GSPO

GRPO

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}\left[q \sim P(Q), \{o_i\}_{i-1}^{G} \sim \pi_{\theta_{old}}(O|q)\right]$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{old}(o_{i,t}|q,o_{i,<t})}\hat{A_{i,t}}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{old}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A_{i,t}}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}$$

GSPO

$$\mathcal{J}_{GSPO}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta old}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(s_i(\theta)\hat{A}_i, \text{clip}(s_i(\theta),1-\epsilon,1+\epsilon)\hat{A}_i\right)\right]$$
$$s_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{old}(y_i|x)}^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{old}(y_{i,t}|x,y_{i,<t})}\right)$$

token-level to seq-level

1. 高方差

2. 优化目标(token-level) 和 奖励(seq-level) 不匹配

GSPO在seq-level粒度计算重要性权重，和奖励粒度对齐

$$s_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{old}(y_i|x)}^{\frac{1}{|y_i|}}$$

分子\分母 表示给定序列x输出序列y的概率, 展开:

$$\pi(y_i|x) = \prod_{t=1}^{|y_i|}\pi(y_{i,t}|x,y_{i,<t})$$

长度归一化

$$s_i(\theta) = (\cdot)^{\frac{1}{|y_i|}}$$

clip粒度

GRPO: 考虑min操作, 当$A < 0$时, 重要性范围波动大, 噪声累计, 可能导致MoE模型崩溃.

GSPO: seq-level, 避免token裁剪带来的噪声和偏差.

消除Routing Replay依赖

Routing Replay

- MoE模型在梯度更新后, 专家激活模式会发生变化.

这导致token重要性权重波动, 触发clip使其梯度消失; 而保留梯度的token带有噪音影响训练.

- Routing Replay强制**当前策略**路由和**旧策略**相同的专家.

GSPO: seq-level, 对单个token波动不敏感.