Teng Xiao, Yige Yuan, Mingxiao Li,
Zhengyu Chen, Vasant G Honavar

On a Connection Between Imitation Learning and RLHF

RLHF is Barely RL, and Secretly Performs Imitation Learning

paper code

Insight: Directly conducting imitation learning

Without Bradley-Terry assumption

$$\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x}) \xrightarrow[\text{reward modelling}]{} \pi_\phi(\mathbf{y} \mid \mathbf{x}) \xrightarrow[\text{policy optimization}]{} \pi_\theta(\mathbf{y} \mid \mathbf{x})$$

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{DIL}}(\boldsymbol{\theta}) = \mathrm{KL}(\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x}))$$

Framework: Unlock family of Alignment Algorithms

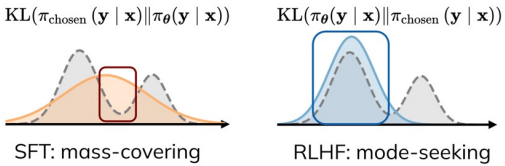
$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})} \left[\log \frac{\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})} - \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})} \right] = \mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}[\log r(\mathbf{x}, \mathbf{y})] - \mathrm{KL}(\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x}))$$

The reward is about the density ratio

Table 1: Summary of the variants of DIL with different h-functions for Bregman divergence: $\mathcal{L}_{\text{DIL}}(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\text{chosen}}(\mathbf{y}|\mathbf{x})}[\ell_1(f_{\boldsymbol{\theta}})] + \mathbb{E}_{\pi_{\text{rejected}}(\mathbf{y}|\mathbf{x})}[\ell_{-1}(f_{\boldsymbol{\theta}})]$ as a function of log ratio $f_{\boldsymbol{\theta}} = \log(\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})/\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x}))$.

h-Bregman Density Ratio Estimation	h-function	$\ell_1(f_\theta)$	$\ell_{-1}(f_\theta)$
LSIF (Kanamori et al., 2009)	$h(r) = (r-1)^2/2$	$-e^{f_\theta}$	$\frac{1}{2}e^{2f_\theta}$
BCE (Hastie et al., 2009)	$h(r) = r\log r - (r+1)\log(r+1)$	$\log(1+e^{-f_\theta})$	$\log(1+e^{f_\theta})$
UKL (Nguyen et al., 2010)	$h(r) = r\log r - r$	$-f_\theta$	e^{f_θ}

SFT vs RLHF: Different Directions



Key Finding: RLHF performs imitation learning, not RL!

Proposition 1. *Suppose the chosen response distribution $\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})$, the EBM $\pi_\phi(\mathbf{y} \mid \mathbf{x})$, and the model $\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$. KL-regularized RLHF with $\beta = 1$ can be viewed as the following problem:*

$$\min_{\boldsymbol{\theta}} \mathrm{KL}(\pi_{\boldsymbol{\theta}} \| \pi_\phi^*) \quad \text{s.t.} \quad \pi_\phi^* = \arg\min_{\pi_\phi} \mathrm{KL}(\pi_{\text{chosen}} \| \pi_\phi),$$

where $\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x}) = \pi_\phi(\mathbf{y} \mid \mathbf{x}) = \pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$ is the equilibrium.

$$\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x}) \xrightarrow[\text{reward modelling}]{} \pi_\phi(\mathbf{y} \mid \mathbf{x}) \xrightarrow[\text{policy optimization}]{} \pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$$

Result: DIL Outperforms SOTAs across Benchmarks

	Model (↓) / Benchmark (→)	MMLU-PRO	BBH	MUSR	MATH	GSM8K	ARC
LLama3-8B-Base	SFT	31.00	46.16	41.27	3.70	46.32	60.15
	DPO (Rafailov et al., 2024)	31.58	47.80	40.48	4.53	38.67	64.42
	SLiC (Zhao et al., 2023)	31.11	46.53	40.55	3.92	48.82	61.43
	f-DPO (Wang et al., 2024a)	30.85	47.55	40.39	4.37	39.55	62.85
	IPO (Azar et al., 2024)	30.18	46.78	39.58	4.02	22.67	62.88
	KTO (Ethayarajh et al., 2024)	31.16	47.92	40.24	4.13	38.99	63.17
	CPO (Xu et al., 2024b)	30.95	47.17	41.59	4.25	46.93	61.69
	SimPO (Meng et al., 2024)	31.61	48.38	40.08	4.23	31.54	65.19
	DIL w/ LSIF	**32.22**	**48.78**	**42.75**	**4.68**	**48.98**	**65.37**

Dataset (→)	TL;DR Summarization			Anthropic-HH		
Method (↓) / Metric (→)	vs SFT	vs Chosen	Average	vs SFT	vs Chosen	Average
DPO (Rafailov et al., 2024)	71.22	57.58	64.40	69.32	59.35	64.34
SLiC (Zhao et al., 2023)	68.61	55.72	62.17	65.52	57.71	61.62
f-DPO (Wang et al., 2024a)	66.19	51.37	58.78	60.21	52.38	56.30
IPO (Azar et al., 2024)	72.17	56.51	64.34	63.19	55.12	59.16
CPO (Xu et al., 2024b)	73.13	58.89	66.01	72.30	63.39	67.86
SimPO (Meng et al., 2024)	69.71	54.38	62.05	67.85	57.51	62.68
DIL w/ LSIF	**75.47**	**60.25**	**67.86**	**73.32**	**65.02**	**69.17**