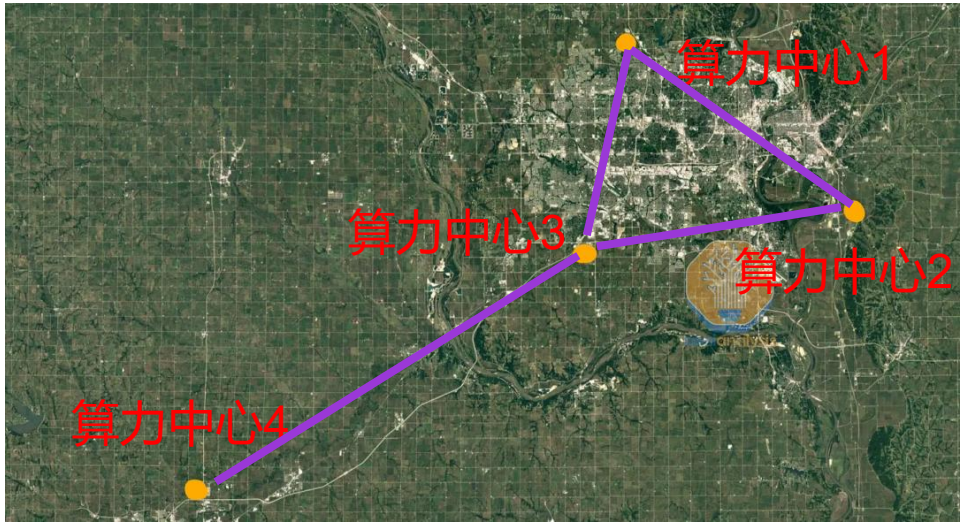
群组期望策略优化
异构算力网络上的稳定强化学习算法

张晗

网络智能研究部/云计算所

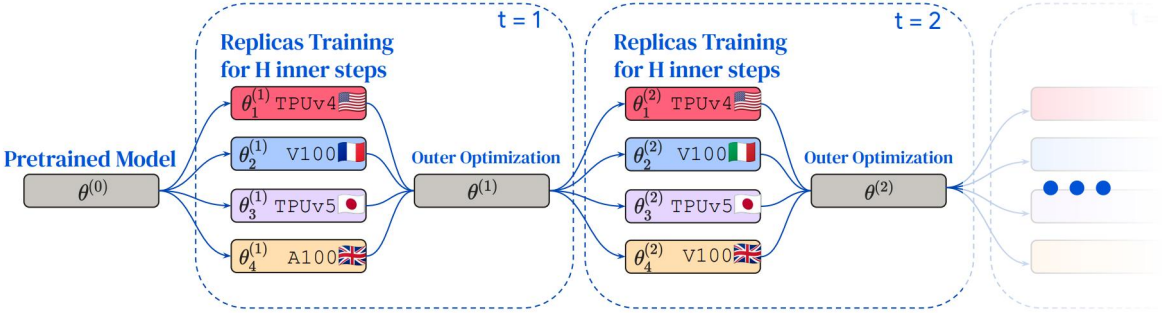
研究背景

- 将异构算力中心连成网络是有效提升算力规模的方式
- 利用算力网研究模型训练具有以下优势和挑战
  - **优势**：提升算力利用率和增大算力规模
  - **劣势**：无法像单算力中心一样高效通信

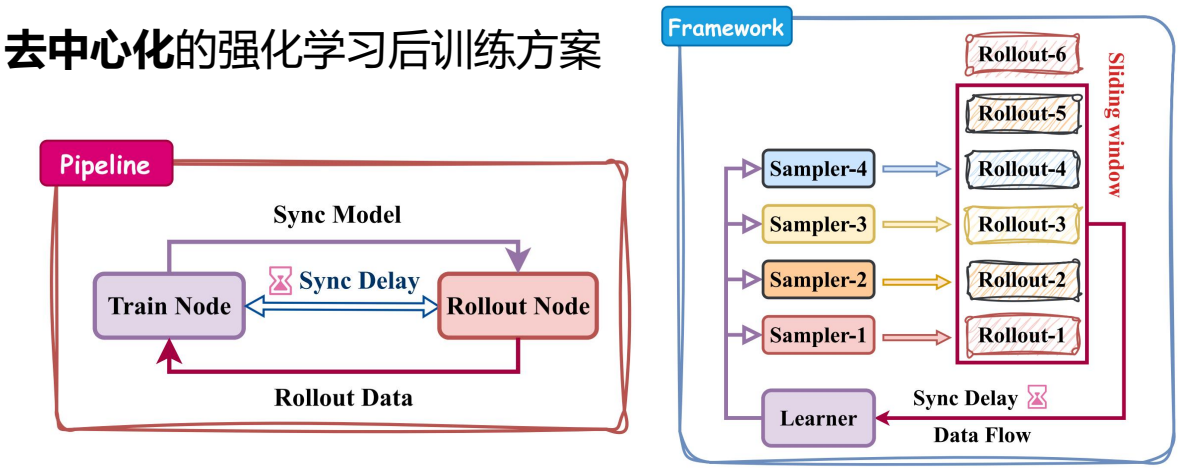


谷歌在俄亥俄州和爱荷华州/内布拉斯加州4个算力中心，目前正在快速扩张和升级高速互联光纤网络，正在进行多中心协同训练探索

去中心化预训练的算法方案：**DiLoCo双层优化器**



去中心化的强化学习后训练方案



研究问题

The goal of HeteroRL is to optimize the policy π_θ to maximize the expected cumulative reward. To reduce gradient variance, an advantage function is used, leading to the objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\theta_k}(\cdot|x)} \left[\frac{\pi_{\theta_{k+\tau}}(y|x)}{\pi_{\theta_k}(y|x)} \cdot A(x,y) \right], \tag{1}$$

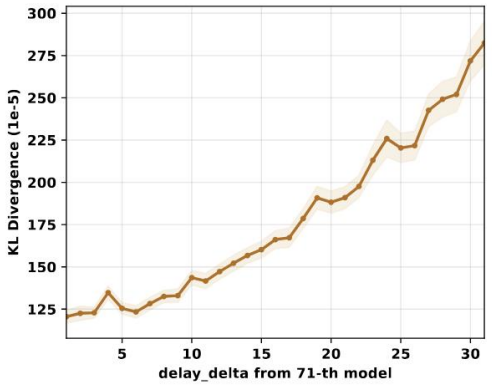
where τ is a random variable. For online RL, $\tau \equiv 0$.

