旨在学习一种能够生成多样化输出的策略。例如，Chen 等人 (2025) 提出使用行列式点过程（DPPs）作为促进多样性的正则项：

$$R_{\mathrm{DPP}}[\pi] := \mathbb{E}_{y_1,\ldots,y_N\sim\pi}\left[\log\det(K(y_1,\ldots,y_N))\right],$$

其中，$[K(y_1,\ldots,y_N)]_{i,j} = k(y_i,y_j)$ 是一个用于衡量输出 $y_1,\ldots,y_N$ 之间相似度的核矩阵。其他可能的正则项还包括 Rao 二次熵：

$$R_{\mathrm{Rao}}[\pi] := -\mathbb{E}_{y,y'\sim\pi}\left[k(y,y')\right].$$

如果核 k 是半正定的，那么上述目标函数关于 π 是凹函数，因为它可以被视为由 k 诱导的二次型。

D. 非线性 GRPO 的实际实现

这里，我们介绍非线性 GRPO 算法的一种实用实现。除使用近似泛函导数 Shao 等人 (2024提出的标准 GRPO 算法几乎完全相同，唯一的区别是将近似泛函导数 $\frac{\partial R}{\partial \pi}[\hat{}\pi_t](y)$ 用作奖励函数。还有许多此处未列出的技术可用于提高 GRPO 的稳定性和效率，也可直接应用于我们的非线性 GRPO。详情请参阅 TRL 文档（von Werra 等人，2020）。

算法 2 非线性 GRPO 的实用实现

1：输入：初始策略 π_0、参考策略 π_{ref}、奖励函数 $\{r_i\}_{i=1}^m$、权重 $\{w_i\}_{i=1}^m$、正则化参数 β、步长 η、迭代次数 T2：对于 $t=0$，1，...，$T-1$，执行 3：从 π_t采样 M 个响应 $\{y_j\}_{j=1}^M$ 4：计算近似泛函导数：$\tilde{r}_t(y) = \frac{\partial R}{\partial \pi}[\hat{}\pi_t](y)$。5：计算优势：$A_j = \frac{\tilde{r}_t(y_j)}{s_t} - b_t$，其中 $b_t = \frac{1}{M}\sum_{j=1}^M \tilde{r}_t(y_j)$ 且 $s_t = \sqrt{\frac{1}{M}\sum_{j=1}^M (\tilde{r}_t(y_j) - b_t)^2}$ 6：定义代理损失：$\hat{\mathcal{L}}[\pi] = \frac{1}{M}\sum_{i=1}^M \min\left[\frac{\pi}{\pi_t}(y_i)\,(y_i)\cdot A_i,\ \cancel{\mathrm{clip}}_\epsilon\left(\frac{\pi}{\pi_t}(y_i)\,(y_i)\right)\cdot A_i\right] + \beta\,\widehat{\mathrm{KL}}[\pi \mid \pi_{\mathrm{ref}}]$，其中 $\hat{}\mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}] = \frac{1}{M}\sum_{i=1}^M \frac{\pi_{\mathrm{ref}}(y_i)}{\pi}(y_i) - \cancel{\log}\frac{\pi_{\mathrm{ref}}}{\pi}(y_i)\,\pi(y_i) - \cancel{1}$7：使用优化器（例如 Adam）近似求解最小化问题 $\min_\pi \hat{\mathcal{L}}[\pi]$，得到 π_{t+1} 8：循环结束 9：返回：最终策略 π_T

D.1. 泛函导数的计算

本节介绍如何计算基于经验分布的 BoN 和 Soft BoN 目标函数的泛函导数。在这两种情况下，计算复杂度取决于样本数量 M，而非 BoN 的规模 N。

BoN 如果我们用由从 π_t 中抽取的 M 个样本 $\{y_i\}_{i=1}^M$ 构造的经验分布 $\hat{}\pi_t$ 来近似泛函导数 $\frac{\partial R}{\partial \pi}[\hat{}\pi_t](y)$，那么计算瓶颈在于对奖励进行排序以计算经验分位数。因此，我们可以在 $O(M\log M)$ 时间内，按如下方式计算这些样本处的泛函导数：

算法 3 BoN 目标的线性化奖励计算

1：输入：奖励值 $\{r_j\}_{j=1}^M$**，BoN 参数** N **2：按奖励值升序排列索引** $\{1,\dots,M\}$**，使得** $r_{(1)} \le r_{(2)} \le \cdots \le r_{(M)}$**，其中 (*i*) 表示第** i **小的奖励所对应的索引。3：初始化累积和** $c \leftarrow 0$ **4：设置参考奖励** $r_{\text{next}} \leftarrow r_{(M)}$ **5：为所有** $j = 1$ **初始化** $\tilde{r}_j \leftarrow 0$**，...，** M **6：对于** $i = M$**、** $M-1$**、...、1，执行 7：计算奖励差值** $\Delta r \leftarrow r_{\text{next}} - r_{(i)}$ **8：计算经验累积分布函数** $q \leftarrow i/M$ **9：更新累积贡献** $c \leftarrow c + Nq^{N-1}\Delta r$ **10：赋值线性化奖励** $\tilde{r}_{(i)} \leftarrow -c$ **11：更新参考奖励** $r_{\text{next}} \leftarrow r_{(i)}$ **12：结束循环 13：返回：** $\{\tilde{r}_j\}_{j=1}^M$

Soft BoN 如果我们使用由从 π_t 中抽取的 M 个样本 $\{y_i\}_{i=1}^M$ 构造的经验分布 $\hat{}\pi_t$，来近似泛函导数 $\frac{\partial R}{\partial \pi}[\hat{}\pi_t](y)$，则可以在 $O(M)$ 时间内按如下方式计算该泛函导数：

算法 4 Soft BoN 目标的线性化奖励计算

1: 输入： 奖励值 $\{r_j\}_{j=1}^M$，温度参数 τ2: 计算归一化常数 $Z \leftarrow \frac{1}{M}\sum_{j=1}^M \exp(r_j/\tau)$ 3: 计算经 softmax 加权的期望奖励 $\bar{r} \leftarrow \frac{1}{M}\sum_{j=1}^M r_j \exp(r_j/\tau)/Z$4: **对于** $j = 1,\dots,M$ **执行**5: 计算线性化奖励 $\tilde{r}_j \leftarrow r_j \exp(r_j/\tau)/Z - \exp(r_j/\tau)\,\bar{r}/Z$6: **结束循环**7: **返回：** 线性化奖励 $\{\tilde{r}_j\}_{j=1}^M$

