Stabilizing Policy Gradients for Sample-Efficient RL in LLM Reasoning

Luckeciano Melo* Alessandro Abate Yarin Gal

University of Oxford

Motivation and Context

- The emergence of reasoning in LLMs represents a major shift in AI research.

- This progress is primarily attributed to advances in scaling RL techniques for LLM post-training, particularly policy gradient methods such as GRPO

- However, RL still faces fundamental challenges that limit its broader practicality and scalability
  - Policy gradients suffer from optimization instabilities driven by the **non stationary RL objective**, **high variance of estimates**, and **pathologies of training deep networks**

Motivation and Context

- These factors lead to several undesired consequences
  - Policy collapse, plasticity loss, sample inefficiency, hyperparameter sensitivity
  - These challenges are potentially more pronounced in the LLM setting

- To overcome these challenges, practitioners typically rely on conservative training regimes to ensure stability, increasing computational costs

- **Stabilizing these algorithms in sample-efficient regimes is crucial to further scale RL for LLM reasoning.**

Key Idea and Contribution

- Explicitly model second-order geometry in the optimization landscape and incorporate this information into policy updates
  - Hessian of the RL objective

$$J(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \underbrace{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})^\top \Delta\boldsymbol{\theta} + \tfrac{1}{2}\, \Delta\boldsymbol{\theta}^\top H(\boldsymbol{\theta})\, \Delta\boldsymbol{\theta}}_{m_H(\Delta\boldsymbol{\theta})} + \mathcal{O}(\|\Delta\boldsymbol{\theta}\|^3)$$

  - Fisher Information Matrix of the policy distribution

$$\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}} \,\|\, \pi_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}) = \underbrace{\tfrac{1}{2}\Delta\boldsymbol{\theta}^\top F(\boldsymbol{\theta})\Delta\boldsymbol{\theta}}_{m_F(\Delta\boldsymbol{\theta})} + \mathcal{O}(\|\Delta\boldsymbol{\theta}\|^3)$$

- We introduce a tractable computational model that estimates these quantities
- Design an algorithm (CAPO) which leverages them to anticipate policy updates that potentially induce sudden shifts in the objective or policy distribution

Methodology

- Computational Model
  - For an LLM with d parameters, Hessian/FIM are d x d matrices
  - **Last-Layer Model**: We only model curvature from the pre-softmax layer
    - Fewer parameters, analytical gradients/curvatures
  - **Direct compute *directional* curvatures**
    - e.g., $\Delta\boldsymbol{\theta}^{\top} F(\boldsymbol{\theta})\Delta\boldsymbol{\theta}$ instead of only $F(\boldsymbol{\theta})$
    - Reduce memory complexity from quadratic to linear w.r.t to last-layer parameters
  - **Exploit Gradient Sparsity:**
    - Standard top-k/nucleus sampling leads to sparse gradients
  - **Model optimizer**
    - Model SGD vs Adam updates under the last-layer model

Methodology

- Curvature-Aware Policy Optimization (CAPO)
  - The curvature estimates allows computing the objective/policy shifts:

$$m_H(\boldsymbol{\psi}) = \tilde{g}(\boldsymbol{\psi})^\top \Delta\boldsymbol{\psi} + \tfrac{1}{2}\Delta\boldsymbol{\psi}^\top \tilde{H}(\boldsymbol{\psi})\Delta\boldsymbol{\psi}, \quad m_F(\boldsymbol{\psi}) = \tfrac{1}{2}\Delta\boldsymbol{\psi}^\top \tilde{F}(\boldsymbol{\psi})\Delta\boldsymbol{\psi}$$

  - CAPO uses this for data selection: mask out tokens that cause abrupt/unstable shifts under the model:

$$\delta_H \leq m_H(\Delta\psi_i) \ \leq \ \delta_H^{high}, \qquad m_F(\Delta\psi_i) \ \leq \ \delta_F$$

- We prove that, under reasonable assumptions, **CAPO presents monotonic policy improvement**