- π_{θ_k} (short for q): the policy used by the *sampler* at time step k to generate rollout trajectories.
- $\pi_{\theta_{k+\tau}}$ (short for p): the latest policy at the *learner* at time step $k+\tau$, used for gradient updates.
- $\tau(\geq 0)$: *policy staleness*, representing the discrepancy in policy versions between the sampler and the learner, caused by `network delays` and `computational asynchrony`.

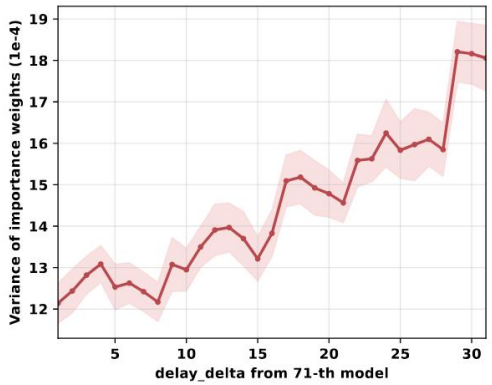
存在挑战：数据延迟对LLM-RL的影响

实验现象

实验发现延迟会导致 KL 散度增加（图 a），进而导致重要性权重方差提升（图 b），最终导致优势函数期望的估计误差增加（图 c）。由于优化目标是最大化优势函数期望的估计，这个估计的误差大将会导致梯度的变化幅度，进而影响训练稳定性和性能。

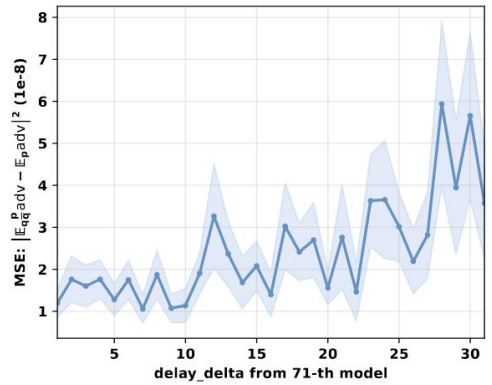


(a) KL 散度随着延迟增加而增加

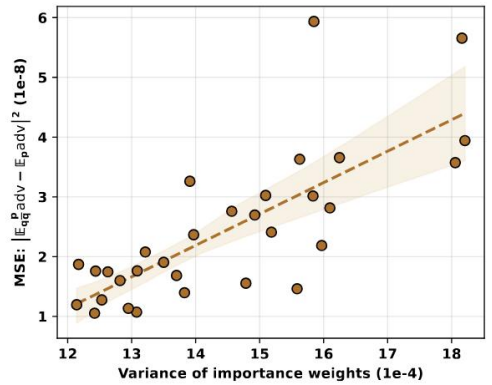


$$\nabla_\theta \left[\frac{p_\theta(y|x)}{q(y|x)} \cdot A(x,y) \right] = \left[\frac{p_\theta(y|x)}{q(y|x)} \cdot A(x,y) \right] \cdot \nabla_\theta \log p_\theta(y|x)$$

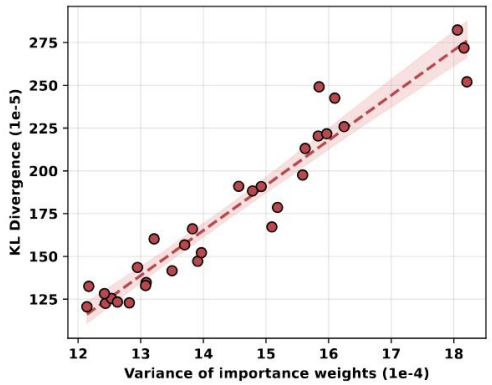
数增加而增加



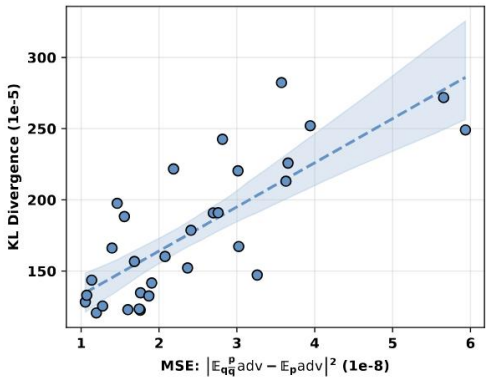
(c) 延迟步数与 $\mathbb{E}[adv(x,y)]$ 的估计误差



(d) 重要性权重方差与 $\mathbb{E}[adv(x,y)]$ 估计误差强相关



(e) 重要性权重方差与 KL 散度强相关 (相关系数 0.96)



(f) $\mathbb{E}[adv(x,y)]$ 的估计误差与 KL 散度强相关

GEPO：群组期望策略优化算法

GRPO

$$\hat{A}_{i,t} = \frac{R(x,y^i) - \text{mean}\{R(x,y^1),\ldots,R(x,y^G)\}}{\text{std}\{R(x,y^1),\ldots,R(x,y^G)}$$

$$\mathcal{L}_{\text{GRPO}_{i,t}} = \min\left[\frac{\pi_\theta(y_t^i \mid x, y_{<t}^i)}{\pi_{\theta_{\text{old}}}(y_t^i \mid x, y_{<t}^i)}\hat{A}_{i,t}, clip_{1\pm\epsilon}\left[\frac{\pi_\theta(y_t^i \mid x, y_{<t}^i)}{\pi_{\theta_{\text{old}}}(y_t^i \mid x, y_{<t}^i)}\hat{A}_{i,t}\right]\hat{A}_{i,t}\right]$$

GSPO

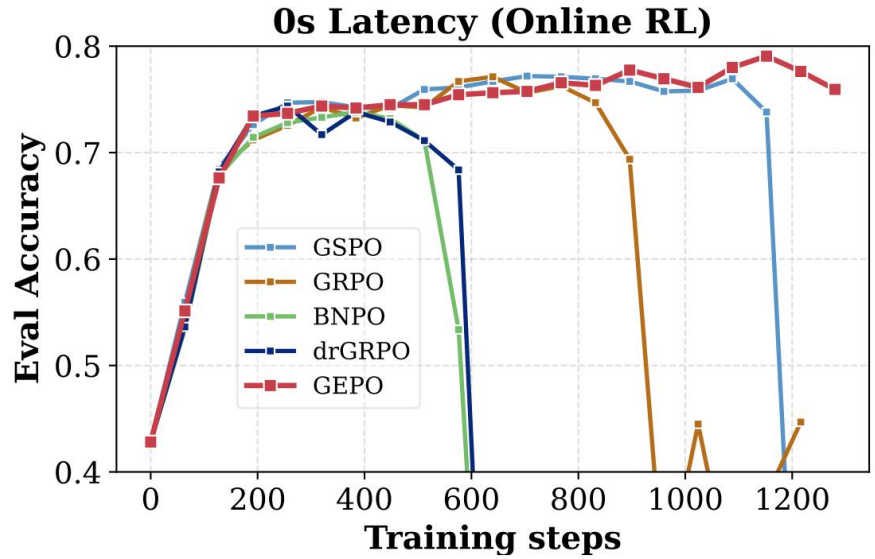
$$\hat{A}_i = \frac{R(x,y^i) - \text{mean}\{R(x,y^1),\ldots,R(x,y^G)\}}{\text{std}\{R(x,y^1),\ldots,R(x,y^G)}$$

