RLHF objective:

non-differentiable

$$\max_{\pi} \mathop{E}_{x,y} \left[\underbrace{r(x,y)}_{\text{frozen}} - \beta D_{KL} \left(\pi(y|x) \,||\, \pi_{ref}(y|x) \right) \right]$$

current aligned LLM

SFT (instruction-tuned LLM)

(x, y)
data used for SFT is different than that used for RLHF,
x
but come from same distribution

- why do we need RL?

DPO (direct preference optimization):

→ no explicit reward model

→ not going to sample outputs $y|x$ from the model
↳ "rollouts"

→ "preference tuning"

$$\max_{\pi} \mathop{E}_{x,y} \left[r(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{ref}(y|x)} \right]$$

$$= \min_{\pi} \mathop{E}_{x,y} \left[\log \frac{\pi(y|x)}{\pi_{ref}(y|x)} - \frac{1}{\beta} r(x,y) \right]$$

---

let's introduce a new policy π^* that incorporates the reward term as well as π_{ref}

$$\pi^*(y|x) = \frac{\pi_{ref}(y|x)\exp\left(\frac{1}{\beta} r(x,y)\right)}{\boxed{\sum_{y} \pi_{ref}(y|x)\exp\left(\frac{1}{\beta} r(x,y)\right)}}$$

$Z(x)$, normalizer / partition function

Substitute $Z(x)$ into our objective:

$$\min_{\pi} \mathop{E}_{x,y} \log \left[\frac{\pi(y|x)}{\boxed{\frac{1}{Z(x)} \pi_{ref}(y|x)\exp\left(\frac{1}{\beta} r(x,y)\right)}} - \log Z \right]$$

$$= \min_{\pi} \mathop{E}_{x,y} \log \left[\boxed{\frac{\pi(y|x)}{\pi^*(y|x)}} \right] - \log Z$$

↳ KL div

$$= \min_{\pi} \mathop{E}_{x} D_{KL}\left(\pi(y|x) \,||\, \pi^*(y|x)\right) - \log Z$$

KL div. is minimized at 0 when $\pi(y|x) = \pi^*(y|x)$

$$\underset{\substack{\uparrow \\ \text{optimal policy}}}{\pi(y|x)} = \pi^*(y|x) = \frac{\pi_{ref}(y|x)\exp\left(\frac{1}{\beta}\,\underline{r(x,y)}\right)}{Z(x)}$$

Solve the above for $r(x,y)$

$$\underline{r(x,y)} = \boxed{\beta \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} + \beta \log \underbrace{Z}}$$

Bradley-Terry pref model:

$$p(y_w > y_L \mid x) = \frac{\exp(r(x,y_w))}{\exp(r(x,y_w)) + \exp(r(x,y_L))}$$

Substitute reward function:

$$p(y_w > y_L \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_L|x)}{\pi_{ref}(y_L|x)} - \beta \log \frac{\pi^*(y_w|x)}{\pi_{ref}(y_w|x)}\right)}$$

Convert to loss fn (neg. log likelihood)

$$L_{DPO}(\pi_\theta, \pi_{ref}) = -\underset{x, y_w, y_L}{E} \log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_L|x)}{\pi_{ref}(y_L|x)}\right)$$

$\downarrow$ aligned model we are training nice properties of DPO:

- no explicit reward model
- no need for rollouts from the policy