E. 辅助引理

引理 E.1. 设 $C[\pi](r)$ 为累积分布函数（*CDF*），定义为 $C[\pi](r) = \int \mathbb{1}[r(y) \le r]\pi(y)\mathrm{d}y$。则 $C[\pi](r)$ 的泛函导数由 $\frac{\partial C}{\partial \pi}[\pi](r) = \mathbb{1}[r(y) \le r]$ 给出。

证明。对于任意 $\pi, \pi' \in \mathcal{P}$ 和 $\epsilon \in \mathbb{R}$，我们有

$$\begin{aligned} C[\pi + \epsilon(\pi' - \pi)](r) - C[\pi](r) &= \int \mathbb{1}[r(y) \le r] \cdot (\pi + \epsilon(\pi' - \pi))(y)\,\mathrm{d}y - \int \mathbb{1}[r(y) \le r] \cdot \pi(y)\,\mathrm{d}y \\ &= \epsilon \int \mathbb{1}[r(y) \le r] \cdot (\pi' - \pi)(y)\,\mathrm{d}y \end{aligned}$$

因此

$$\left.\frac{\mathrm{d}C[\pi + \epsilon(\pi' - \pi)](r)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = \int \mathbb{1}[r(y) \le r] \cdot (\pi' - \pi)(y)\,\mathrm{d}y.$$

因此，我们有 $\frac{\partial C}{\partial \pi}[\pi](r) = \mathbb{1}[r(y) \le r]$。 □

引理 E.2。对于奖励函数 $r(y)$ 和基础策略 π，假设当 $y \sim \pi$ 时，$r(y)$ 的分布是连续的。那么，*BoN* 策略 $\mathrm{BoN}_N[\pi]$ 满足

$$\mathrm{BoN}_N[\pi](y) = N \cdot C[\pi](r(y))^{N-1} \cdot \pi(y)$$

证明。 参见 Balashankar 等人(2025) 的引理 5。 □

引理 E.3 （三点不等式）。 设 $\mathcal{G}$、 ψ、 ϕ 为凸泛函。那么，对于任意 μ、 ν 和 $\bar{\nu}$：$=\arg\min \mathcal{G}[\nu]+\beta D_{\psi}$（$\nu \mid \pi_{\text{ref}}$）$+LD_{\phi}$（$\nu \mid \mu$） $\in \mathcal{P}$，我们有

$$\begin{aligned}\mathcal{G}[\bar{\nu}]+\beta D_{\psi}(\bar{\nu} \mid \pi_{\text{ref}})+LD_{\phi}(\bar{\nu} \mid \mu) \leq \mathcal{G}[\nu]+\beta D_{\psi}(\nu \mid \pi_{\text{ref}})+LD_{\phi}(\nu \mid \mu) \\ -LD_{\phi}(\nu \mid \bar{\nu})-\beta D_{\psi}(\nu \mid \bar{\nu}).\end{aligned}$$

证明。 这是对 Aubin-Frankowski 等人 （2022）中引理 3 的扩展，我们沿用相同的证明策略。令$f=\mathcal{G}[\nu]+\beta D_{\psi}$（$\nu \mid \pi_{\text{ref}}$） $+LD_{\phi}$（$\nu \mid \mu$）。注意， f是凸的，因为 $\mathcal{G}$、 D_{ψ}和 D_{ϕ}均为凸的。于是，我们有

$$\begin{aligned}D_f(\nu \mid \bar{\nu}) &= D_{\mathcal{G}}(\nu \mid \bar{\nu})+\beta D_{\psi}(\nu \mid \bar{\nu})+LD_{\phi}(\nu \mid \bar{\nu}) \\ &\geq \beta D_{\psi}(\nu \mid \bar{\nu})+LD_{\phi}(\nu \mid \bar{\nu})\end{aligned}$$

根据 Bregman 散度的非负性。

根据 $\bar{\nu}$ 的最优性，我们有 $D_f(\nu \mid \bar{\nu}) = f(\nu) - f(\bar{\nu}) - \langle \frac{\partial f}{\partial \pi}[\bar{\nu}], \nu - \bar{\nu} \rangle \leq f(\nu) - f(\bar{\nu})$。因此，我们有

$$f(\nu)-f(\bar{\nu}) \geq \beta D_{\psi}(\nu \mid \bar{\nu})+LD_{\phi}(\nu \mid \bar{\nu}).$$

重新整理上述不等式，可得所需结果。 □

F. 命题4.2和4.3的证明

情况 1：BoN BoN 将对齐模型定义为

$$\text{BoN}_N[\pi](y) = f(C[\pi](r(y))) \cdot \pi(y),$$

其中 $f(r) = N \cdot r^{N-1}$。因此，目标函数为

$$\begin{aligned}R(\pi) &= \mathbb{E}_{y \sim \text{BoN}_N[\pi]}[r(y)] \\ &= \int r(y) \cdot f(C[\pi](r(y))) \cdot \pi(y)\,\mathrm{d}y \\ &= \int_0^{r_{\max}} r \cdot f(C[\pi_r](r)) \cdot \pi_r(r)\,\mathrm{d}r \\ &= [r \cdot F(C[\pi_r](r))]_0^{r_{\max}} - \int_0^{r_{\max}} F(C[\pi_r](r))\,\mathrm{d}r\end{aligned}$$

其中， π_r 是 $y \sim \pi$ 时 $r(y)$ 的分布， F 是 f 的原函数。此处，最后一个不等式由分部积分得到。然后，泛函导数可计算为

$$\begin{aligned}\frac{\delta R[\pi]}{\delta \pi}(z) &= -\int f(C[\pi_r](r)) \frac{\delta C[\pi_r]}{\delta \pi}(z)\,\mathrm{d}r \\ &= -\int f(C[\pi_r](r)) \cdot \mathbb{1}_{r \geq r(z)}\,\mathrm{d}r \\ &= -\int_{r(z)}^{r_{\max}} f(C[\pi_r](r))\,\mathrm{d}r.\end{aligned}$$

证明完毕。

情形 2：Soft BoN 设 $\lambda = 1/\tau$。Soft BoN 将对齐模型定义为

$$\text{SoftBoN}_{\tau}[\pi](y) = \frac{\exp(\lambda r(y)) \cdot \pi(y)}{\int \exp(\lambda r(z)) \cdot \pi(z)\,\mathrm{d}z}.$$