Theorem 5.1 (Monotonic improvement under CAPO). *Fix thresholds $\delta_H > 0$ and $\delta_F > 0$. Let $\mathcal{B}$ be a batch of sampled trajectories. Split $\mathcal{B}$ into disjoint N subsets $b_i \subset \mathcal{B}$, and propose candidate subset updates $\{\Delta\theta_i\}_{i:N}$. Retain those satisfying:*

$$m_H(\Delta\theta_i) \ \geq \ \delta_H = \omega + \tfrac{1}{2}Mr^2, \qquad m_F(\Delta\theta_i) \ \leq \ \delta_F, \tag{12}$$

with $\omega > 0$ and M, r defined as in Assumption E.1. Let $\mathcal{B}_{acc}$ denote the superset of the B accepted subsets, and define the aggregated update: $\Delta\theta = \frac{1}{B}\sum_{i\in\mathcal{B}_{acc}} \Delta\theta_i$. Then, for two policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}$, with $|A^\pi(s,a)| \leq \epsilon$, we obtain:

$$J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta) \ \geq \ \omega \ - \ C\sqrt{\delta_F}, \quad C = \frac{2\gamma}{(1-\gamma)^2}\epsilon\sqrt{2}. \tag{13}$$

Thus choosing $\omega \geq C\sqrt{\delta_F}$ guarantees monotonic improvement: $J(\pi_{\theta+\Delta\theta}) \geq J(\pi_\theta)$.

