Group-Normalized Implicit Value Optimization for Language Models

Yunseon Choi, Junyoung Jang, Chaeyoung Oh, Minchan Jeong,
Doohwan Hwang, Kee-Eung Kim

KAIST

National AI Research Lab

ICLR

The Challenge of Step-level Alignment

- Traditional Reinforcement Learning (RL) treats the entire generated response as a single action.
- This approach provides a weak learning signal for multi-step tasks like mathematical reasoning.
- Existing solutions often require a complex "critic" network, adding significant computational overhead and memory costs.

KL-Regularized Policy Optimization

▪ KL-regularized policy optimization is a widely used approach for refining language model.

▪ The goal is to train a policy π_θ that maximizes the expected reward while staying close to a reference policy $\pi_{\theta_{\text{old}}}$,

$$\max_\theta \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[R(x,y) - \alpha \log \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}\right]$$

where x is query and y is the completion generated by the policy.

- The optimal policy has a known closed-form solution:

$$\pi_{\theta^\star}(y|x) = \frac{\pi_{\theta_{\text{old}}}(y|x)\mathrm{e}^{R(x,y)/\alpha}}{Z(x)}$$

where $Z(x) = \mathbb{E}_{\pi_{\theta_{\text{old}}}(y|x)}\left[\mathrm{e}^{R(x,y)/\alpha}\right]$ is the partition function.

KL-Regularized RL with Sparse Reward

- In language tasks, the reward is sparse, typically given only upon completion of the entire generation.

Question: *Let $T = 11$. Compute the value of x that satisfies $\sqrt{20+\sqrt{T+x}} = 5$.* x

step1: Restate the problem.
We need to find the value of x that satisfies the equation $\sqrt{20+\sqrt{T+x}} = 5$, where $T = 11$.

step2: Substitute $T = 11$ into the equation.
The equation becomes $\sqrt{20+\sqrt{11+x}} = 5$.

... $\pi_{\theta^\star}(y_{<t}|x) \propto \pi_{\theta_{\text{old}}}(y_{<t}|x)$ {??} $y_{<t}$

step7: Verify the solution by substituting $x = 14$ back into the original equation.

$$\sqrt{20+\sqrt{11+14}} = \sqrt{20+\sqrt{25}} = \sqrt{20+5} = \sqrt{25} = 5$$

The solution satisfies the original equation.
Therefore, the value of x is $\boxed{14}$. y

$\pi_{\theta^\star}(y|x) \propto \pi_{\theta_{\text{old}}}(y|x)\mathrm{e}^{R(x,y)/\alpha}$

Toward Step-level Value

- (Theorem) Suppose that $\pi_{\theta^\star}$ and $\pi_{\theta_{\text{old}}}$ are autoregressive policies, and that satisfy the optimal condition for complete response y. For any t, $y_{<t}$ distribution of these policies satisfy:

$$\pi_{\theta^\star}(y_{<t}|x) = \frac{\pi_{\theta_{\text{old}}}(y_{<t}|x)\mathrm{e}^{V(x,y_{<t})}}{Z(x)}$$

where a soft value function $V(x, y_{<t})$, which represents the expected future reward from $y_{<t}$, is defined as:

$$V(x, y_{<t}) := \begin{cases} R(x,y)/\alpha & t = T, \\ \log \mathbb{E}_{\pi_{\theta_{\text{old}}}(y|y_{<t},x)}\left[\mathrm{e}^{R(x,y)/\alpha}\right] & t < T. \end{cases}$$

Toward Critic-free Algorithm

- While the soft value function can be modeled with an explicit network, we pursue a direct policy optimization that avoids explicit value modeling.
- We can express the value function implicitly in terms of a policy ratio:

$$V(x, y_{<t}) = \log Z(x) + \log \frac{\pi_{\theta^\star}(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)}$$

- The mean-squared error (MSE)[1][2] between the policy-defined value and the true value function can be an objective for direct policy training loss.