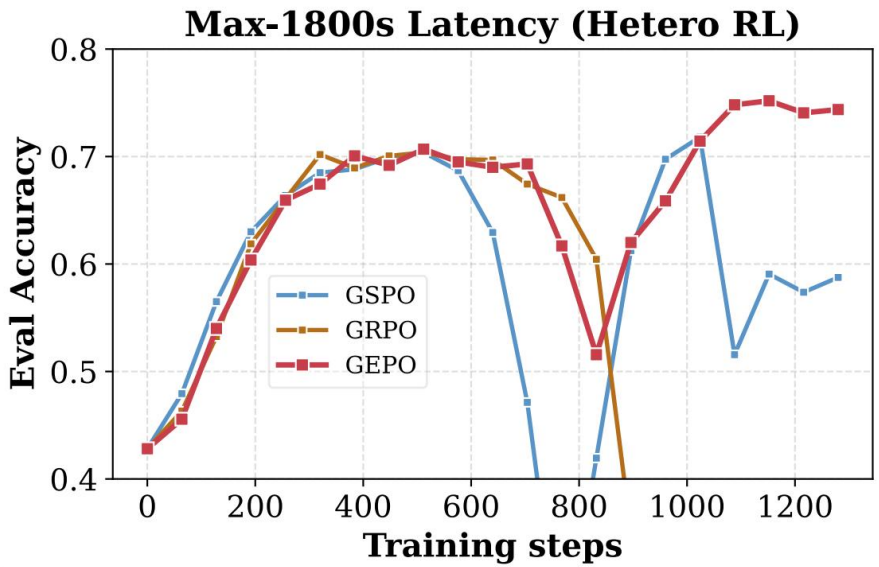
$$\mathcal{L}_{\text{GSPO}_i} = \min\left[\frac{\pi_\theta(y^i \mid x)}{\pi_{\theta_{\text{old}}}(y^i \mid x)}\hat{A}_i, clip_{1\pm\varepsilon}\left[\frac{\pi_\theta(y^i \mid x)}{\pi_{\theta_{\text{old}}}(y^i \mid x)}\right]\hat{A}_i\right]$$

GEPO

$$\hat{A}_i = \frac{R(x,y^i) - \text{mean}\{R(x,y^1),\ldots,R(x,y^G)\}}{\cancel{\text{std}\{R(x,y^1),\ldots,R(x,y^G)}}}$$

$$\mathcal{L}_{\text{GEPO}_i} = \min\left[\frac{\pi_\theta(y^i \mid x)}{\underbrace{\mathbb{E}_{\pi_{\theta_{\text{old}}}(\cdot|x)}\pi_{\theta_{\text{old}}}(y \mid x)}_{\text{Group Expectation}}}\hat{A}_i, clip_{1\pm\varepsilon}\left[\frac{\pi_\theta(y^i \mid x)}{\mathbb{E}_{\pi_{\theta_{\text{old}}}(\cdot|x)}\pi_{\theta_{\text{old}}}(y \mid x)}\right]\hat{A}_i\right]$$





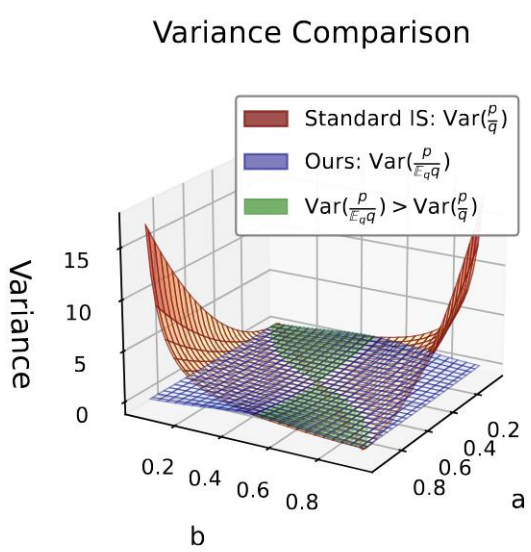
理论支撑

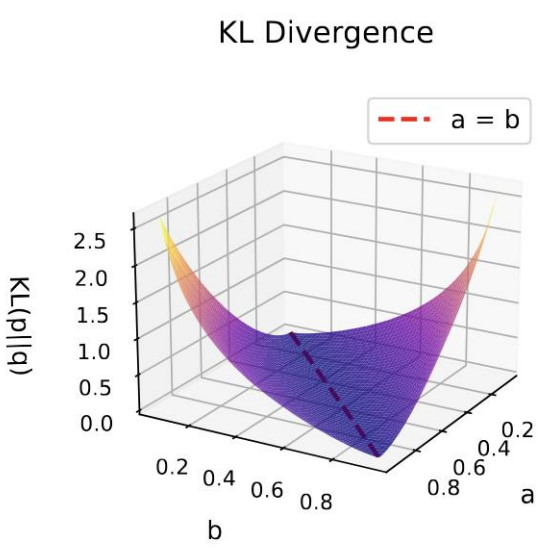
定理 1. 设 p, q 为离散概率分布。则存在常数 C，使得：

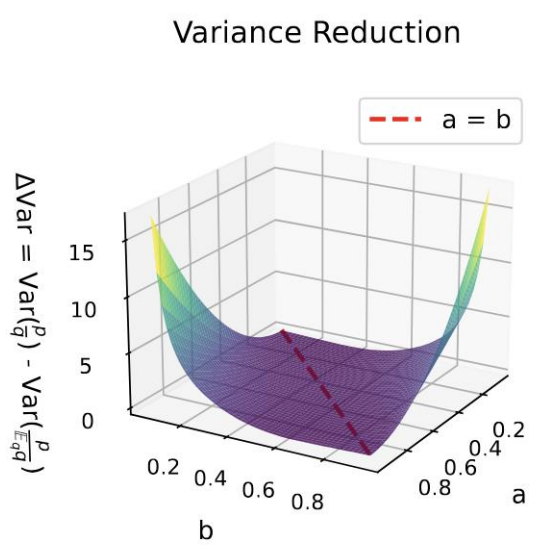
$$Var(\frac{p}{q}) - Var(\frac{p}{\mathbb{E}_q q}) \geq \exp\left(D_{\mathrm{KL}}(p\|q)\right) - C.$$

