Not All Rollouts are Useful: Down-Sampling Rollouts in LLM Reinforcement Learning

Yixuan (Even) Xu Yash Savani Fei Fang J. Zico Kolter **Carnegie Mellon University**

Contact us: yixuanx@cs.cmu.edu

ICML International Conference On Machine Learning

TRANSACTIONS tmlr on ML RESEARCH

RLVR training has ***two phases***

Inference	Policy Update
To generate rollouts	To update parameters
Embarrassingly **parallel**	Needs **synchronization**
Modest in memory	**Intense** in memory

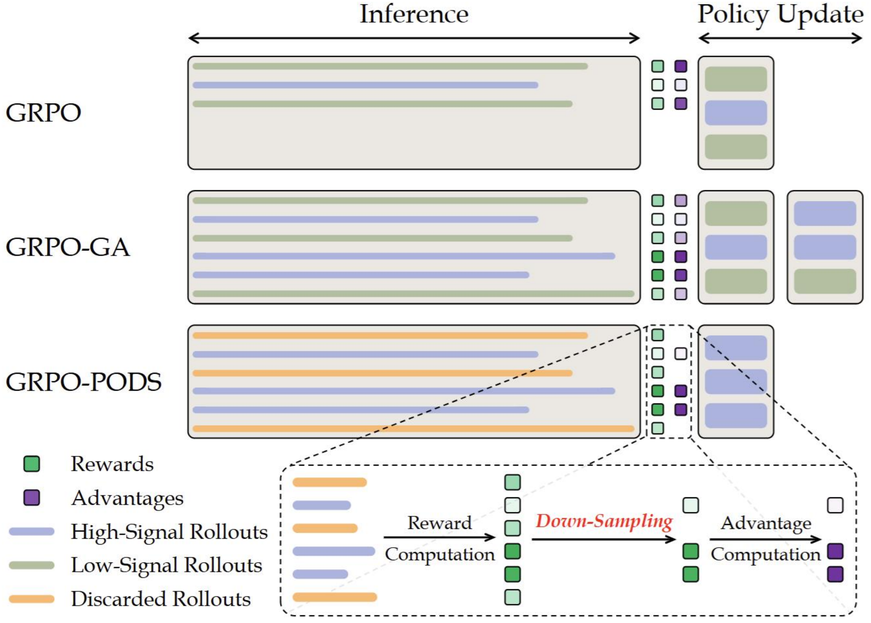
To tackle the ***computation asymmetry***

Grad Accumulation	PODS
Splitting rollouts into multiple update steps	Strategically **discarding** some of the rollouts
Fully utilizes the GPU at inference time	Also fully utilizes the GPU at inference time
All generate rollouts contribute to updates	*"Not all rollouts are useful"* Only uses the selected

Max-Variance Down-Sampling

Choose the subset **maximizing variance** in rewards
Intuition: Captures both **positive** & **negative** learning signals
Example: {**0.2**, 0.4, 0.6, **0.8**} → {**0.2**, **0.8**}

Theorem: This set contains k highest & $m-k$ lowest rewards
Which gives us an $O(n \log n)$ algorithm for computing this set



Experiments

We evaluate PODS across diverse **hardware** settings, model **architectures**, model **scales**, and **domains**
(a) to (d) compares **GRPO** with **GRPO-PODS**
(e) to (f) compares **GRPO-GA** with **GRPO-PODS**

Setting	Benchmark	Model	Parameters	GPUs	Fine-tuning Method
(a)	GSM8K	Qwen2.5	3B	1 L40S	LoRA (rank 64, $\alpha = 64$)
(b)	GSM8K	Llama3.2	3B	1 L40S	LoRA (rank 64, $\alpha = 64$)
(c)	MATH	Qwen2.5	3B	1 L40S	LoRA (rank 64, $\alpha = 64$)
(d)	Chemistry	Qwen2.5	3B	1 L40S	LoRA (rank 64, $\alpha = 64$)
(e)	GSM8K	Qwen2.5	3B	8 H100s	Full-Parameter
(f)	GSM8K	Qwen2.5	7B	8 A100s	Full-Parameter

With PODS, RL ***converges faster,*** *and often to a* ***higher accuracy***