Online Preference Alignment for Language Models via Count-based Exploration

Chenjia Bai [1,6] **Yang Zhang** [2,1] **Shuang Qiu** [3] **Qiaosheng Zhang** [4] **Kang Xu** [5] **Xuelong Li** [1*]

1. The Institute of Artificial Intelligence (TeleAI), China Telecom 2. Tsinghua University 3. City University of Hong Kong 4. Shanghai Artificial Intelligence Laboratory
5. Tencent AI Lab 6. Shenzhen Research Institute of Northwestern Polytechnical University.

Motivation

- Existing RLHF methods rely on **a fixed dataset**, which can be limited in **data coverage**, and the resulting reward model is hard to **generalize in out-of-distribution (OOD) responses**.
- Offline RLHF lacks systematic exploration, leading to the learned reward model being inaccurate for OOD prompt-response pairs.
- The central problem in an online RLHF process is how to explore the prompt-response space in each iteration.

Contributions

- We propose "Count-based Online Preference Optimization (COPO)", integrating count-based exploration with DPO for online RLHF and encouraging the LLMs to balance exploration and preference optimization.
- We construct a lightweight pseudo-counting module with several fully-connected layers based on the LLM, which is theoretically grounded in policy optimization of online RLHF.
- we conduct RLHF experiments of COPO and several strong online RLHF baselines on Zephyr and Llama-3 models. The results of instruction-following and academic benchmarks show better performance.

Theoretical Analysis

Assumption. Linear reward $r_\theta(x,y) = \theta^T\varphi(x,y)$ with bounded features ($\|\varphi(x,y)\|_2 \le 1$) and parameter $\theta \epsilon \Theta_B$

Confidence Set. For any $\lambda > 0$, letting $\gamma = \frac{1}{(2+e^{-B}+e^{B})}$,with probability at least $1-\delta$, we have:

$$\|\hat{\theta}_{\text{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}_t}+\lambda I} \le C \cdot \sqrt{\frac{d+\log(\frac{1}{\delta})}{\gamma^2 n} + \lambda B^2}$$

UCB Objective. We thus construct the optimistic expected value function $\hat{J}_\beta(\pi)$which takes the upper confidence bound (UCB) as the reward estimate, as

$$\hat{J}_\beta(\pi) = J_\beta(\pi) + \xi \, \|\mathbb{E}_{x\sim\rho}[\phi(x,\pi(x))]\|_{\left(\Sigma_{\mathcal{D}_t}+\lambda I\right)^{-1}}$$

where $\xi = C\sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}$

Regret Bound. Under linear rewards and Assumption, the total regret over T iterations is:

$$\text{Regret}(T) \le \sqrt{T}\cdot C_1 \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2} + \lambda B^2} \cdot \sqrt{dt},$$

Methodology

- **Iterative Data Collection.** For each iteration t, sample prompts x from $\widetilde{D}_t$, generate responses $y \sim \pi_{ref}(\cdot\,|x)$, and rank $\{y, y_w, y_l\}$using a reward model to form D_t with best and worst pair.
- **Pseudo-Count Estimation (CFN).** Train a Coin-Flipping Network f_ϑ via equation below to estimate pseudo-counts $\hat{N}(x,y) \approx d/\|f_\vartheta(s_i)\|^2$, where s_i is the LLM's hidden state for (x,y) and $c_i \sim\{-1,1\}^d$ are random labels.

$$\min_{f_\vartheta} J_{\text{cfn}}(f_\vartheta;\mathcal{D}_{\text{cfn}}) = \mathbb{E}_{(s_i,s_i^{\text{label}})\sim\mathcal{D}_{\text{cfn}}}[\mathcal{L}(s_i,s_i^{\text{label}})] = \arg\min_{\vartheta} \sum_{i=1}^{|\mathcal{D}_{\text{cfn}}|} \|\mathbf{c_i} - f_\vartheta(s_i)\|^2,$$