$$\mathcal{L}_{\text{MSE}}(Z,\theta) = \left(\log Z(x) + \log \frac{\pi_\theta(y_{<t}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}|x)} - \log \mathbb{E}_{y\sim\pi_{\theta_{\text{old}}}(\cdot|y_{<t},x)}\left[e^{R(x,y)/\alpha}\right]\right)^2$$

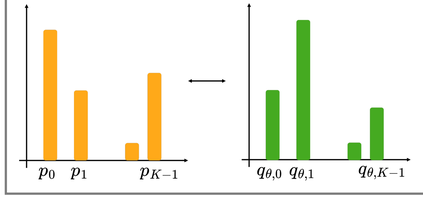
→ A key challenge remains: the partition function still requires an auxiliary network to approximate.

The Group-normalized objective for V_θ

$$\min_{V_\theta} \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\,|\,x)}}\left[\left(\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}\right)\cdot -\left(\sum_{i=0}^{K-1} \underbrace{\frac{e^{V\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}}}_{p_i} \log \underbrace{\frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}}}_{q_{\theta,i}}\right)\right]$$

Cross-Entropy Loss

$(p_i = q_{\theta,i} \quad \text{for } \forall\, i)$



$$\frac{e^{V\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}} = \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}}$$

$$\rightarrow \frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{e^{V\left(x,y_{<t}^{(i)}\right)}} = \frac{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}}{\sum_{j=1}^{K-1} e^{V\left(x,y_{<t}^{(j)}\right)}} = C_t(x)$$

- *(Theorem 2, Consistency up to constant shift)* Assume unlimited model capacity and data. For any $K > 1$ and $t \in \{1,\cdots,T\}$, the minimizer V_{θ^*} of *the group normalized objective* recover the soft value function V up to an additive, $y_{<t}$-independent offset $C_t(x)$:

$$V_{\theta^*}(x, y_{<t}) = V(x, y_{<t}) + \log C_t(x),$$

equivalently, $e^{V_{\theta^*}(x,y_{<t})} = C_t(x)e^{V(x,y_{<t})}$.

Group-normalized Implicit Value Optimization

$$e^{V_{\theta^*}(x,y_{<t})} = C_t(x)Z(x)\frac{\pi_{\theta^*}(y_{<t}\,|\,x)}{\pi_{\theta_{\text{old}}}(y_{<t}\,|\,x)} \quad \rightarrow \frac{\cancel{C_t(x)}\cancel{Z(x)}\frac{\pi_\theta(y_{<t}^{(i)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(i)}|x)}}{\sum_{j=0}^{K-1}\cancel{C_t(x)}\cancel{Z(x)}\frac{\pi_\theta(y_{<t}^{(j)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(j)}|x)}} = \frac{\frac{\pi_\theta(y_{<t}^{(i)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(i)}|x)}}{\sum_{j=0}^{K-1}\frac{\pi_\theta(y_{<t}^{(j)}|x)}{\pi_{\theta_{\text{old}}}(y_{<t}^{(j)}|x)}}$$

$$\min_{V_\theta} \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y^{(0:K-1)},\, y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\,|\,x)}}\left[-\sum_{i=0}^{K-1} e^{R(x,y^{(i)})/\alpha}\log\frac{e^{V_\theta\left(x,y_{<t}^{(i)}\right)}}{\sum_{j=0}^{K-1} e^{V_\theta\left(x,y_{<t}^{(j)}\right)}}\right]$$

$$\mathcal{L}_{GN-IVO}(\theta) = \mathbb{E}_{\substack{x\sim\mathcal{D},t\sim U\{1,\ldots,T\},\\ y_{<t}^{(0:K-1)}\sim\pi_{\text{old}}(\cdot\,|\,x)}}\left[-\sum_{i=0}^{K-1} e^{R(x,y^{(i)})/\alpha}\left(\log\frac{\pi_\theta\left(y_{<t}^{(i)}\,\middle|\,x\right)}{\pi_{\theta_{\text{old}}}\left(y_{<t}^{(i)}\,\middle|\,x\right)} - \log\sum_{j=0}^{K-1}\frac{\pi_\theta\left(y_{<t}^{(j)}\,\middle|\,x\right)}{\pi_{\theta_{\text{old}}}\left(y_{<t}^{(j)}\,\middle|\,x\right)}\right)\right]$$