因此，目标函数为

$$\begin{aligned} R(\pi) &= \mathbb{E}_{y\sim \mathrm{SoftBoN}_\lambda}\left[r(y)\right] \\ &= \int r(y)\cdot \frac{\exp(\lambda r(y))\cdot \pi(y)}{\int \exp(\lambda r(z))\cdot \pi(z)\,\mathrm{d}z}\,\mathrm{d}y \\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\log\left(\int \exp(\lambda r(z))\cdot \pi(z)\,\mathrm{d}z\right). \end{aligned}$$

于是，泛函导数可计算为

$$\begin{aligned} \frac{\partial R}{\partial \pi}[\pi](y) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\partial}{\partial \pi}\log\left(\int \exp(\lambda r(z))\cdot \pi(z)\,\mathrm{d}z\right)\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\exp(\lambda r(y))}{\int \exp(\lambda r(z))\cdot \pi(z)\,\mathrm{d}z}\right) \\ &= \frac{r(y)\exp(\lambda r(y))}{\int \exp(\lambda r(z))\cdot \pi(z)\,\mathrm{d}z} - \frac{\exp(\lambda r(y))\cdot \int r(z)\exp(\lambda r(z))\cdot \pi(z)\,\mathrm{d}z}{\left(\int \exp(\lambda r(z))\cdot \pi(z)\,\mathrm{d}z\right)^2}. \end{aligned}$$

至此，证明完毕。

G. 第3节的证明

G.1. 命题3.1的证明

令 Y_1，……，$Y_N \overset{\text{i.i.d.}}{\sim} \pi$ 为来自 π 的 N 独立同分布样本。由于 r_1 严格递减且 r_2 严格递增，使 r_1 和 r_2 达到最大值的样本分别是最小和最大次序统计量，

$$\begin{aligned} Y_{(1)} &:= \min_{k\le N} Y_k, \\ Y_{(N)} &:= \max_{k\le N} Y_k. \end{aligned}$$

因此，目标函数为

$$R[\pi] := \mathbb{E}[-Y_{(1)}^2] + \mathbb{E}[-(1-Y_{(N)})^2].$$

对于任意 $y\in[0,1]$，

$$\begin{aligned} \mathrm{Prob}\left(Y_{(1)} > y\right) &= (1-C[\pi](y))^N, \\ \mathrm{Prob}\left(Y_{(N)} \le y\right) &= C[\pi](y)^N. \end{aligned}$$

利用 $\mathbb{E}[X^2] = \int_0^1 2t\,\mathbb{P}(X>t)\,dt$ 对于 $X\in[0,1]$，并进行变量替换，

$$R[\pi] = -2\int_0^1 \left[y(1-C[\pi](y))^N + (1-y)C[\pi](y)^N\right] dy.$$

因此，最大化 $R[\pi]$ 等价于最小化

$$I(\pi) := \int_0^1 \phi_y(C[\pi](y))\,dy,$$

其中 $\phi_y(p) := y(1-p)^N + (1-y)p^N$。对于固定的 $y\in(0,1)$，我们有

$$\phi_y''(p) = N(N-1)\left[y(1-p)^{N-2} + (1-y)p^{N-2}\right] > 0$$

对于 $p \in (0,\ 1)$。因此， ϕ_y 在 (0，1) 上严格凸，并且有唯一的极小值点满足 g

$$0 = \phi_y'(p) = -Ny(1-p)^{N-1} + N(1-y)p^{N-1}$$
$$\iff \left(\frac{p}{1-p}\right)^{N-1} = \frac{y}{1-y}.$$

也就是说，最小化解为

$$C^*(y) = \frac{y^\alpha}{y^\alpha + (1-y)^\alpha}, \qquad \alpha := \frac{1}{N-1}.$$

此外， $C^*(y)$ 严格递增，且 $C^*(0) = 0$和 $C^*(1) = 1$，这意味着 $\pi_* := \frac{dC^*}{dy}$ 属于 $\mathcal{P}$，并且是 $\mathcal{I}(\pi)$ 的一个最小化解。证毕。

H. 第5节的证明

H.1. 定理5.2的证明

为简化符号，我们定义

$$F[\pi] := -R[\pi],$$
$$\mathcal{L}[\pi] = F[\pi] + \beta \mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}].$$

注意到，由于 R 是凹且 L-光滑的，因此 F 是凸且 L-光滑的。

由 $F[\pi]$ 的 L-光滑性可得

$$\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] \le \langle \mathrm{d}F[\pi_t], \pi_{t+1} - \pi_t \rangle + L\mathrm{KL}[\pi_{t+1} \mid \pi_t] + \beta \mathrm{KL}[\pi_{t+1} \mid \pi_{\mathrm{ref}}] - \beta \mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}].$$

应用引理 E.3 于凸泛函 $\langle \mathrm{d}F[\pi_t],\ \pi - \pi_t \rangle$可得

$$\begin{aligned}\langle \mathrm{d}F[\pi_t], \pi_{t+1} - \pi_t \rangle + L\mathrm{KL}[\pi_{t+1} \mid \pi_t] + \beta \mathrm{KL}[\pi_{t+1} \mid \pi_{\mathrm{ref}}] \le \langle \mathrm{d}F[\pi_t], \pi - \pi_t \rangle + L\mathrm{KL}[\pi \mid \pi_t] + \beta \mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}] \\ - (L+\beta)\mathrm{KL}[\pi \mid \pi_{t+1}]\end{aligned}$$

对于任意 $\pi \in \mathcal{P}$。结合上述两个不等式，可得以下不等式:

$$\begin{aligned}\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] \le \langle \mathrm{d}F[\pi_t], \pi - \pi_t \rangle \\ + L\mathrm{KL}[\pi \mid \pi_t] \\ + \beta \mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}] \\ - (L+\beta)\mathrm{KL}[\pi \mid \pi_{t+1}] \\ - \beta \mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}].\end{aligned}$$

令 $\pi = \pi_t$，可得

$$\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] \le -(L+\beta)\mathrm{KL}[\pi_t \mid \pi_{t+1}] \le 0,$$

这意味着 $\mathcal{L}[\pi_t]$的单调性。

另一方面，通过设定 $\pi = \pi_*$，我们有

$$\begin{aligned}\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] \le \langle \mathrm{d}F[\pi_t], \pi_* - \pi_t \rangle \\ + L\mathrm{KL}[\pi_* \mid \pi_t] \\ + \beta \mathrm{KL}[\pi_* \mid \pi_{\mathrm{ref}}] \\ - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}] \\ - \beta \mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}].\end{aligned}$$

