DPO

- **核心思想：**直接使用偏好数据进行策略优化，省去 `reward` 模型策略优化。
- **技术背景知识：**

  首先给定prompt x，生成两个答案$(y_1, y_2)\ \Pi^{SFT}(y|x)$ ，并通过人工标注对比y_1, y_2 ，获得偏好结果(preference) $y_w \succ y_l|x$，其中w和l表示`win`和`lose`。

  引入奖励模型r ，$y_1 > y_2$ 的概率可以表示为

$$p(y_1 > y_2) = \frac{r^*(x,y_1)}{r^*(x,y_1)+r^*(x,y_2)}$$

  为使得奖励函数均为正数，引入`Bradley-Terry` 模型。

  - **Bradley-Terry**：

$$p^*(y_w \succ y_l|x) = \frac{exp(r^*(x,y_1))}{exp(r^*(x,y_1))+exp(r^*(x,y_2))}$$

    交叉熵：

    令$a_x = exp(r^*(x,y_1)),\ \ a_y = exp(r^*(x,y_2))$

$$\begin{aligned} Loss &= -E_{(a_x,a_y)\sim D}[ln\frac{a_x}{a_x+a_y}] \\ &= -E_{(x,y_w,y_l)\sim D}[ln\frac{exp(r^*(x,y_w))}{exp(r^*(x,y_w))+exp(r^*(x,y_l))}] \\ &= -E_{(x,y_w,y_l)\sim D}[ln\frac{1}{1+exp(r^*(x,y_l)-r^*(x,y_w))}] \\ &= -E_{(x,y_w,y_l)\sim D}[ln\sigma(r^*(x,y_w)-r^*(x,y_l))] \end{aligned}$$

  - **KL 散度：**

$$KL(P||Q) = \sum_{x\in X} P(X)log(\frac{P(X)}{Q(X)})$$

    $P(x), Q(x)$ 分别是数据真实分布和模型预测分布。

- `DPO` **目标函数**:获取更多的奖励，并尽可能保证与基准模型一致。

$$\begin{aligned} &\max_\pi E_{x\in X,y\in\pi}[r(x,y)] - \beta\cdot\mathbb{D}_{KL}[\pi(y|x)||\pi_{ref}(y|x)] \\ &= \max_\pi E_{x\in X,y\in\pi}[r(x,y)] - E_{x\in X,y\in\pi}[\beta\cdot log\frac{\pi(y|x)}{\pi_{ref}(y|x)}] \\ &= \max_\pi E_{x\in X,y\in\pi}[r(x,y) - \beta\cdot log\frac{\pi(y|x)}{\pi_{ref}(y|x)}] \\ &= \max_\pi E_{x\in X,y\in\pi}[log\frac{\pi(y|x)}{\pi_{ref}(y|x)} - \frac{1}{\beta}r(x,y))] \\ &= \min_\pi E_{x\in X,y\in\pi}[log\frac{\pi(y|x)}{\pi_{ref}(y|x)} - log\ \ exp(\frac{1}{\beta}r(x,y))] \\ &= \min_\pi E_{x\in X,y\in\pi}[log\frac{\pi(y|x)}{\pi_{ref}(y|x)\cdot exp(\frac{1}{\beta}r(x,y))}] \\ &= \min_\pi E_{x\in X,y\in\pi}[log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{ref}(y|x)\cdot exp(\frac{1}{\beta}r(x,y))} - log\ \ Z(x)] \end{aligned}$$

  令$Z(x)$ 表示如下：

$$Z(x) = \sum_y \pi_{ref}(y|x)exp(\frac{1}{\beta}r(x,y))$$

令：

$$\frac{1}{Z(x)}\pi_{ref}(y|x)\cdot exp(\frac{1}{\beta}r(x,y)) = \frac{\pi_{ref}(y|x)\cdot exp(\frac{1}{\beta}r(x,y))}{\sum_y \pi_{ref}(y|x)exp(\frac{1}{\beta}r(x,y))}$$
$$= \pi^*(y|x)$$