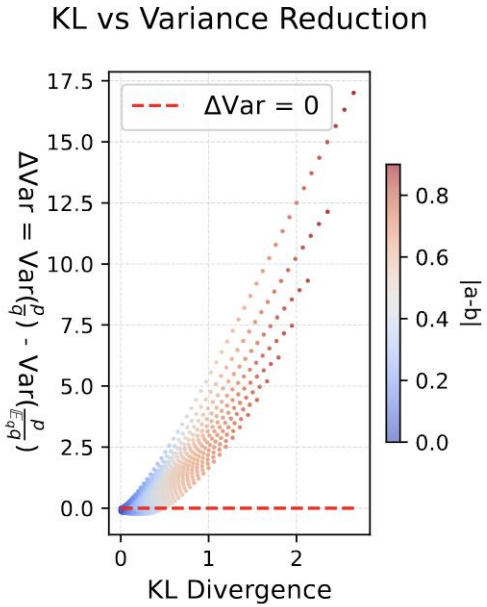
特别地，当 $D_{\mathrm{KL}}(p\|q) > \log C$ 时，有 $Var(\frac{p}{q}) > Var(\frac{p}{\mathbb{E}_q q})$。

可视化结果









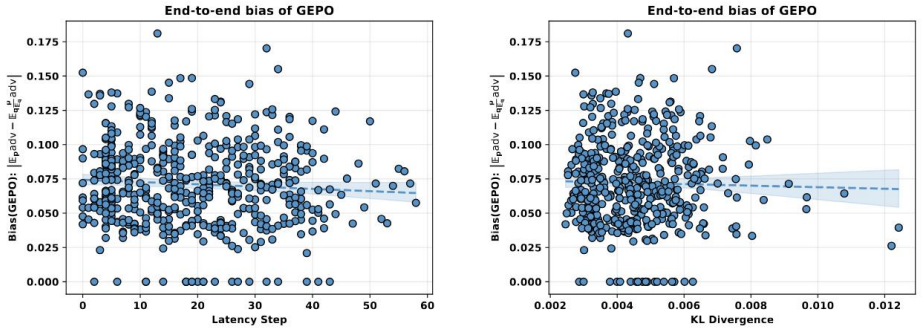
结论： KL散度越大，GEPO的重要性权重方差相比于GSPO/GRPO呈指数衰减

偏差-方差分析

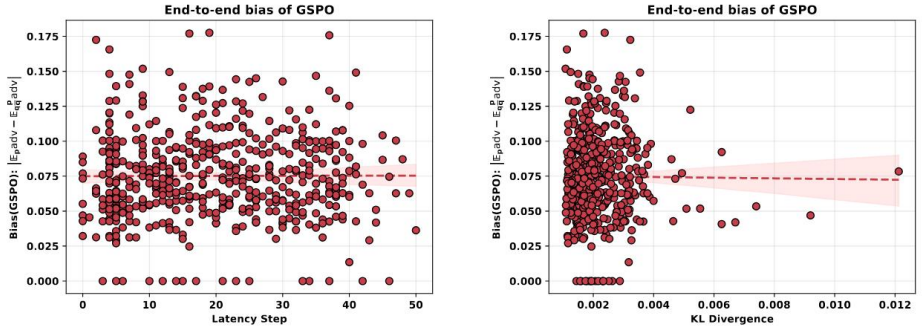
Theorem 2 (Bias of GEPO). *Let $p(y|x)$ and $q(y|x)$ be discrete probability distributions, and let $A(y|x)$ be a bounded advantage function. The bias of GEPO can be bounded by $\frac{\|p\|_2}{\|q\|_2}$:*

$$Bias(GEPO) = \left| \mathbb{E}_p \left[A(y|x) \right] - \mathbb{E}_q \left[\frac{p(y|x)}{\mathbb{E}_q[q]} A(y|x) \right] \right| < \frac{\|p\|_2}{\|q\|_2}, \tag{39}$$

where $\frac{\|p\|_2}{\|q\|_2}$ denotes the L^2-norm ratio of p and q.



(a) Bias-GEPO vs Latency (b) Bias-GEPO vs KL



(c) Bias-GSPO vs Latency (d) Bias-GSPO vs KL

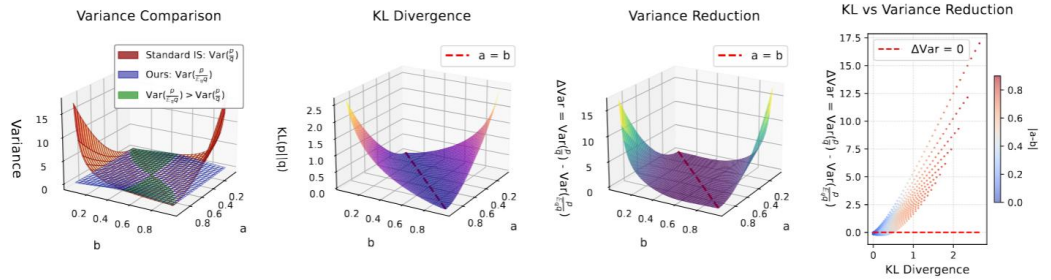
Figure 12: The Bias of GEPO and GSPO. Over 512 global training steps, $Bias(GEPO)$ has a mean of 0.072 and variance of 0.001, while $Bias(GSPO)$ has a mean of 0.075 and variance of 0.001.

