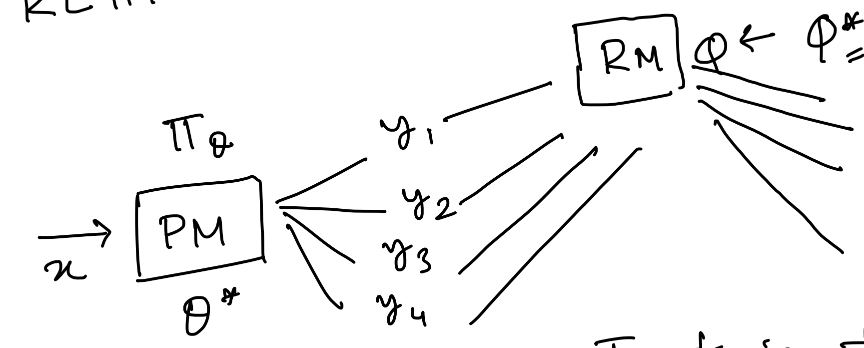
Alignment of LM (DPO: Direct Preference optimization)

RLHF



(x, y_+, y_-)

$$\left[\sum_{(x, y_+, y_-) \in D} \log \sigma\left(r_\varphi(x, y_+) - r_\varphi(x, y_-) \right) \right] \text{ O.F. for the RM}$$

maximize $\downarrow \varphi^*$

To train the policy

$$\theta^* \leftarrow \arg\max \mathbb{E}_{\pi_\theta(y|x)} \left[\underbrace{r_{\varphi^*}(x, y) - \beta \, \mathrm{KL}\left(\pi_\theta(y|x) \,\|\, \pi_{ref}(y|x) \right)}_{\text{Regularized Reward}} \right]$$

$\boxed{\text{RM}}$

$\boxed{\text{PM}}$

$\boxed{\text{Ref.}}$

DPO

- No RM (Explicit)
- No RL
- Policy model will act as a RM

$$r^* \leftarrow \arg\max_{r} \sum_{(x, y_+, y_-) \in D} \log \sigma \left(r(x, y_+) - r(x, y_-) \right)$$

optimal Policy model

$$\Rightarrow \pi^* \leftarrow \arg\max_{\pi} \mathbb{E}_{\pi(y|x)} \; r^*(x, y) - \beta \cdot \mathrm{KL}\left(\pi^*(y|x) \,\|\, \pi_{ref}(y|x) \right) -$$

$$\text{s.t.} \sum_{y \in Y} \pi^*(y|x) = 1$$

$$\nabla_{\pi(y_0|x)} \mathcal{L}(\pi, \lambda) = 0$$

$$\mathcal{L}(\pi, \lambda) = \sum_{y \in Y} \pi^*(y|x) \, r^*(x, y) - \sum_{y \in Y} \pi^*(y|x) \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} + \lambda \left(\sum_{y \in Y} \pi^*(y|x) - 1 \right)$$

$$\nabla_{\pi(y_0|x)} \mathcal{L} = r^*(x, y) - \left[1 + \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} \right] + \lambda = 0$$

$$\Rightarrow r^*(x, y) + \lambda - 1 = \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)}$$

$$\Rightarrow r^*(x, y) + \bar{\lambda} = \log \pi^* / \pi_{ref} \Longleftarrow (1)$$

where $\bar{\lambda} = \lambda - 1$

$$\Rightarrow e^{(r^*(x, y) + \bar{\lambda})} = \pi^* / \pi_{ref}$$

$$\Rightarrow \underline{\pi^*(y|x)} = \pi_{ref} \cdot \exp\left(r^*(x,y) + \bar{\lambda}\right) \quad \text{— } (1')$$

This is the optimal policy in term of optimal reward.

$$\sum_{y \in Y} \pi^*(y|x) = 1 \quad \Rightarrow \sum \pi_{ref}(y|x) \exp\left(r^*(x,y) + \bar{\lambda}\right) = 1$$

$$\Rightarrow \underline{\exp(\bar{\lambda})} = \frac{1}{\sum \pi_{ref}(y|x) \exp\left(r^*(x,y)\right)} = \frac{1}{Z}$$

For (1), $\boxed{\pi^*(y|x) = \pi_{ref} \frac{\exp\left(r^*(x,y)\right)}{Z}}$ — (1) policy model in term of RM

we will write the reward in term of the policy.

