GRPO with Binary Rewards Is an Adaptive Weighted Contrastive Loss

Youssef Mroueh[a]

[a]IBM Research

February 6, 2025

Abstract

The goal of this short note is to understand GRPO that was used successfully to train deepseek models.

Understanding GRPO with Rule Based (Binary) Reward

GRPO has been successfully in DeepSeek (v3,math, and R1) especially with rule based rewards, the goal of this note is to understand it. GRPO optimizes the following objective:

$$\max_{\theta} \mathbb{E}_q \mathbb{E}_{o\sim \pi_{\theta_{\text{old}}}(.|q)} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, A(o)\right) - \beta \text{KL}(\pi_\theta || \pi_{\text{ref}})$$

Where

$$A(o) = \frac{\left(r(o) - \mathbb{E}_{o'\sim \pi_{\theta_{\text{old}}}(.|q)} r(o')\right)}{\sqrt{Var_{o'\sim \pi_{\theta_{\text{old}}}(.|q)} r(o'))}} \tag{1}$$

and

$$f(x, y) = \min(xy, \text{clip}(x, 1-\varepsilon, 1+\varepsilon)y)$$

Note that in our context $x = \frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)} > 0$ and the advantage $A(o)$ can be positive or negative and hence if $A(o) > 0$ we have :

$$\begin{aligned} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, A(o)\right) &= A(o) \min\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, \text{clip}(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1-\varepsilon, 1+\varepsilon)\right) \\ &= A(o) \min(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1+\varepsilon) \end{aligned}$$

and if $A(o) < 0$

$$\begin{aligned} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, A(o)\right) &= A(o) \max\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, \text{clip}(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1-\varepsilon, 1+\varepsilon)\right) \\ &= A(o) \max(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1-\varepsilon) \end{aligned}$$

We will assume that we have a rule based reward that evaluates correctness of a reasoning or the execution of the code meaning that $r(o) \in \{0, 1\}$. We note :

$$p = p_{\theta_{\text{old}}}(.|q) = \mathbb{P}_{o \sim \pi_{\theta_{\text{old}}}}(r(o) = 1) \quad \text{probability of success}$$

Hence we have for mean and variance of a Bernoulli random variable :

$$\mathbb{E}_{o' \sim \pi_{\theta_{\text{old}}}(.|q)} r(o') = p \text{ and } Var_{o' \sim \pi_{\theta_{\text{old}}}(.|q)} r(o') = p(1-p)$$

and hence replacing mean and variance in the advantage function (1) :

$$A(o) = \begin{cases} \frac{1-p}{\sqrt{p(1-p)}} & \text{if } r(o) = 1, \\ -\frac{p}{\sqrt{p(1-p)}} & \text{if } r(o) = 0. \end{cases}$$

which simplifies to :

$$A(o) = \begin{cases} \sqrt{\frac{1-p}{p}} & \text{if } r(o) = 1, \\ -\sqrt{\frac{p}{(1-p)}} & \text{if } r(o) = 0. \end{cases}$$

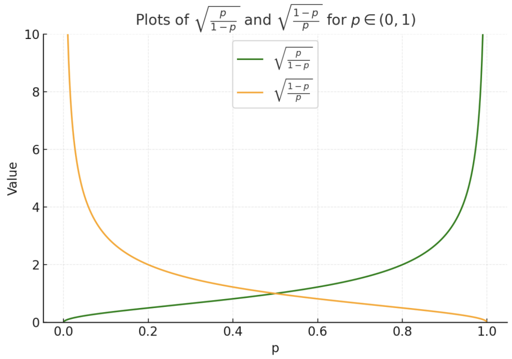

Figure 1: weighting of GRPO with success probability of old policy.

Hence we have conditioning on q:

$$\begin{aligned}
&\mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, A(o)\right) \\
&= \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, A(o)\right) 1_{r(o)=1} + \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, A(o)\right) 1_{r(o)=0} \\
&= \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, \sqrt{\frac{1-p}{p}}\right) 1_{r(o)=1} + \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} f\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, -\sqrt{\frac{p}{(1-p)}}\right) 1_{r(o)=0} \\
&= \sqrt{\frac{1-p}{p}} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q), r(o)=1} \min\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1 + \varepsilon\right) - \sqrt{\frac{p}{(1-p)}} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q), r(o)=0} \max\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1 - \varepsilon\right) \\
&= \sqrt{\frac{1-p}{p}} (1+\varepsilon) \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} 1_{r(o)=1, \pi_\theta(o|q) \geq (1+\varepsilon)\pi_{\theta_{\text{old}}}(o|q)} + \sqrt{\frac{1-p}{p}} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} \frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)} 1_{r(o)=1, \pi_\theta(o|q) < (1+\varepsilon)\pi_{\theta_{\text{old}}}(o|q)} \\
&- \sqrt{\frac{p}{(1-p)}} (1-\varepsilon) \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} 1_{r(o)=0, \pi_\theta(o|q) \leq (1-\varepsilon)\pi_{\theta_{\text{old}}}(o|q)} - \sqrt{\frac{p}{(1-p)}} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(.|q)} \frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)} 1_{r(o)=0, \pi_\theta(o|q) > (1-\varepsilon)\pi_{\theta_{\text{old}}}(o|q)}
\end{aligned}$$

and hence the overall cost is obtained by taking expectation over q, note that $p = p_{\theta_{\text{old}}}(q)$:

$$E_q \sqrt{\frac{1 - p_{\theta_{\text{old}}}(q)}{p_{\theta_{\text{old}}}(q)}} E_{o \sim \pi_{\theta_{\text{old}}}(.|q)} \min \left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1 + \varepsilon \right) 1_{r(o)=1}$$

$$-E_q \sqrt{\frac{p_{\theta_{\text{old}}}(q)}{(1 - p_{\theta_{\text{old}}}(q))}} E_{o \sim \pi_{\theta_{\text{old}}}(.|q)} \max \left(\frac{\pi_\theta(o|q)}{\pi_{\theta_{\text{old}}}(o|q)}, 1 - \varepsilon \right) 1_{r(o)=0}$$

$$-\beta \text{KL}(\pi_\theta || \pi_{\text{ref}})$$

We see that GRPO is effectively a weighted contrastive loss that is weighted by ratio depending on the probability of succes of $\pi_{\theta_{\text{old}}}(.|q)$: We see from the weights plots that :

- if the success probability of old policy is high (say $>$0.5), the weighting for points with success is low since the old policy is already good, and for failing point the weight is high and hence they are more penalized
- If the success probability of old policy is low (say $<$0.5), the weighting for points with success is high since we want to reinforce those successes, and for failing points these are still penalized but with a small weight

More observations due to clipping:

- for correct outputs the cost is constant $(1 + \varepsilon)$ if $\pi_\theta(o|q) \geq (1 + \varepsilon)\pi_{\theta_{\text{old}}}(o|q)$
- for wrong outputs the cost is $(1 - \varepsilon)$ if $\pi_\theta(o|q) \leq (1 - \varepsilon)\pi_{\theta_{\text{old}}}(o|q)$,

In summary, the standardized reward or the advantage function used in GRPO results in an interesting adaptive weighted contrastive loss : if the probability of success of the old policy is high, the wrong answers are more penalized than the correct ones are reinforced. If the probability of success of old policy is low , the correct answers are more reinforced than the wrong answers are penalized.