Theorem 3 (Variance of GEPO). *Let $p(y|x)$ and $q(y|x)$ be discrete probability distributions, and let $A(y|x)$ be a bounded advantage function. Assume that on the support of interest, $|A(y|x)| \geq A_{\min} > 0$, and define $A_{\max} = \max_y |A(y|x)|$. Then there exists a constant $C_{\text{adv}} = \frac{A_{\max}^2}{\|q\|_2^2}$ such that:*

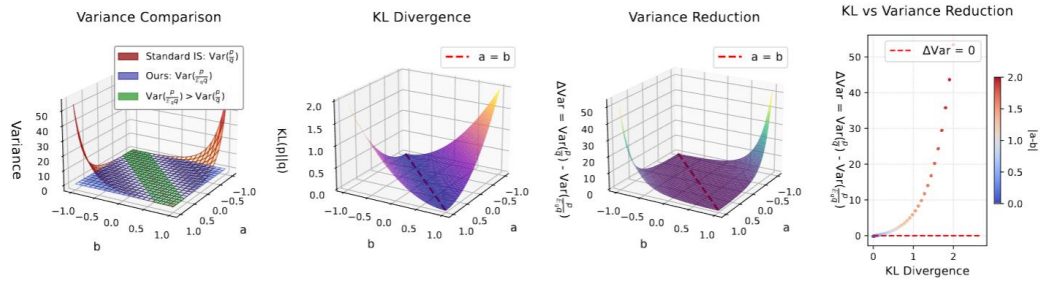
$$\mathrm{Var}_q\left[A(y|x) \cdot \frac{p(y|x)}{q(y|x)}\right] - \mathrm{Var}_q\left[A(y|x) \cdot \frac{p(y|x)}{\mathbb{E}_q[q(y|x)]}\right] \geq \boxed{A_{\min}^2 \cdot \exp\left(D_{\mathrm{KL}}(p\|q)\right)} - C_{\text{adv}}. \tag{48}$$

In particular, when $D_{\mathrm{KL}}(p\|q) > \log\left(\frac{C_{\text{adv}}}{A_{\min}^2}\right)$, it holds that

$$\mathrm{Var}_q\left[A(y|x) \cdot \frac{p(y|x)}{q(y|x)}\right] > \mathrm{Var}_q\left[A(y|x) \cdot \frac{p(y|x)}{\mathbb{E}_q[q(y|x)]}\right].$$



(a) Variance comparison of $\frac{p}{q}$ and $\frac{p}{\mathbb{E}_q[q]}$ under Bernoulli distributions, where $p \sim \text{Bernoulli}(a)$ and $q \sim \text{Bernoulli}(b)$.



(b) Variance comparison of $\frac{p}{q}$ and $\frac{p}{\mathbb{E}_q[q]}$ under Gaussian distributions, where $p \sim \mathcal{N}(a, 1)$ and $q \sim \mathcal{N}(b, 1)$.

Figure 2: In high-KL regions, $\mathrm{Var}[\frac{p(y|x)}{\mathbb{E}_q[q(y|x)}] \ll \mathrm{Var}[\frac{p(y|x)}{q(y|x)}]$.

实现方式与梯度分析

```python
if self.loss_type in ["grpo","dr_grpo","bnpo"]: # Token level
    coef_1 = learner_token_p / sampler_token_p
elif self.loss_type == "gspo":  # Sequence level
    coef_1 = learner_seq_p / sampler_seq_p
elif self.loss_type == "gepo": # Group level
    hat_q = sampler_seq_p.detach() / (sampler_seq_p.sum().detach())
    coef_1 = learner_seq_p / (hat_q * sampler_seq_p).sum()
```

Listing 1: Coefficient computation for different policy optimization methods

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{A} \odot \underbrace{\begin{bmatrix} \frac{p'_{1,1}(\boldsymbol{\theta})}{q_{1,1}} & \cdots & \frac{p'_{1,T}(\boldsymbol{\theta})}{q_{1,T}} \\ \vdots & \ddots & \vdots \\ \frac{p'_{G,1}(\boldsymbol{\theta})}{q_{G,1}} & \cdots & \frac{p'_{G,T}(\boldsymbol{\theta})}{q_{G,T}} \end{bmatrix}}_{\texttt{GRPO}} \texttt{or} \underbrace{\begin{bmatrix} \frac{p'_{1,1}(\boldsymbol{\theta})}{q_{1}} & \cdots & \frac{p'_{1,T}(\boldsymbol{\theta})}{q_{1}} \\ \vdots & \ddots & \vdots \\ \frac{p'_{G,1}(\boldsymbol{\theta})}{q_{G}} & \cdots & \frac{p'_{G,T}(\boldsymbol{\theta})}{q_{G}} \end{bmatrix}}_{\texttt{GSPO}} \texttt{or} \underbrace{\begin{bmatrix} \frac{p'_{1,1}(\boldsymbol{\theta})}{\mathbb{E}_q q} & \cdots & \frac{p'_{1,T}(\boldsymbol{\theta})}{\mathbb{E}_q q} \\ \vdots & \ddots & \vdots \\ \frac{p'_{G,1}(\boldsymbol{\theta})}{\mathbb{E}_q q} & \cdots & \frac{p'_{G,T}(\boldsymbol{\theta})}{\mathbb{E}_q q} \end{bmatrix}}_{\texttt{GEPO (ours)}}, \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{G \times T}$ is token-level advantages matrix, $\odot$ denotes Hadamard product, $q_{i,t} = q(y_t^i \mid x^i, y_{<t}^i)$, $q_i = q(y^i \mid x)$, and $\mathbb{E}_q q = \widehat{\mathbb{E}}_q[q(y|x)]$. From the perspective of gradient

Method	AMC2023		AIME2024		AIME2025		MATH500		Average	
	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last
Qwen3-1.7B	25.6	-	1.6	-	3.9	-	54.7	-	21.5	-
Max Tolerable Delay 0 (Online RL)										
BNPO	54.3	0.0	18.4	0.0	19.1	0.0	78.7	0.0	42.6	0.0
Dr.GRPO	53.4	14.3	19.1	1.6	18.8	2.0	78.6	35.9	42.5	13.5
GRPO	56.3	23.4	20.7	0.4	19.9	2.3	79.8	49.7	44.2	19.0
GSPO	54.1	27.8	**23.8**	3.1	**20.7**	4.3	79.9	62.1	44.6	24.3
GEPO (ours)	**56.9**	**56.9**	21.9	**16.4**	20.3	**14.1**	**80.4**	**78.1**	**44.9**	**41.4**
Max Tolerable Delay 64 (Hetero RL)										
BNPO	45.0	43.1	12.1	11.3	12.5	10.1	71.1	69.3	35.2	33.5
Dr.GRPO	48.4	48.4	17.2	17.2	14.8	14.8	73.9	73.9	38.6	38.6
GRPO	46.6	46.6	19.1	14.5	14.8	14.8	74.9	74.9	38.9	37.7
GSPO	**54.4**	23.8	17.6	1.6	17.6	2.7	78.2	55.6	42.0	20.9
GEPO (ours)	53.8	**53.8**	**21.9**	**21.9**	**18.8**	**18.8**	**79.6**	**79.6**	**43.5**	**43.5**

