Implicit Reward as the Bridge A Unified View of SFT and DPO Connections

Yang

2025 年 7 月 21 日

1 Introduction

模型训练后为了培养其在某一方面比较突出的技能通常会使用 SFT 和 RFT(DPO) 进行学习，但是两者之间的关系是什么尚未研究清楚。本论文认为 SFT 和 DPO 都是通过操控 implicit reward 来获得成长，DPO(preference learning) 在之前的研究中就已经阐明了作用。

首先对所有可能的 reward function 的 optimal policy 构建 policy-reward subspace，SFT 和 DPO 都在这个子空间中分布。关键发现是 SFT 的 KL 散度是一个 0 阶量，即在微分之后对模型的提升没有任何限制且与 π 相关。

论文观点梳理：

- 在数学上证明了 SFT 同样是操控 implicit reward function 达成训练效果，和 DPO 有相似的结构，可以统一辨别 SFT 和 DPO 的理论观点
- 在 SFT 阶段降低学习率来减轻 KL 项缺失造成的影响。KL 项确保策略不会过度偏离基础模型从而促进稳定高效的学习
- 选用了一些其他的以 f-divergence function 的 SFT 进行训练，再进行 DPO，得到的效果更优
- 将 LLM logits 和 Q-function 之间的关系从 DPO 拓展到 SFT

2 Unified View of SFT and DPO

2.1 Markov Decision Process

定义 state: $s \in S$，action: $a \in A$，前者表示 context，后者表示可能的下一个 token，$P(a|s)$ 表示决策的已知概率，$T(s'|s,a)$ 已知，由此实现 token 层面的决策和状态转移。模型具备 $r(s,a)$ 作为决策的 reward，γ 是 discount factor。

定义一个 policy π 的 occupancy measure 表示为：

$$\rho(s) = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i P(s_i = s|\pi)$$

其中 $P(s_i = s|\pi)$ 表示基于 policy π，第 i 步的状态为 s 的概率

定义一个 policy π 的 state-action distribution：

$$\mu(s,a) = \pi(a|s)\rho(s)$$

2.2 Distribution Matching in Post-Training

首先阐述 imitation learning 中两个重要的概念：expert 指高质量策略，模型模仿的标杆，但是 expert 的策略是黑盒，难以显示表示。比如 Alphago 先模仿人类棋谱 (expert policy)，再强化学习可以变得更好；policy 可以抽象为概率分布，就是 $\pi(a|s)$ 的集合体

Post-Training 的关键目的就是为了降低 μ_{Expert} 和 μ_π 之间的 f-divergence，但是只用 entropy 可能会破坏 base model 的 natural language priors，因此正则化的标准改为：

$$min_\pi D_f(\mu_\pi||\mu_{Expert}) + \beta D_{KL}(\pi||\pi_{ref})$$

从两个角度来看：

- $D_{KL} = -H(\pi) + H(\pi, \pi_{ref})$，后者不会很大

2.3 Important Theorem

上述 post-training 的目的可以等价为首先基于任意的奖励函数学习最优策略，然后最优化奖励函数：

$$-min_r[E_{\mu_E}[f^*(-r)] + max_\pi E_{\mu_\pi}[r] - \beta D_{KL}(\pi||\pi_{ref})]$$

其中 f^* 是基于选择的 f-divergence 产生的 convex conjugate function(凸共轭函数)，r 是 reward function。这个 formulation 建立了找到合适的 policy 和选择 reward function 之间的关系。后半部分 max 的部分具备 closed-form solution，令

$$J(\pi) = E_{\mu_\pi}[r] - \beta D_{KL}(\pi||\pi_{ref})$$

解表示为 $\pi^* = argmax_\pi J(\pi), \quad J(\pi^*) = V^*(s_0)$