- **Policy Optimization.** Update the LLM policy π_φ by maximizing the COPO objective function, combining the DPO loss with an exploration bonus, then set $\pi_{ref} \leftarrow \pi_\varphi$for the next iteration

$$\max_{\pi_\varphi} J_{\text{copo}}(\pi_\varphi,\mathcal{D}_t) = -\mathcal{L}_{\text{DPO}}(\pi_\varphi;\mathcal{D}_t,\beta) + \alpha \underbrace{\mathbb{E}_{x\sim\mathcal{D}_t,y\sim\pi_\varphi(y|x)}\left[1/\left(\sqrt{N_{\mathcal{D}_t}(x,y;\vartheta)+\lambda}\right)\right]}_{\text{optimistic term of COPO}},$$

- **Optimism in the face of uncertainty.** The count-based bonus encourages active exploration toward not only high-reward but also more uncertain regions with respect to the regions the LLM has already confirmed.

Experiments

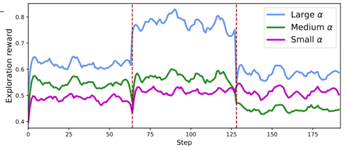
Results on AlpacaEval 2.0 and MT-Bench

Model	AlpacaEval 2.0 LC Win Rate	AlpacaEval 2.0 Win Rate	AlpacaEval 2.0 Avg. len	MT-Bench Avgerage	MT-Bench 1st Turn	MT-Bench 2nd Turn
Zephyr-7B-SFT	8.01	4.63	916	5.30	5.63	4.97
Zephyr-7B-DPO	15.41	14.44	1752	7.31	7.55	7.07
DPO Iter 1 (Zephyr)	20.53	16.69	1598	7.53	7.81	7.25
DPO Iter 2 (Zephyr)	22.12	19.82	1717	7.55	7.85	7.24
DPO Iter 3 (Zephyr)	22.19 (↑14.18)	19.88 (↑15.25)	1717	7.46 (↑2.16)	7.85	7.06
SELM Iter 1 (Zephyr)	20.52	17.23	1624	7.53	7.74	7.31
SELM Iter 2 (Zephyr)	21.84	18.78	1665	7.61	7.85	7.38
SELM Iter 3 (Zephyr)	24.25 (↑16.24)	21.05 (↑16.42)	1694	7.61 (↑2.31)	7.74	7.49
COPO Iter 1 (Zephyr)	26.43	21.61	1633	7.68	7.72	7.64
COPO Iter 2 (Zephyr)	**27.21** (↑19.20)	22.61	1655	7.78	7.85	7.71
COPO Iter 3 (Zephyr)	26.91	**23.60** (↑18.97)	1739	**7.79** (↑2.49)	7.89	7.69
Llama-3-8B-Instruct	22.92	22.57	1899	7.93	8.47	7.38
DPO Iter 1 (Llama3-It)	30.89	31.60	1979	8.07	8.44	7.70
DPO Iter 2 (Llama3-It)	33.91	32.95	1939	7.99	8.39	7.60
DPO Iter 3 (Llama3-It)	33.17 (↑10.25)	32.18 (↑9.61)	1930	8.18 (↑0.25)	8.60	7.77
SELM Iter 1 (Llama3-It)	31.09	30.90	1956	8.09	8.57	7.61
SELM Iter 2 (Llama3-It)	33.53	32.61	1919	8.18	8.69	7.66
SELM Iter 3 (Llama3-It)	34.67 (↑11.75)	**34.78** (↑12.21)	1948	8.25 (↑0.32)	8.53	7.98
COPO Iter 1 (Llama3-It)	33.68	33.15	1959	8.12	8.38	7.86
COPO Iter 2 (Llama3-It)	34.30	33.31	1939	8.25	8.49	8.01
COPO Iter 3 (Llama3-It)	**35.54** (↑12.62)	32.94 (↑10.37)	1930	**8.32** (↑0.39)	8.53	8.11
SPIN	7.23	6.54	1426	6.54	6.94	6.14
Orca-2.5-SFT	10.76	6.99	1174	6.88	7.72	6.02
DNO (Orca-2.5-SFT)	22.59	24.97	2228	7.48	7.62	7.35
Mistral-7B-Instruct-v0.2	19.39	15.75	1565	7.51	7.78	7.25
SPPO (Mistral-it)	28.53	31.02	2163	7.59	7.84	7.34
Yi-34B-Chat	27.19	21.23	2123	7.90	-	-
Llama-3-70B-Instruct	33.17	33.18	1919	9.01	9.21	8.80
GPT-4 Turbo (04/09)	55.02	46.12	1802	9.19	9.38	9.00