与异步RL算法对比

异步RL方法从**权重截断**的角度进行优化，训练结果表明在64步延迟场景下仍然存在不稳定风险

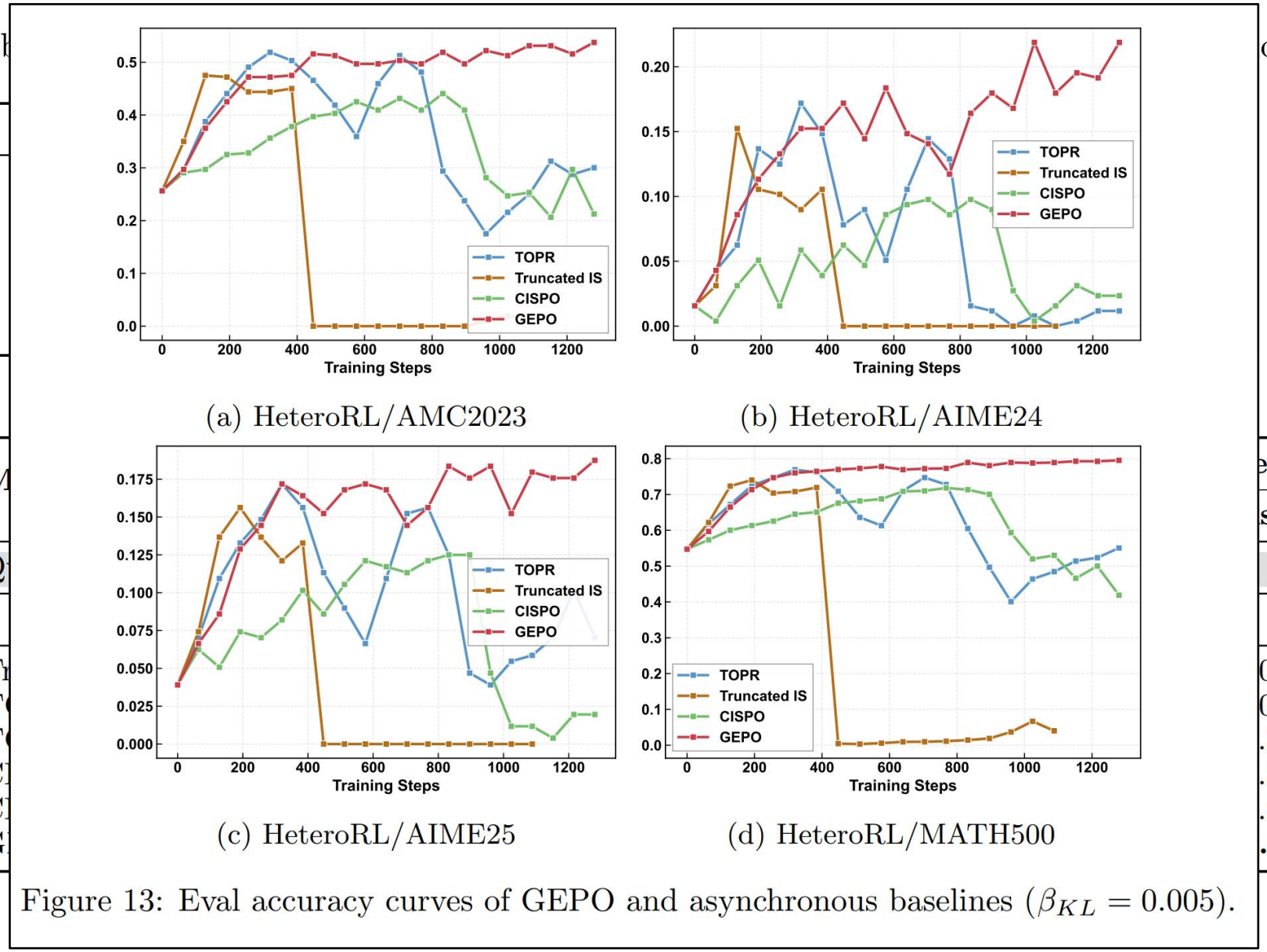
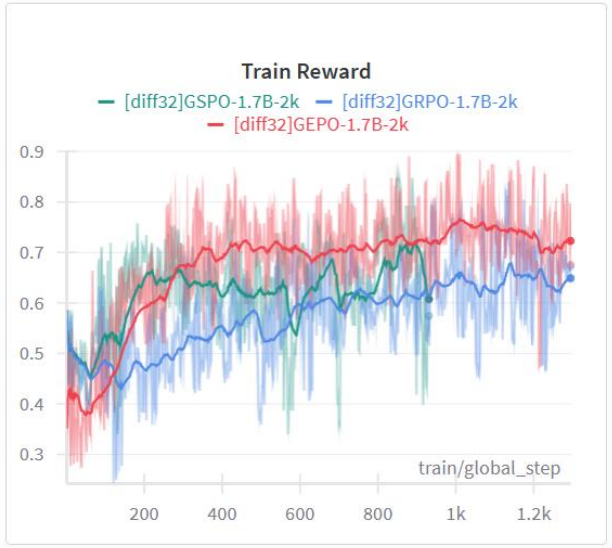


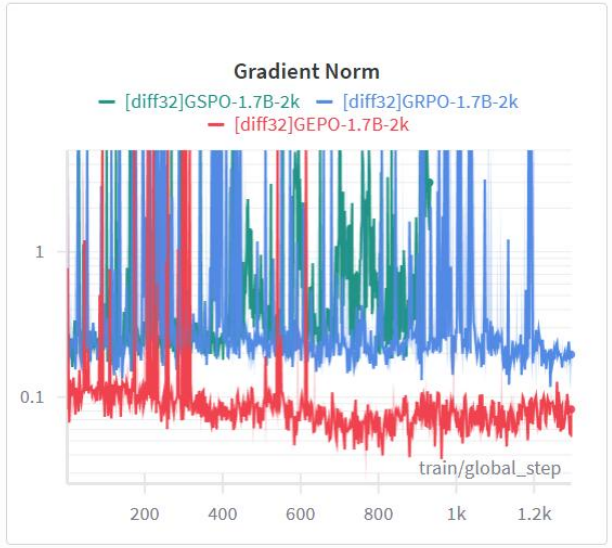
Figure 13: Eval accuracy curves of GEPO and asynchronous baselines ($\beta_{KL} = 0.005$).