由于 F 是凸的，我们有

$$\langle \mathrm{d}F[\pi_t], \pi_* - \pi_t \rangle \leq F(\pi_*) - F(\pi_t).$$

因此，我们得到

$$\begin{aligned}
\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] \leq\ & F(\pi_*) - F(\pi_t) \\
& + L\mathrm{KL}[\pi_* \mid \pi_t] \\
& + \beta\mathrm{KL}[\pi_* \mid \pi_{\mathrm{ref}}] \\
& - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}] \\
& - \beta\mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}] \\
=\ & \mathcal{L}[\pi_*] - \mathcal{L}[\pi_t] \\
& + L\mathrm{KL}[\pi_* \mid \pi_t] \\
& - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}].
\end{aligned}$$

重新整理上述不等式，可得

$$\mathcal{L}[\pi_{t+1}] \leq \mathcal{L}[\pi_*] + L\mathrm{KL}[\pi_* \mid \pi_t] - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}].$$

将上述不等式从 $t=0$ 到 $T-1$ 求和，可得

$$\begin{aligned}
\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathcal{L}[\pi_t] &\leq \sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathcal{L}[\pi_*] + L\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathrm{KL}[\pi_* \mid \pi_{t-1}] - (L+\beta)\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathrm{KL}[\pi_* \mid \pi_t] \\
&= \sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathcal{L}[\pi_*] + (L+\beta)\left[\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^{t-1} \mathrm{KL}[\pi_* \mid \pi_{t-1}] - \sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathrm{KL}[\pi_* \mid \pi_t]\right] \\
&= \sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathcal{L}[\pi_*] + (L+\beta)\left[\sum_{t=0}^{T-1}\left(\frac{L+\beta}{L}\right)^t \mathrm{KL}[\pi_* \mid \pi_t] - \sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathrm{KL}[\pi_* \mid \pi_t]\right] \\
&= \sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathcal{L}[\pi_*] + (L+\beta)\left[\mathrm{KL}[\pi_* \mid \pi_0] - \left(\frac{L+\beta}{L}\right)^T \mathrm{KL}[\pi_* \mid \pi_T]\right] \\
&\leq \sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathcal{L}[\pi_*] + (L+\beta)\mathrm{KL}[\pi_* \mid \pi_0].
\end{aligned}$$

将两边同时除以 $\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t$，可得

$$\begin{aligned}
\mathcal{L}[\pi_T] - \mathcal{L}[\pi_*] &\leq \frac{\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t \mathcal{L}[\pi_t]}{\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t} - \mathcal{L}[\pi_*] \\
&\leq \frac{(L+\beta)\mathrm{KL}[\pi_* \mid \pi_0]}{\sum_{t=1}^{T}\left(\frac{L+\beta}{L}\right)^t} \\
&= \frac{(L+\beta)\mathrm{KL}[\pi_* \mid \pi_0]}{\frac{(L+\beta)}{L} \cdot \frac{((L+\beta)/L)^T - 1}{(L+\beta)/L - 1}} \\
&= \frac{\beta \cdot \mathrm{KL}[\pi_* \mid \pi_0]}{((L+\beta)/L)^T - 1}.
\end{aligned}$$

第一个不等式由 $\mathcal{L}[\pi_t]$ 的单调性得出。证明完毕。

H.2. 定理5.3的证明

设经验梯度为 $\hat{}g_t := \mathrm{d}F[\hat{}\pi_t]$，并定义估计误差 $e_t := \hat{}g_t - \mathrm{d}F[\pi_t]$。由 F 的 L-光滑性可得

$$\begin{aligned}\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] &\le \mathrm{d}F[\pi_t](\pi_{t+1} - \pi_t) + L\,\mathrm{KL}[\pi_{t+1} \mid \pi_t] + \beta\,\mathrm{KL}[\pi_{t+1} \mid \pi_{\mathrm{ref}}] - \beta\,\mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}] \\ &= \hat{g}_t(\pi_{t+1} - \pi_t) + L\,\mathrm{KL}[\pi_{t+1} \mid \pi_t] + \beta\,\mathrm{KL}[\pi_{t+1} \mid \pi_{\mathrm{ref}}] - \beta\,\mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}] - e_t(\pi_{t+1} - \pi_t).\end{aligned}$$

由于采用了非精确优化，存在一个残差项 r_t，其定义为

$$r_t := \hat{g}_t + \frac{1}{\eta}\log\frac{\pi_{t+1}}{\pi_t} + \beta\log\frac{\pi_{t+1}}{\pi_{\mathrm{ref}}}.$$

注意，如果优化是精确的，则 r_t 为恒定值。这意味着 π_{t+1} 是以下优化问题的最优解：

$$\min_{\pi\in\mathcal{P}} \langle \hat{g}_t - r_t, \pi - \pi_t\rangle + \frac{1}{\eta}\mathrm{KL}[\pi \mid \pi_t] + \beta\mathrm{KL}[\pi \mid \pi_{\mathrm{ref}}].$$

将引理 E.3 应用于凸泛函 $\langle \hat{}g_t - r_t,\ \pi - \pi_t\rangle$，可得

$$\begin{aligned}&\langle \hat{g}_t - r_t, \pi_{t+1} - \pi_t\rangle + L\mathrm{KL}[\pi_{t+1} \mid \pi_t] + \beta\mathrm{KL}[\pi_{t+1} \mid \pi_{\mathrm{ref}}] \\ &\quad= \langle \hat{g}_t, \pi_{t+1} - \pi_t\rangle + L\mathrm{KL}[\pi_{t+1} \mid \pi_t] + \beta\mathrm{KL}[\pi_{t+1} \mid \pi_{\mathrm{ref}}] - \langle r_t, \pi_{t+1} - \pi_t\rangle \\ &\quad\le \langle \hat{g}_t - r_t, \pi_* - \pi_t\rangle + L\mathrm{KL}[\pi_* \mid \pi_t] + \beta\mathrm{KL}[\pi_* \mid \pi_{\mathrm{ref}}] - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}] \\ &\quad= \langle \hat{g}_t, \pi_* - \pi_t\rangle + L\mathrm{KL}[\pi_* \mid \pi_t] + \beta\mathrm{KL}[\pi_* \mid \pi_{\mathrm{ref}}] - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}] - \langle r_t, \pi_* - \pi_t\rangle.\end{aligned}$$