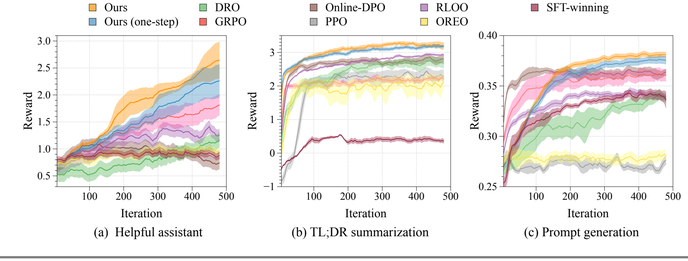
Algorithm 1 Group-Normalized Implicit Value Optimization

Input: Reward function R, learning rate η, the policy π_θ and set $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
for iterations **do**
  Sample a query $x \sim \mathcal{D}$ and generate K responses $y^{(0:K)} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$
  Evaluate reward $R(x, y^{(i)})$ for all $i \in \{0, \ldots, K-1\}$
  Update θ by optimizing the following loss in Eq. 9, $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{GN-IVO}}(\theta)$
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
end for

Table 1. Comparison of our method against baselines on the math reasoning task. The Pass@3 (P@3) metric is calculated over three trials per query.

Method	AMC2023		Minerva Math		Olympiad-Bench		AIME2024		AIME2025	
	P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3
Llama-3.1-8B-Instruct	27.5	37.5	25.7	32.7	15.6	24.2	3.3	10.0	0.0	0.0
SFT-winning	27.5	35.0	24.2	35.6	16.0	27.1	6.6	6.6	0.0	0.0
Online DPO	22.5	33.1	25.3	30.5	15.1	26.2	3.3	13.3	0.0	0.0
PPO	25.0	35.0	21.7	34.9	15.7	26.2	3.3	**16.6**	**3.3**	0.0
DRO	22.5	35.0	23.1	33.8	15.5	25.6	3.3	0.0	0.0	**6.6**
OREO	27.5	32.5	25.7	35.3	15.7	26.8	3.3	6.6	0.0	**6.6**
RLOO	35.0	40.0	**26.1**	34.1	17.9	26.1	6.6	**16.6**	0.0	0.0
GRPO	35.0	37.5	25.3	35.3	**18.8**	25.3	6.6	**16.6**	**3.3**	0.0
Ours	**42.5**	**45.0**	**26.1**	**36.0**	17.3	**27.8**	**10.0**	**16.6**	**3.3**	3.3
Qwen2.5-Math-7B	52.5	70.0	27.0	36.0	37.7	46.2	23.0	26.6	6.6	13.3
SFT-winning	57.5	62.5	30.5	36.0	38.8	49.0	23.3	26.6	**13.3**	13.3
Online DPO	57.5	70.0	27.2	37.1	36.2	49.2	23.3	30.0	10.0	13.3
PPO	47.5	67.5	28.3	37.8	38.6	49.0	23.3	30.0	10.0	13.3
DRO	55.0	67.5	31.2	37.1	37.7	48.7	23.3	33.3	10.0	13.3
OREO	55.0	70.0	**31.6**	38.6	38.5	49.1	16.6	30.0	10.0	13.3
RLOO	57.5	72.5	30.1	40.4	38.8	48.8	23.3	36.6	**13.3**	16.6
GRPO	60.0	70.0	29.7	37.8	39.2	**49.4**	26.6	33.3	6.6	13.3
Ours	**62.5**	**75.0**	**31.6**	**41.9**	**39.8**	49.0	**30.0**	**40.0**	**13.3**	**23.3**

Figure 1. Training curves for our methods and baselines on the Llama-3.2-3B-Instruct model.



(a) Helpful assistant (b) TL;DR summarization (c) Prompt generation