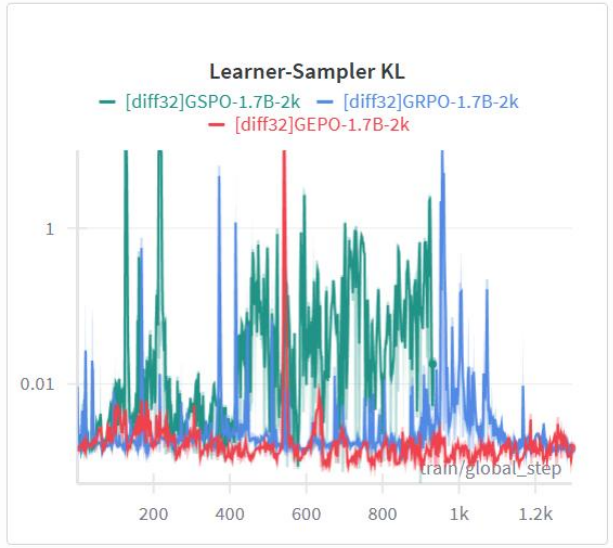
训练过程对比

主流方法

GEPO vs GSPO vs GRPO

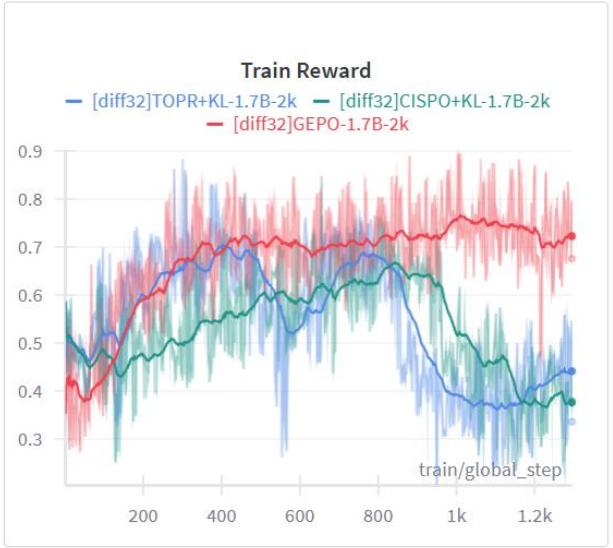


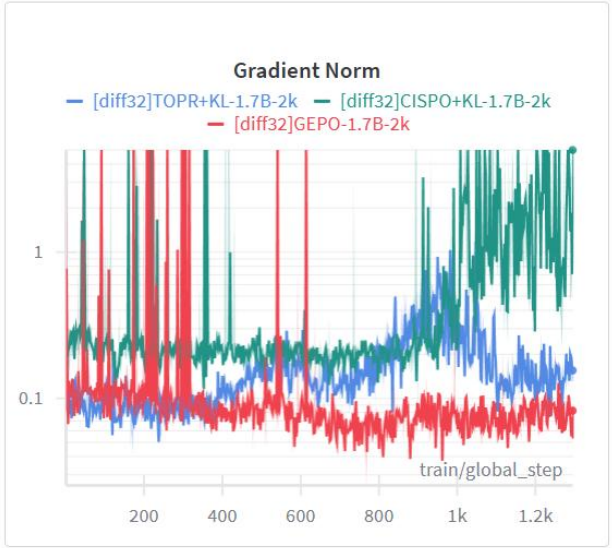


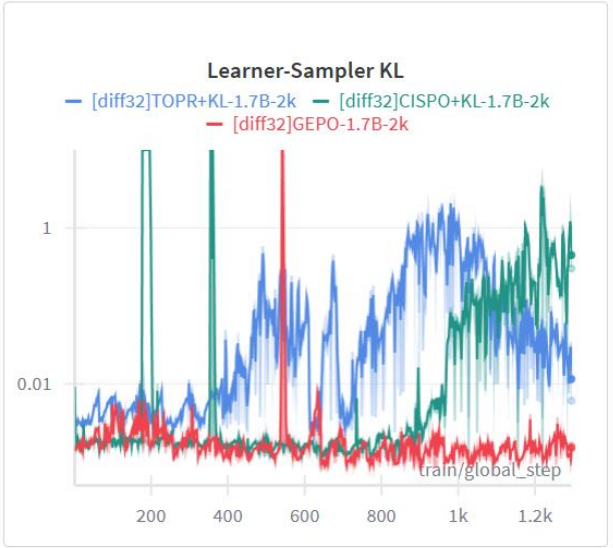


异步方法

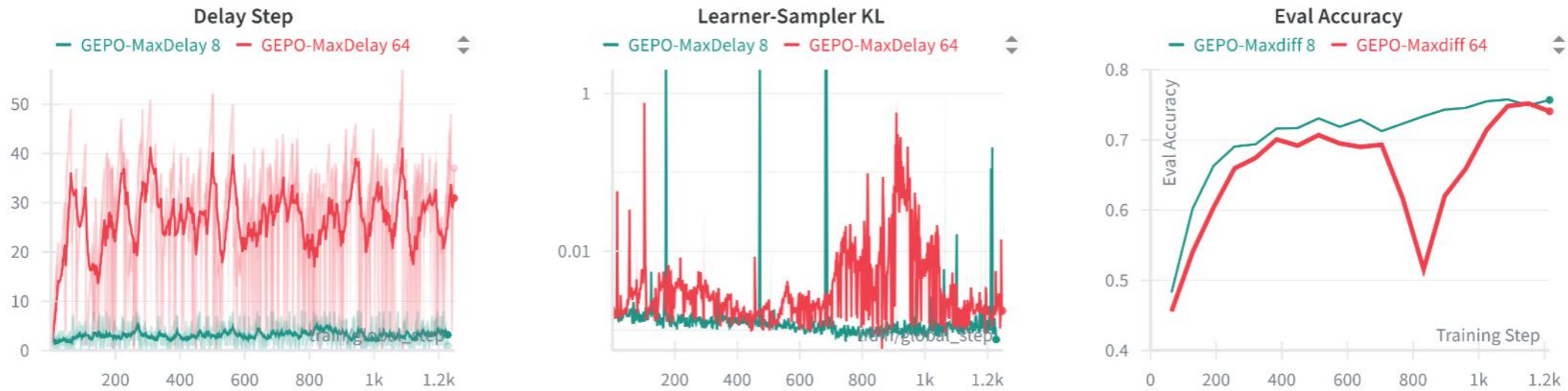
GEPO vs TOPR vs CISPO







高延迟的挑战 为了证明高延迟会降低训练稳定性，我们对比了最大延迟为 8 个 step 和 64 的 step 的训练过程。如图 8所示，在最大延迟为 64 的情况下，尤其是延迟达到最大时（900 步左右），KL 散度快速变大，评估精度显著降低。这个实验证明了我们关于延迟会加大不稳定性的猜想。尽管 GEPO 相比 GRPO 等方法可以显著提升训练稳定性，其仍然会受到延迟造成的影响（900 步左右的性能下降），说明高延迟场景下的异构强化学习仍是一个有挑战的研究方向。更好的方法理应可以将低延迟对训练造成的负面影响，使训练过程更加稳定。