结合上述两个不等式，可得

$$\begin{aligned}\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] \le\ & \hat{g}_t(\pi_* - \pi_t) + L\mathrm{KL}[\pi_* \mid \pi_t] + \beta\mathrm{KL}[\pi_* \mid \pi_{\mathrm{ref}}] \\ & - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}] - \beta\mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}] - \langle e_t, \pi_{t+1} - \pi_t\rangle - \langle r_t, \pi_* - \pi_{t+1}\rangle. \qquad (2)\end{aligned}$$

根据 F 的凸性，

$$\mathrm{d}F[\pi_t](\pi_* - \pi_t) \le F[\pi_*] - F[\pi_t],$$

因此

$$\hat{g}_t(\pi_* - \pi_t) = \mathrm{d}F[\pi_t](\pi_* - \pi_t) + \langle e_t, \pi_* - \pi_t\rangle \le F[\pi_*] - F[\pi_t] + \langle e_t, \pi_* - \pi_t\rangle.$$

将其代入（2）可得

$$\begin{aligned}\mathcal{L}[\pi_{t+1}] - \mathcal{L}[\pi_t] \le\ & F[\pi_*] - F[\pi_t] + L\mathrm{KL}[\pi_* \mid \pi_t] + \beta\mathrm{KL}[\pi_* \mid \pi_{\mathrm{ref}}] \\ & - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}] - \beta\mathrm{KL}[\pi_t \mid \pi_{\mathrm{ref}}] + \langle e_t, \pi_* - \pi_{t+1}\rangle - \langle r_t, \pi_* - \pi_{t+1}\rangle.\end{aligned}$$

重新整理上述不等式，可得

$$\mathcal{L}[\pi_{t+1}] \le \mathcal{L}[\pi_*] + L\mathrm{KL}[\pi_* \mid \pi_t] - (L+\beta)\mathrm{KL}[\pi_* \mid \pi_{t+1}] + \langle e_t, \pi_* - \pi_{t+1}\rangle - \langle r_t, \pi_* - \pi_{t+1}\rangle. \qquad (3)$$

令 $c = (\sup_y e_t(y) + \inf_y e_t(y))/2$。由 Pinsker 不等式和 H¨older 不等式，

$$\begin{aligned}\langle e_t, \pi_* - \pi_{t+1}\rangle &= \langle e_t - c, \pi_* - \pi_{t+1}\rangle \\ &\le \|e_t - c\|_\infty \|\pi_* - \pi_{t+1}\|_1 \\ &= \|e_t\|_{\mathrm{sp}} \cdot \|\pi_* - \pi_{t+1}\|_1 \\ &\le \|e_t\|_{\mathrm{sp}} \cdot \sqrt{2\,\mathrm{KL}[\pi_* \mid \pi_{t+1}]}.\end{aligned}$$

第二个等号由以下事实得出： $\|e_t - c\|_\infty = \max(\sup_y e_t(y) - c,\ c - \inf_y e_t(y)) = (\sup_y e_t(y) - \inf_y e_t(y))/2 = \|e_t\|_{\mathrm{sp}}$。应用杨氏不等式（$ab \le \frac{a^2}{2\alpha} + \frac{\alpha b^2}{2}$），可得 — —

$$\langle e_t, \pi_* - \pi_{t+1}\rangle \le \frac{\|e_t\|_{\mathrm{sp}}^2}{2\alpha} + \alpha \mathrm{KL}[\pi_* \mid \pi_{t+1}]$$

对于任意 $\alpha > 0$。

类似地，我们有

$$-\langle r_t, \pi_* - \pi_{t+1}\rangle \le \frac{\|r_t\|_{\mathrm{sp}}^2}{2\alpha} + \alpha \mathrm{KL}[\pi_* \mid \pi_{t+1}]$$

对于任意 $\alpha > 0$。

将这些代入（3），得到

$$\mathcal{L}[\pi_{t+1}] \le \mathcal{L}[\pi_*] + L\,\mathrm{KL}[\pi_* \mid \pi_t] - (L + \beta - 2\alpha)\,\mathrm{KL}[\pi_* \mid \pi_{t+1}] + \frac{\|e_t\|_{\mathrm{sp}}^2 + \|r_t\|_{\mathrm{sp}}^2}{2\alpha}. \tag{4}$$

令 $\alpha = \beta/4$（使得 $\beta - 2\alpha = \beta/2$），我们有

$$\mathcal{L}[\pi_{t+1}] \le \mathcal{L}[\pi_*] + L\mathrm{KL}[\pi_* \mid \pi_t] - (L + \beta/2)\,\mathrm{KL}[\pi_* \mid \pi_{t+1}] + \frac{2(\|e_t\|_{\mathrm{sp}}^2 + \|r_t\|_{\mathrm{sp}}^2)}{\beta}.$$

令 $\mu := L + \beta/2$。对 $t = 1, \ldots, T$ 求和，并赋予权重 $(\mu/L)^t$，得到

$$\sum_{t=1}^{T} \left(\frac{\mu}{L}\right)^t \mathcal{L}[\pi_t] \le \sum_{t=1}^{T} \left(\frac{\mu}{L}\right)^t \mathcal{L}[\pi_*] + \mu \left[\mathrm{KL}[\pi_* \mid \pi_0] - \left(\frac{\mu}{L}\right)^T \mathrm{KL}[\pi_* \mid \pi_T]\right] + \sum_{t=1}^{T} \left(\frac{\mu}{L}\right)^t \frac{2(\|e_{t-1}\|_{\mathrm{sp}}^2 + \|r_{t-1}\|_{\mathrm{sp}}^2)}{\beta}.$$

令 $S_T := \sum_{t=1}^{T} (\mu/L)^t = \frac{\mu}{L} \cdot \frac{(\mu/L)^T - 1}{(\mu/L) - 1}$。则

$$\begin{aligned}\frac{\sum_{t=1}^{T} \left(\frac{\mu}{L}\right)^t \mathcal{L}[\pi_t]}{S_T} - \mathcal{L}[\pi_*] &= \sum_{t=1}^{T} \mathrm{Prob}\left(\hat{t} = t\right) \mathcal{L}[\pi_t] - \mathcal{L}[\pi_*] \\ &\le \frac{\mu\,\mathrm{KL}[\pi_* \mid \pi_0]}{S_T} + \frac{2}{\beta} \cdot \frac{\sum_{t=1}^{T} (\mu/L)^t (\|e_{t-1}\|_{\mathrm{sp}}^2 + \|r_{t-1}\|_{\mathrm{sp}}^2)}{S_T}.\end{aligned}$$