Results

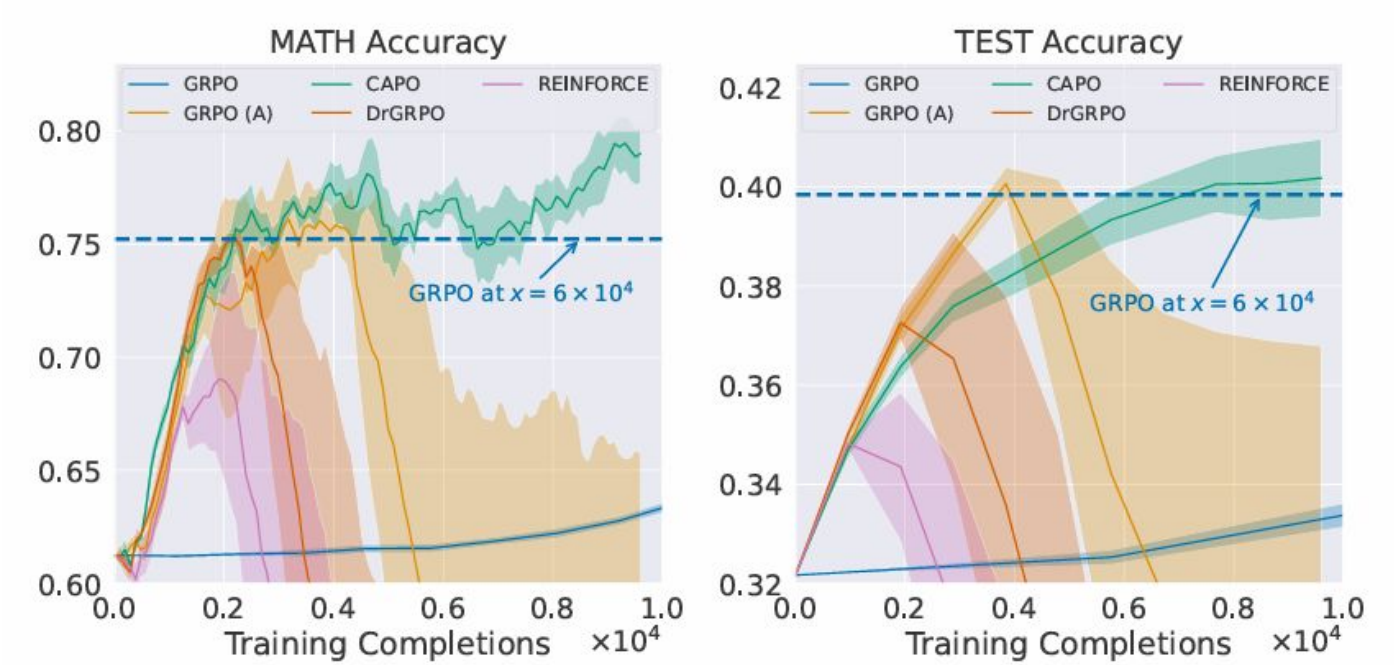
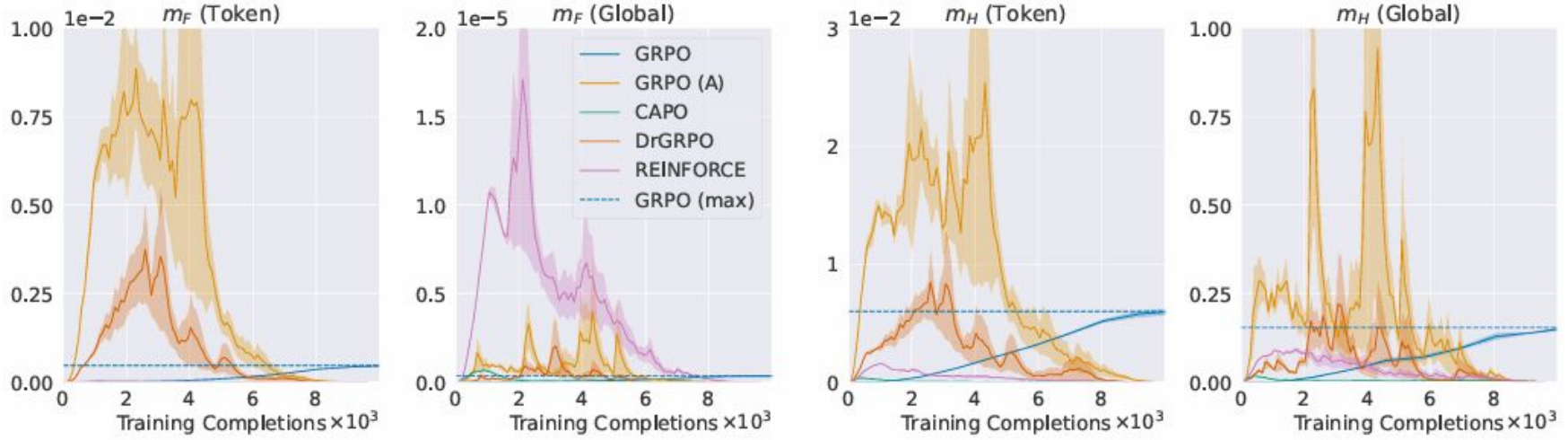
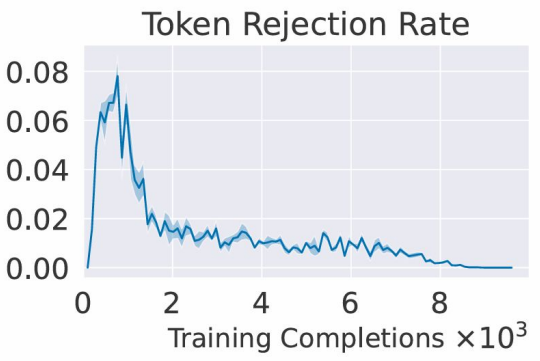