(a) 采样器的延迟步数　(b) KL 散度变化　(c) 评估精度变化曲线

图 8: 不同延迟下训练过程

项目开源

技术报告

arxiv:

https://arxiv.org/abs/2508.17850

启智社区

https://openi.pcl.ac.cn/Hanlard/HeteroRL.git

开源社区

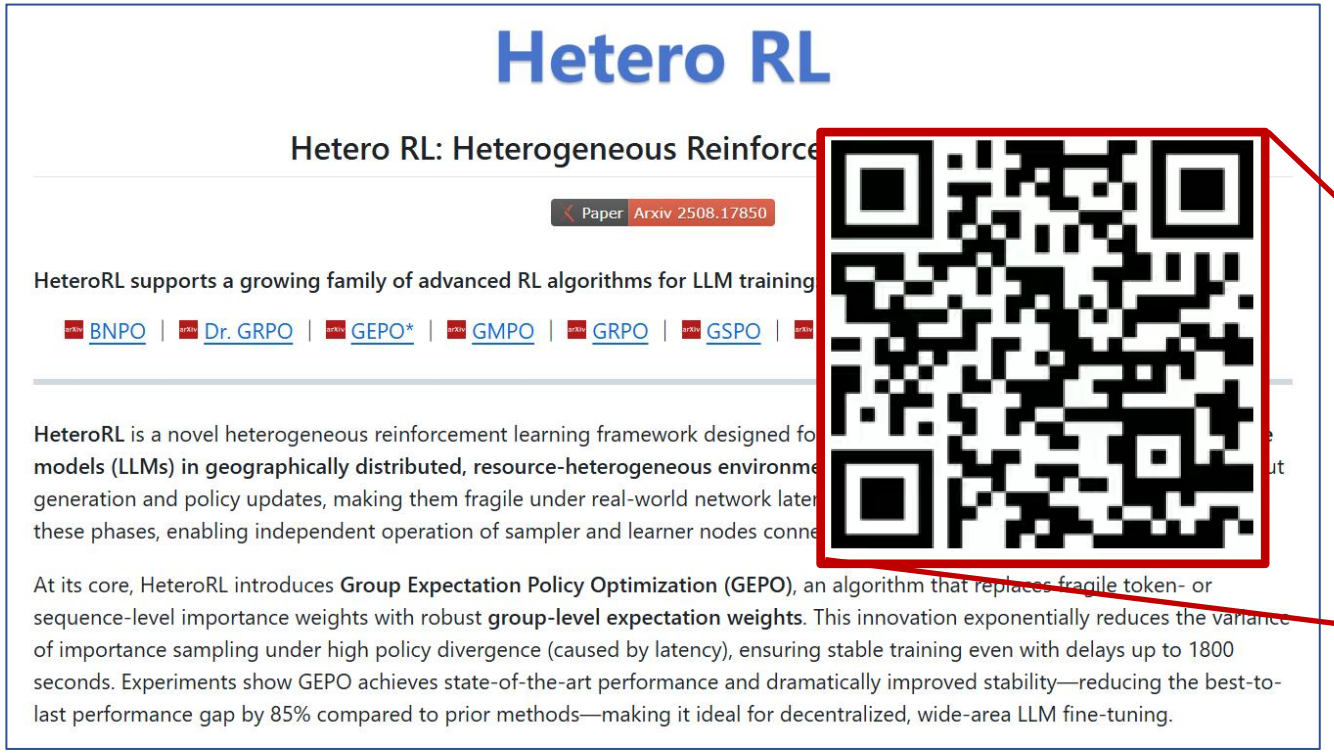
https://github.com/HanlardResearch/Hetero-RL.git

Pengcheng Laboratory / Heterogeneous Large Model Research Team

GEPO: GROUP EXPECTATION POLICY OPTIMIZATION FOR STABLE HETEROGENEOUS REINFORCEMENT LEARNING

Han Zhang*, Ruibin Zheng*, Zexuan Yi, Zhuo Zhang, Hanyang Peng, Hui Wang, Zike Yuan, Cai Ke, Shiwei Chen, Jiacheng Yang, Yangning Li, Xiang Li, Jiangyue Yan, Yaoqi Liu, Liwen Jing, Jiayin Qi, Ruifeng Xu, Binxing Fang, Yue Yu†

https://github.com/HanlardResearch/Hetero-RL.git



Hetero RL

Hetero RL: Heterogeneous Reinforce

Paper Arxiv 2508.17850

HeteroRL supports a growing family of advanced RL algorithms for LLM training

BNPO | Dr. GRPO | GEPO* | GMPO | GRPO | GSPO |

HeteroRL is a novel heterogeneous reinforcement learning framework designed fo
models (LLMs) in geographically distributed, resource-heterogeneous environme
generation and policy updates, making them fragile under real-world network late
these phases, enabling independent operation of sampler and learner nodes conne

At its core, HeteroRL introduces **Group Expectation Policy Optimization (GEPO)**, an algorithm that replaces fragile token- or sequence-level importance weights with robust **group-level expectation weights**. This innovation exponentially reduces the variance of importance sampling under high policy divergence (caused by latency), ensuring stable training even with delays up to 1800 seconds. Experiments show GEPO achieves state-of-the-art performance and dramatically improved stability—reducing the best-to-last performance gap by 85% compared to prior methods—making it ideal for decentralized, wide-area LLM fine-tuning.

Text Note

盘点各类GxPO的梯度估计方差降低思路

共13张

酸楂

盘点各类GxPO的梯度估计方差降低思路

降低梯度估计方差是RL方法稳定训练的关键，各类GxPO的改进方法最终都回归到降低训练的梯度估计的方差，本帖从降低方差的角度盘点一下近半年来各类GxPO方法的思路

#强化学习 #大模型 #算法 #科研学习 #秋招面试 #面试技巧 #面试题 #面试热点

2025-09-26 16:11 156 298 11

小红书

长按扫描二维码查看笔记

参考文献

[1] GRPO: DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

[2] GSPO: Group Sequence Policy Optimization

[3] TOPR: TAPERED OFF-POLICY REINFORCE Stable and efficient reinforcement learning for LLMs

[4] CISPO: MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention

[5] BNPO: Beta Normalization Policy Optimization

[6] DrGRPO: Understanding R1-Zero-Like Training: A Critical Perspective

谢 谢 观 看

欢迎学术合作