代入 $\mu = L + \beta/2$ 得到

$$\begin{aligned}\frac{\mu\,\mathrm{KL}[\pi_* \mid \pi_0]}{S_T} &= \frac{(L + \beta/2)\,\mathrm{KL}[\pi_* \mid \pi_0]}{\frac{(L+\beta/2)}{L} \cdot \frac{((L+\beta/2)/L)^T - 1}{((L+\beta/2)/L) - 1}} \\ &= \frac{(\beta/2) \cdot \mathrm{KL}[\pi_* \mid \pi_0]}{((L + \beta/2)/L)^T - 1}.\end{aligned}$$

根据假设 $\mathbb{E}[\|e_t\|_{\mathrm{sp}}^2] \le \varepsilon$ 和 $\mathbb{E}[\|r_t\|_{\mathrm{sp}}^2] \le \delta$，我们有

$$\mathbb{E}\left[\frac{\sum_{t=1}^{T} (\mu/L)^t (\|e_{t-1}\|_{\mathrm{sp}}^2 + \|r_{t-1}\|_{\mathrm{sp}}^2)}{S_T}\right] \le \frac{\sum_{t=1}^{T} (\mu/L)^t (\varepsilon + \delta)}{S_T} = \varepsilon + \delta.$$

因此，我们有

$$\mathbb{E}\left[\mathcal{L}[\pi_{\hat{t}}] - \mathcal{L}[\pi_*]\right] \le \frac{(\beta/2) \cdot \mathrm{KL}[\pi_* \mid \pi_0]}{((L + \beta/2)/L)^T - 1} + \frac{2(\varepsilon + \delta)}{\beta}.$$

H.3. 引理证明5.4

证明。令 F 为 f 的原函数。由于 f 是非递减的，因此 F 是凸的。通过分部积分，我们有 e

$$\begin{aligned} R[\pi] &= \int r(y) f(C[\pi](r(y)))\pi(y)\,\mathrm{d}y \\ &= \int_0^\infty r f(C[\pi](r))\pi_r(r)\,\mathrm{d}r \\ &= \int_0^\infty r \cdot \frac{\mathrm{d}}{\mathrm{d}r} F(C[\pi](r))\,\mathrm{d}r \\ &= [rF[\pi](r)]_0^{r_{\max}} - \int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r \\ &= r_{\max}F(1) - \int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r. \end{aligned}$$

因此，只需证明 $P[\pi] \coloneqq \int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r$ 在 π 上是凸的。对于任意 π， $\pi' \in \mathcal{P}$ 和 $\lambda \in [0,1]$，有

$$\begin{aligned} P[(\lambda\pi + (1-\lambda)\pi')] &= \int_0^{r_{\max}} F(C[(\lambda\pi + (1-\lambda)\pi')](r))\,\mathrm{d}r \\ &= \int_0^{r_{\max}} F\left(\lambda C[\pi](r) + (1-\lambda)C[\pi'](r)\right)\mathrm{d}r \\ &\le \int_0^{r_{\max}} \left(\lambda F(C[\pi](r)) + (1-\lambda)F(C[\pi'](r))\right)\mathrm{d}r \\ &= \lambda P[\pi] + (1-\lambda)P[\pi']. \end{aligned}$$

这里，不等式由 F 的凸性得出。证毕。 □

H.4. 引理证明5.5

证明。 由分部积分可得

$$\begin{aligned} R(\pi) &= \int_0^{r_{\max}} r(y) f(C[\pi](r))\,\mathrm{d}\pi(y) \\ &= [rF(C[\pi](r))]_0^{r_{\max}} - \int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r, \\ &= r_{\max}F(1) - \int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r, \end{aligned}$$

其中 F是 f 的原函数。由于第一项是恒定的，只需证明 $\int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r$ 相对于 KL 散度是 $L_f \cdot r_{\max}$-光滑的。

首先，我们证明对所有 r， $F(C[\pi](r))$ 相对于 KL 散度是 L_f-光滑的。对任意 π, π'，我们有

$$\begin{aligned}
&F(C[\pi'](r)) - F(C[\pi](r)) - \int f(C[\pi](r)) \cdot \mathbb{1}_{r(y)\le r} \cdot (\pi'(y) - \pi(y))\,\mathrm{d}y \\
&= F(C[\pi'](r)) - F(C[\pi](r)) - f(C[\pi](r)) \cdot (C[\pi'](r) - C[\pi](r)) \\
&\le \frac{L_f}{2}(C[\pi'](r) - C[\pi](r))^2 \\
&= \frac{L_f}{2}\left(\int (\pi'(y) - \pi(y)) \cdot \mathbb{1}_{r(y)\le r}\,\mathrm{d}y\right)^2 \\
&\le \frac{L_f}{2}\left(\int |\pi'(y) - \pi(y)| \cdot \mathbb{1}_{r(y)\le r}\,\mathrm{d}y\right)^2 \\
&\le \frac{L_f}{2}\left(\int |\pi'(y) - \pi(y)|\,\mathrm{d}y\right)^2 \\
&\le 2L_f d_{\mathrm{TV}}(\pi, \pi')^2 \\
&\le L_f \mathrm{KL}[\pi' \mid \pi].
\end{aligned}$$

此处，第一个不等式由 F 的 L_f-光滑性得出，最后一个不等式由 Pinsker 不等式得出。因此，对所有 r， $F(C[\pi](r))$ 相对于 KL 散度是 L_f- 光滑的。

接下来，对两边关于 $r \in [0, r_{\max}]$ 积分，我们得到

$$\begin{aligned}
&\int_0^{r_{\max}} F(C[\pi'](r))\,\mathrm{d}r - \int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r - \int_0^{r_{\max}} \int f(C[\pi](r)) \cdot \mathbb{1}_{r(y)\le r} \cdot (\pi'(y) - \pi(y))\,\mathrm{d}y\,\mathrm{d}r \\
&\le L_f \cdot r_{\max} \cdot \mathrm{KL}[\pi' \mid \pi].
\end{aligned}$$

因此，$\int_0^{r_{\max}} F(C[\pi](r))\,\mathrm{d}r$ 相对于 KL 散度是 $L_f \cdot r_{\max}$-光滑的。 □

H.5. 引理的证明5.6

由于跨度半范数满足 $\|f\|_{\mathrm{sp}} \le \|f\|_\infty$，只需界定无穷范数即可。一阶变分可计算为

$$\frac{\partial R}{\partial \pi}[\pi](y) = -\int_{r(y)}^{r_{\max}} f(C[\pi](r))\,\mathrm{d}r.$$