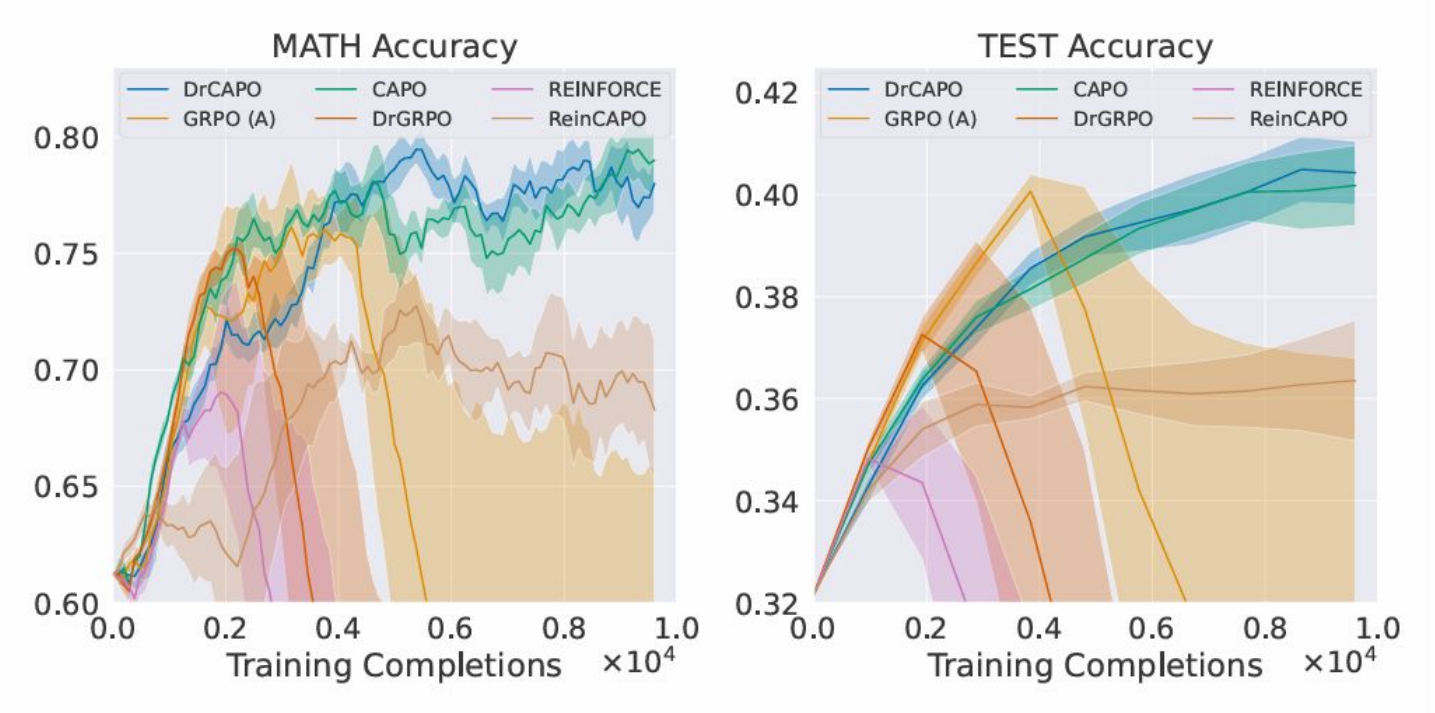
Figure 2: **Comparison with baseline methods on policy gradient stability.** While the setup with more aggressive updates makes all methods more sample-efficient, it also leads the baselines to policy collapse. In contrast, CAPO prevents collapse and achieves up to 30× greater sample efficiency than GRPO under aggressive updates.

Results


1e−2
m_F (Token)
1.00
0.75
0.50
0.25
0.00
0
2
4
6
8
Training Completions ×10³
1e−5
m_F (Global)
2.0
1.5
1.0
0.5
0.0
GRPO
GRPO (A)
CAPO
DrGRPO
REINFORCE
GRPO (max)
1e−2
m_H (Token)
3
2
1
0
m_H (Global)


Token Rejection Rate
0.08
0.06
0.04
0.02
0.00
0
2
4
6
8
Training Completions ×10³

Results


MATH Accuracy
DrCAPO
GRPO (A)
CAPO
DrGRPO
REINFORCE
ReinCAPO
0.80
0.75
0.70
0.65
0.60
0.0
0.2
0.4
0.6
0.8
1.0
Training Completions
×10^4
TEST Accuracy
DrCAPO
GRPO (A)
CAPO
DrGRPO
REINFORCE
ReinCAPO
0.42
0.40
0.38
0.36
0.34
0.32
0.0
0.2
0.4
0.6
0.8
1.0
Training Completions
×10^4

Stabilizing Policy Gradients for Sample-Efficient Reinforcement Learning in LLM Reasoning

Luckeciano Carvalho Melo · Alessandro Abate · Yarin Gal

Luckeciano Melo Alessandro Abate Yarin Gal

University of Oxford