GRPO Algorithm

Traditional Reinforcement learning depends on absolute reward score, where the agent gets reward based on current performance of the task and the past history of agent is not taken into consideration. This can cause reward fluctuations and hence the model might not improve or produce better results. This occurs as the agent lacks a baseline model which shows the current state of the agent.

Hence, we need to introduce a Value function that would act as a baseline. We want to know how much better have we performed from the baseline model which would give us the advantage parameter.

$$A_t = r_t - V_\psi(s_t).$$

Where we need to optimize A. This also reduced the variance while training.

$$\mathcal{J}_{\text{adv}}(\theta) = \mathbb{E}\big[A(o)\big], \quad \text{where } A(o) = r(o) - V_\psi(o).$$

Now we need to make sure that we have not exponential overperformed and exponentially underperformed, hence we need to use the clip function.

Policy is basically the strategy the agent would use to perform the action to achieve the set of tasks in the environment.

$$r_t(\theta) = \frac{\pi_\theta(o_t \mid s_t)}{\pi_{\theta_{\text{old}}}(o_t \mid s_t)},$$

Represents the ratio between new policy and old policy. We clip this ratio between [1-e,1+e] .

Why not use raw probabilities (instead of ratios)?

- Fail to normalize updates across varying policy changes.
- Lack a mechanism to penalize drastic policy shifts, leading to instability

The advantage term A serves as directional guide and quantifies whether an action a in state s is better or worse than the average action the policy would take. (+ve, -ve)

$$\min\left(r_t(\theta)A_t,\ \text{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)A_t\right)$$

But we still need a mechanism, so that we don't deviate too much from the original policy, that's where KL divergence comes to play.

KL divergence basically tells us about how 2 probability distributions are similar to each other.

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)} \mathrm{d}x$$

In our case we would need something like

$$-\beta\, \mathbb{D}_{\mathrm{KL}}\left(\pi_\theta \parallel \pi_{\mathrm{ref}}\right)$$

Combining everything we get

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}\left[\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta\left(o_i\right)}{\pi_{\theta_{\text{old}}}\left(o_i\right)}A_i,\ \text{clip}\left(\frac{\pi_\theta\left(o_i\right)}{\pi_{\theta_{\text{old}}}\left(o_i\right)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\,\mathbb{D}_{KL}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\right)\right],$$

where

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}$$

Reference:

https://huggingface.co/blog/NormalUhr/grpo