From Eq (11) $\quad r^*(x,y) + \bar{\lambda} = \log \pi^* / \pi_{ref} \qquad \bar{\lambda} = -\log Z \simeq \log \frac{1}{Z}$

$$\Rightarrow r^*(x,y) = \log \pi^* / \pi_{ref} - \bar{\lambda}$$

$$\rightarrow \boxed{r^*(x,y) = \log \frac{\pi^*}{\pi_{ref}} - \log Z} \quad \text{reward in term of policy}$$

the parametric policy & reward

$$r_\theta(x,y) = \beta . \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} - \log Z_x(\theta) \quad \text{— } (3)$$

Training the reward fnc:

$$(x, y+, y-) \quad \arg\max \sum_{(x,y+,y-) \in D} \log \sigma\left(r_\theta(x,y+) - r_\theta(x,y-)\right)$$

From Eq (3)

$$r_\theta(x,y+) = \beta \log \frac{\pi_\theta(y+|x)}{\pi_{ref}(y+|x)} - \log Z_x(\theta)$$

$$r_\theta(x,y-) = \beta \log \frac{\pi_\theta(y-|x)}{\pi_{ref}(y-|x)} - \log Z_x(\theta)$$

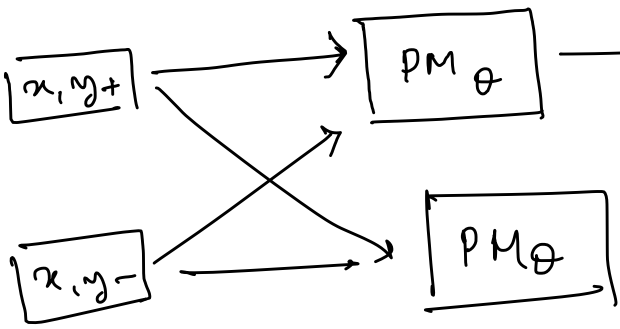
$$\log \sigma\left(\left[\beta \log \frac{\pi_\theta(y+|x)}{\pi_{ref}(y+|x)} - \cancel{\log Z_x(\theta)}\right] - \left[\beta \log \frac{\pi_\theta(y-|x)}{\pi_{ref}(y-|x)} - \cancel{\log Z_x(\theta)}\right]\right)$$

$$= \log \sigma\left(\beta\left[\log \frac{\pi_\theta(y+|x)}{\pi_{ref}(y+|x)} - \log \frac{\pi_\theta(y-|x)}{\pi_{ref}(y-|x)}\right]\right)$$

$$=$$

$$= \log \frac{\exp\left(\beta \log \frac{\pi_\theta(y_+|x)}{\pi_{ref}(y_+|x)}\right)}{\exp\left(\beta \log \frac{\pi_\theta(y_+|x)}{\pi_{ref}(y_+|x)}\right) + \exp\left(\beta \log \frac{\pi_\theta(y_-|x)}{\pi_{ref}(y_-|x)}\right)}$$

DPO Objective



$$r_+ = \log \frac{\pi_\theta(y_+|x)}{\pi_{ref}(y_+|x)}$$

$$r_- = \log \frac{\pi_\theta(y_-|x)}{\pi_{ref}(y_+|x)}$$

$$\max \log \sigma(r_+ - r_-)$$

$\uparrow \dfrac{\pi_\theta(y_+|x) \approx .9}{\pi_{ref}(y_+|x) \approx 0.8}$ $\quad y_+ \rightarrow \quad y_+ \approx .1$

$$\boxed{\frac{.9}{.8} \approx \frac{.1}{.1}}$$

$\beta \rightarrow (0.3 - 0.003)$

DPO is prone to generating a biased-policy that favours out-of-distribution responses