因此，近似误差可界定如下：

$$\begin{aligned}
\left|\frac{\partial R}{\partial \pi}[\hat{\pi}_t](y) - \frac{\partial R}{\partial \pi}[\pi_t](y)\right| &\le \int_0^{r_{\max}} |f(C[\hat{\pi}_t](r)) - f(C[\pi_t](r))|\,\mathrm{d}r \\
&\le L_f r_{\max} \|C[\hat{\pi}_t] - C[\pi_t]\|_\infty
\end{aligned}$$

令 $Z = \|C[\hat{\pi}_t] - C[\pi_t]\|_\infty$ 为一个随机变量。根据 Dvoretzky-Kiefer-Wolfowitz（DKW）不等式，我们有

$$\mathrm{Prob}\left(Z \ge \varepsilon\right) \le 2e^{-2M\varepsilon^2},$$

这意味着

$$\begin{aligned}
\mathbb{E}\left[Z^2\right] &= 2\int_0^1 \varepsilon \mathrm{Prob}\left(Z \ge \varepsilon\right)\mathrm{d}\varepsilon \le 2\int_0^1 \varepsilon \cdot 2e^{-2M\varepsilon^2}\,\mathrm{d}\varepsilon \\
&\le \frac{1}{M}.
\end{aligned}$$

因此，我们有

$$\mathbb{E}\left[\left\|\frac{\partial R}{\partial \pi}[\hat{\pi}_t](y) - \frac{\partial R}{\partial \pi}[\pi_t](y)\right\|_\infty^2\right] \le \frac{L_f^2 r_{\max}^2}{M},$$

由此完成证明。

I. 实验细节

这里，我们补充说明实验的更多细节。实验中使用的超参数汇总于表2。

计算资源 我们的实验在 Intel(R) Xeon(R) Silver 4316 CPU @ 2.30GHz 和 8 块 NVIDIA A100-SXM4-80GB GPU 上进行。

长度奖励实验 我们将响应数量设为 $M = 8$，学习率设为 $1e-6$，批次大小设为 48，迭代次数设为 $T = 3000$，并设定 $\beta = 1e-4$。我们采用恒定学习率调度，并在 GRPO 中关闭奖励缩放，以稳定训练。其他超参数采用 TRL 库的默认值（von Werra 等人，2020）。

HH-RLHF 实验 我们将响应数量设为 $M = 8$，学习率设为 $1e-6$，批次大小设为 48，迭代次数设为$T = 1000$，并设定初始 $\beta = 1e-2$。我们采用对数空间比例控制器（Ziegler 等人，2019）在训练过程中调整 KL 正则化系数 β，具体如下：

$$e_t = \text{clip}\left(\frac{\text{KL}[\pi_t \mid \pi_{\text{ref}}] - \text{KL}_{\text{target}}}{\text{KL}_{\text{target}}}, [-0.2, 0.2]\right),$$
$$\beta_{t+1} = \beta_t \cdot (1 + 0.1 \cdot e_t)$$

目标 KL散度 $\text{KL}_{\text{target}} = 0.1$。为稳定训练，我们在 GRPO 中关闭奖励缩放。其他超参数均采用 TRL 库的默认值（von Werra 等人，2020）。在评估中，我们使用 HH-RLHF 数据集验证集中的前 1000 个样本，以降低计算成本。

表2。 我们实验中使用的超参数。

超参数	长度奖励实验	HH-RLHF 实验
响应数量 M	8	8
学习率	10^{-6}	10^{-6}
每设备批次大小	8	8
迭代次数 T	3000	1000
初始 KL 系数 β	10^{-4}	10^{-2}
目标 KL KL散度$_{\text{target}}$	N/A	10^{-1}
学习率调度器	恒定	线性
bf16 训练	True	True
优化器	AdamW	AdamW
裁剪 ϵ	0.2	0.2

J. 其他实验结果

J.1. 响应数量的影响 M

为研究响应数量 M 对线性化损失估计的影响，我们开展了额外的合成实验。由于通常难以获得推理时对齐下的真实期望奖励，我们考虑一个易于处理的设定：奖励分布为 $[0,1]$ 上的贝塔分布，推理时对齐则通过 BoN 使用 N 个响应来完成。设 $\pi_\alpha(y)$ 为 $[0,1]$ 上形状参数为 $(\alpha,\ 1)$ 的贝塔分布。那么，$\pi_1(y)$ 等价于 $[0,1]$ 上的均匀分布。众所周知，$\text{BoN}_N[\pi_\alpha](y)$ 服从贝塔分布 $\pi_{\alpha N}(y)$，其期望奖励为 $R_{\text{bon}}[\pi_1] = \frac{\alpha N}{\alpha N+1}$。因此，我们可以解析地计算期望奖励的真实差值，即 $R_{\text{bon}}[\pi_\alpha] - R_{\text{bon}}[\pi_1] = \frac{N\alpha}{N\alpha+1} - \frac{N}{N+1}$。我们预期，当 M 足够大且 $\alpha \simeq 1$ 时，估计得到的线性化损失 $\langle \nabla R[\hat{\pi}],\ \pi_\alpha - \pi_1 \rangle$ 能够很好地近似这一真实差值。具体而言，我们通过从 π_1 中抽取 $M = 2$、……、32 个样本来构造 $\hat{\pi}$，并估计 $\alpha = 1 + 10^{-4}$ 的线性化损失。随后，在 100,000 次试验中计算估计的线性化损失与期望奖励真实差值之间的均方误差。为降低方差，线性化