接下来继续对`dpo`目标函数进行化简：

$$\min_\pi E_{x\in X, y\in\pi}[log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{ref}(y|x)\cdot exp(\frac{1}{\beta}r(x,y))} - log\ \ Z(x)]$$
$$= \min_\pi E_{x\in X, y\in\pi}[log\frac{\pi(y|x)}{\pi^*(y|x)} - log\ \ Z(x)]$$

由于$Z(x)$表达式与π 不相关，优化可以直接省去。

$$\min_\pi E_{x\in X, y\in\pi}[log\frac{\pi(y|x)}{\pi^*(y|x)} - log\ \ Z(x)]$$
$$= \min_\pi E_{x\in X, y\in\pi}[log\frac{\pi(y|x)}{\pi^*(y|x)}]$$
$$= \min_\pi E_{x\sim D}[\mathbb{D}_{KL}(\pi(y|x)||\pi^*(y|x))]$$

当 目标函数最小化，也就是$\mathbb{D}_{KL}$最小化，所满足的条件为：

$$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{ref}(y|x)\cdot exp(\frac{1}{\beta}r(x,y))$$

反解奖励函数$r(x,y)$

$$r(x,y) = \beta\frac{\pi(y|x)}{\pi_{ref}(y|x)} + \beta\cdot ln\mathbb{Z}(x)$$

求解奖励函数隐式表达后，带入`Bradley-Terry`交叉熵函数：

$$Loss = -E_{(x,y_w,y_l)\sim D}[ln\sigma(r^*(x,y_w) - r^*(x,y_l))]$$
$$= -E_{(x,y_w,y_l)\sim D}[ln\sigma(\beta log\frac{\pi(y_w|x)}{\pi_{ref}(y_w|x)} - \beta log\frac{\pi(y_l|x)}{\pi_{ref}(y_l|x)})]$$

到此，整个数学部分已推导完毕，不得不说句牛逼plus。

- **梯度表征：**

  将上述损失进行梯度求导

$$\nabla_\theta Loss(\pi_\theta;\pi_{ref}) = -E_{(x,y_w,y_l)\sim D}[\beta\sigma(\beta log\frac{\pi(y_w|x)}{\pi_{ref}(y_w|x)} - \beta log\frac{\pi(y_l|x)}{\pi_{ref}(y_l|x)})[\nabla_\theta log\pi(y_w|x) - \nabla_\theta log\pi(y_l|x)]]$$

  再令：

$$\hat{r}(x,y) = \beta\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$$

  最终形式：

$$\nabla_\theta Loss(\pi_\theta;\pi_{ref}) = -\beta E_{(x,y_w,y_l)\sim D}[\underbrace{\sigma(\hat{r}^*(x,y_w) - \hat{r}^*(x,y_l))}_{higher\ weight\ when\ reward\ estimate\ is\ wrong}[\underbrace{\nabla_\theta log\pi(y_w|x)}_{increase\ likelihood\ of\ y_w} - \underbrace{\nabla_\theta log\pi(y_l|x)}_{decrease\ likelihood\ of\ y_l}]]$$

- **改进方法ODPO**

  `dpo`缺陷主要是：采用`Bradley-Terry model`只给出了一个`response`比另一个`response`好的概率，而没有告诉我们好的程度。

`odpo` 核心思想： 把这个好的程度的差距信息引入到偏好的建模里，应该能带来收益，及在`dpo`损失里添加`margin`，这相当于要求`偏好回应`的评估分数要比`非偏好回应`的评估分数大，且要大`offset`值这么多。目的是：加大对靠得比较近的数据对的惩罚力度。

$$Loss^{odpo} = -E_{(x,y_w,y_l)\sim D}[ln\sigma(r^*(x,y_w) - r^*(x,y_l)) - \delta_r]$$
$$\delta_r = \alpha\ log(r(y_w) - r(y_l))$$

- **相似改进方法：**

  `IPO` `KTO` 都是不需要奖励模型的；