Results on academic benchmarks

Models	GSM8K (8-s CoT)	HellaSwag (10-s)	ARC (25-s)	TruthfulQA (0-s)	EQ (0-s)	OBQA (10-s)	Average
Zephyr-7B-SFT	43.8	82.2	57.4	43.6	39.1	35.4	50.3
Zephyr-7B-DPO	47.2	84.5	61.9	45.5	65.2	38.0	57.0
DPO Iter 1 (Zephyr)	45.5	85.2	62.1	52.4	68.4	39.0	58.8
DPO Iter 2 (Zephyr)	44.9	85.4	62.0	53.1	69.3	39.4	59.0
DPO Iter 3 (Zephyr)	43.2	85.2	60.8	52.5	69.1	39.6	58.4
SELM Iter 1 (Zephyr)	46.3	84.8	62.9	52.9	68.8	39.6	59.2
SELM Iter 2 (Zephyr)	46.2	85.4	62.1	53.1	69.3	39.6	59.3
SELM Iter 3 (Zephyr)	43.8	85.4	61.9	52.4	69.9	39.8	58.9
COPO Iter 1 (Zephyr)	46.8	85.0	62.4	53.0	68.7	39.3	59.2
COPO Iter 2 (Zephyr)	46.7	85.3	62.5	53.3	69.1	39.8	59.5
COPO Iter 3 (Zephyr)	47.0	85.4	62.9	53.4	69.9	40.3	59.9
Llama-3-8B-Instruct	76.7	78.6	60.8	51.7	61.8	38.0	61.3
DPO Iter 1 (Llama3-It)	78.5	81.7	63.9	55.5	64.1	42.6	64.4
DPO Iter 2 (Llama3-It)	79.4	81.7	64.4	56.4	64.3	42.6	64.8
DPO Iter 3 (Llama3-It)	80.1	81.7	64.1	56.5	64.1	42.6	64.8
SELM Iter 1 (Llama3-It)	78.7	81.7	64.5	55.4	64.1	42.4	64.5
SELM Iter 2 (Llama3-It)	79.3	81.8	64.7	56.5	64.2	42.6	64.9
SELM Iter 3 (Llama3-It)	80.1	81.8	64.3	56.5	64.2	42.8	65.0
COPO Iter 1 (Llama3-It)	79.1	81.7	64.3	56.4	64.3	43.0	64.8
COPO Iter 2 (Llama3-It)	79.3	81.8	64.6	56.4	64.4	43.2	65.0
COPO Iter 3 (Llama3-It)	80.2	81.8	64.7	56.5	64.4	43.6	65.2
SPIN	44.7	85.9	65.9	55.6	54.4	39.6	57.7
Mistral-7B-Instruct-v0.2	43.4	85.3	63.4	67.5	65.9	41.2	61.1
SPPO (Mistral-it)	42.4	85.6	65.4	70.7	56.5	40.0	60.1

Different α



Different Exploration Factor

Factor	Iter1	Iter2	Iter3
0.01	32.89	33.18	33.76
0.10	33.68	34.30	35.54
0.50	33.70	34.61	34.82