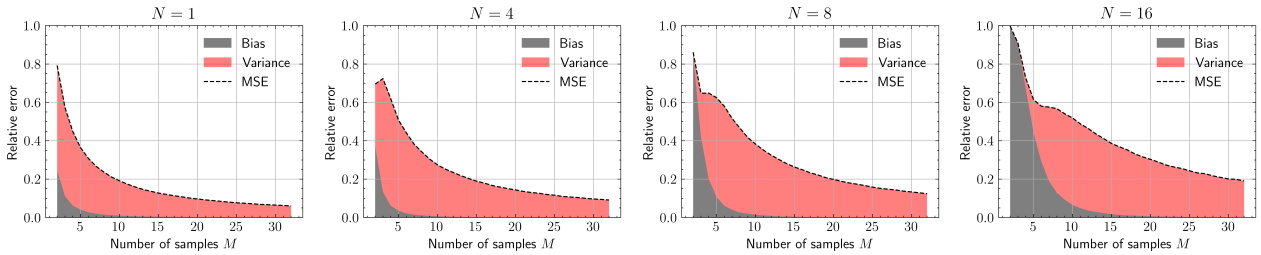


图 *4.* 不同 M和 N值下线性化损失估计的期望误差。从左至右， $N = 1$、4、8、16。阴影区域表示期望误差的偏差-方差分解。

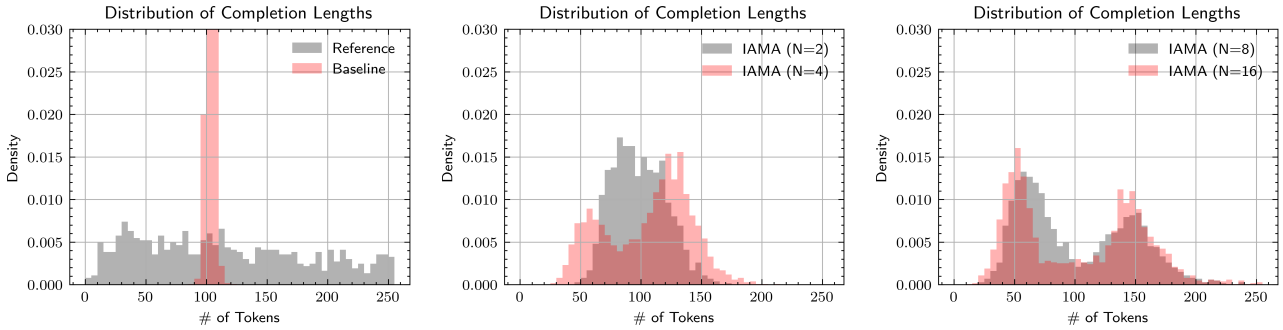


图 *5.* 参考模型 π_{ref}、通过标准 GRPO 对齐的基线模型，以及使用 BoN（$N = 2$、4、8、16）目标函数训练的 IAMA 模型的补全长度分布。

奖励按 GRPO 算法中的做法，通过减去均值进行中心化。图 4 展示了不同 M和 N取值下的比较结果。可以观察到，较大的 N需要更大的 M才能达到较低的期望误差，但与 $M < N$相比，即使在 $N = 1$情况下，仍能取得合理的准确度；后者对应标准 GRPO。特别是，偏差项随着 M增大而迅速减小，并在 $M \geq N$时几乎消失。

J.2. 长度奖励实验

除了图 $N = 2$、4中的结果外，3a我们还在图 $N = 8$、16中提供了 5的结果。我们还展示了参考模型（即训练前的初始模型）的分布。我们观察到，参考模型的分布近似均匀，而基线模型生成的响应长度平平，集中在约 100 个 token 左右。另一方面，IAMA 模型会生成多样化的响应，并且随着 N增加，分布会更加两极化，生成较短和较长的响应。这与第 3.1 节的结果一致。

J.3. HH-RLHF 实验

我们为 HH-RLHF 实验提供了额外的实验结果。

实验结果 $N = 8$图 6 （右）展示了使用 $N = 8$的基于人类反馈的强化学习结果。与 $N = 8$图 3b中的情况类似，与基线方法相比，我们的方法显著推动了帕累托前沿。

在 Mistral 7B 上进行的实验 我们使用 Mistral 7B Instruct（Jiang 等人，2023）进行了额外实验。Mistral 7B Instruct 是最先进的开源大语言模型之一。除了 $\beta = 0.01$和 KL散度$_{\text{target}} = 0.1$外，我们采用了与 Alpaca 7B 相同的实验设置。图 7 展示了 IAMA 与基线方法在使用和不使用 BoN 采样（$N = 4$）时的帕累托前沿。我们观察到的趋势与 Alpaca 7B 的结果类似。

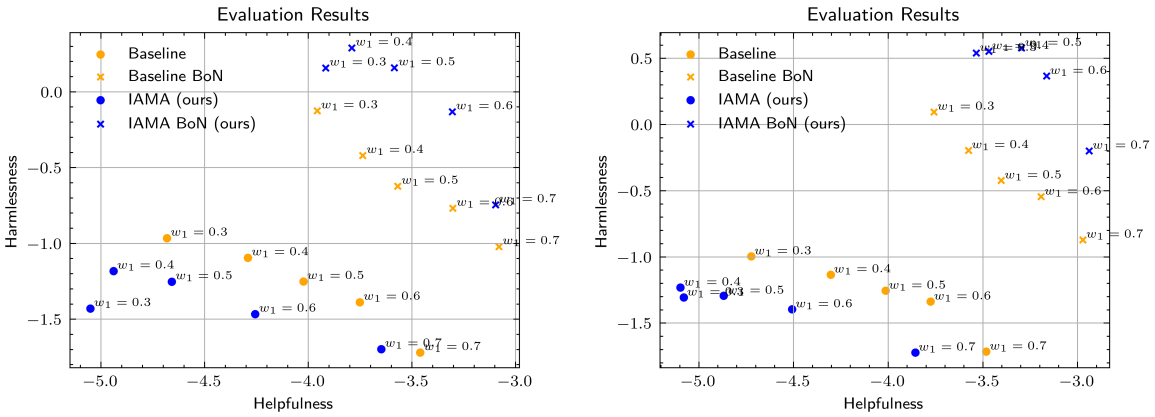


图 6. 使用 $N = 4$（左）和 $N = 8$（右）的基于人类反馈的强化学习结果。

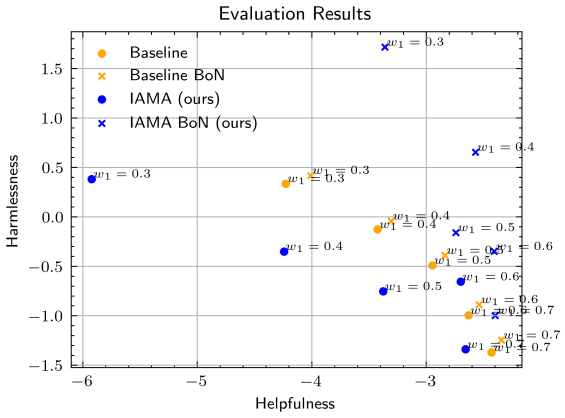


图 7. 使用 $N = 4$ 在 Mistral 7B 上进行 HH-RLHF 的结果。