Multi-Step Preference Optimization via Two-Player Markov Games

Yongtao Wu*[1], Luca Viano*[1], Yihang Chen[2], Zhenyu Zhu[1], Quanquan Gu†[1], and Volkan Cevher†[2]

[1]EPFL [2]UCLA

Introduction

Our contributions:

- We formulate multi-step preference optimization as a two-player partially observable Markov game. Unlike [1, 2, 3] who focus on the preference feedback at the final state, we assume that the preference signal is received at each step. Such feedback allows the model to better identify which steps are correct or erroneous, potentially enhancing learning efficiency and accuracy.
- We propose Multi-step Preference Optimization (`MPO`) based on the natural actor-critic framework and Optimistic Multi-step Preference Optimization (`OMPO`), built upon the optimistic online gradient descent. Theoretically, we show that `OMPO` requires $\mathcal{O}(\epsilon^{-1})$ policy updates to converge to an ϵ-approximate Nash equilibrium, compared to $\mathcal{O}(\epsilon^{-2})$ by the algorithms provided in [1, 2, 3]. Our result cannot be trivially extended by [4] due to the partially observable nature of Markov game. We bypass this difficulty by parameterizing the game over occupancy measures.
- We provide practical implementations of both `MPO` and `OMPO` for LLM alignment. Numerical results show that the proposed methods achieve considerable improvement on MT-bench-101, compared to the multi-step variant of DPO.

Multi-step alignment as two-player Markov games

We can cast the multi-step alignment process as a finite-horizon MDP.

- We define $s_h = [x_1, a_1, \ldots, x_{h-1}, a_{h-1}, x_h]$ as the state at $h > 1$.
- We define the action a_h as the answer given s_h.
- Particularly, we have $s_1 = x_1$. The prompt in the next state is sampled under the transition $x_{h+1} \sim f(\cdot|s_h, a_h)$, which is equivalent to $s_{h+1} \sim f(\cdot|s_h, a_h)$. The terminal state is s_{H+1}.
- We define the pair-wise reward function of two state-action pairs as the preference of two trajectories: $r(s_h, a_h, s'_h, a'_h) = \mathbb{P}([s_h, a_h] \succ [s'_h, a'_h])$.
- We define the initial state distribution ν_1 is a distribution over the initial prompt x_1. Note that each state in $\mathcal{S}$ is a pair of s_h and s'_h generated by two policies.

The multi-step setting covers a number of alignment problems:

Example 1 (Single-step alignment)**.** A language model receives one prompt and outputs one answer. Our framework covers the single-step alignment by dissecting the answer into single tokens.

Example 2 (Chain-of-thought reasoning alignment)**.** The horizon H denotes the reasoning step, where x_1 is the initial prompt and $x_2, \ldots, x_{H+1}$ are empty. Each a_h corresponds to a reasoning step.

Example 3 (Mutli-turn conversation alignment)**.** The horizon H denotes the total turn of conversation. In the h-th turn, x_h is the prompt, and a_h is the answer.

Our goal is to identify the Nash equilibrium of the following two-player Markov game:

$$(\pi^*, \pi^*) = \arg\max_{\pi} \min_{\pi'} \mathbb{E}_{s_1 \sim \nu_1, s_h, a_h, s'_h, a'_h} \left[\sum_{h=1}^{H} r(s_h, a_h, s'_h, a'_h) \right],$$

where $s_1 = s'_1 = x_1, a_h \sim \pi(\cdot|s_h)$, $a'_h \sim \pi'(\cdot|s'_h)$, $s_h \sim f(\cdot|s_{h-1}, a_{h-1})$, $s'_h \sim f(\cdot|s'_{h-1}, a'_{h-1})$.

* Equal contribution † Equal mentorship

Method

Algorithm 1 `MPO` (Theory Version)

input: reference policy π^1, preference oracle $\mathbb{P}$, learning rate $\beta = \sqrt{\frac{\log \underline{\pi}^{-1}}{TH^2}}$, total iteration T
for $t = 1, 2, \ldots, T$ **do**

$$\pi_h^{t+1}(a|s) \propto \pi_h^t(a|s) \exp\left[\beta \mathbb{E}_{s',a' \sim d_h^{\pi^t}|_{s_1(s)}} Q_h^{\pi^t, \pi^t}(s, a, s', a')\right] \quad \forall h \in [H], \quad \forall s, a.$$

end for
output: $\bar{\pi}^T$ (such that $d_h^{\bar{\pi}^T} = \frac{1}{T} \sum_{t=1}^{T} d_h^{\pi^t}$, $\forall h \in [H]$.).

Algorithm 2 `OMPO` (Theory Version)

input: occupancy measure of reference policy π^1 denoted as d^1, preference oracle $\mathbb{P}$ (i.e. reward function r), learning rate β, Bregman divergence $\mathbb{D}$, iteration T
for $t = 1, 2, \ldots, T$ **do**

$$d_h^{t+1} = \arg\max_{d \in \mathcal{F}_{s_1}} \beta \left\langle d, 2\mathbb{E}_{s',a' \sim d_h^t} r(\cdot, \cdot, s', a') - \mathbb{E}_{s',a' \sim d_h^{t-1}} r(\cdot, \cdot, s', a') \right\rangle - \mathbb{D}(d, d_h^t) \quad \forall h \in [H] \ \forall s_1.$$

end for
$\pi_h^{\text{out}}(a|s) = \frac{\bar{d}_h(s,a|s_1)}{\sum_a \bar{d}_h(s,a|s_1)}$ with $\bar{d}_h = T^{-1} \sum_{t=1}^{T} d_h^t$ for all $h \in [H]$ for the unique s_1 from which s is reachable.
Output : π^{out}

Theorem 4. *Consider Alg.1 and assume that the reference policy is uniformly lower bounded by $\underline{\pi}$, then there exists a policy $\bar{\pi}^T$ such that $d_h^{\bar{\pi}^T} = \frac{1}{T} \sum_{t=1}^{T} d_h^{\pi^t}, \forall h \in [H]$, and it holds that for $T = \frac{16H^4 \log \underline{\pi}^{-1}}{\epsilon^2}$ the policy pair $(\bar{\pi}^T, \bar{\pi}^T)$ is an ϵ-approximate Nash equilibrium. Therefore, Alg.1 outputs an ϵ-approximate Nash equilibrium after $\frac{16H^4 \log \underline{\pi}^{-1}}{\epsilon^2}$ policy updates.*

Theorem 5 (Convergence of `OMPO`)**.** *Consider Alg.2 and assume the occupancy measure of the reference policy is uniformly lower bounded by $\underline{d}$. Moreover, let $\mathbb{D}$ be $1/\lambda$ strongly convex, i.e. $\mathbb{D}(p||q) \geq \frac{\|p-q\|_1^2}{2\lambda}$. Then, setting $T = \frac{10H \log \underline{d}^{-1}}{\beta\epsilon}$ and $\beta \leq \frac{1}{\sqrt{2\lambda}}$, we ensure that the output of Alg.2 is an ϵ-approximate Nash equilibrium. Therefore, we need at most $\frac{10H \log \underline{d}^{-1}}{\beta\epsilon}$ policy updates.*

Experiments

Table 1: Evaluation results on MT-bench-101 dataset. We can observe that both of the proposed algorithms `MPO` and `OMPO` considerably outperform the baseline in terms of the score.

Model	Avg.	Perceptivity: Memory CM	Perceptivity: Understanding SI	Perceptivity: Understanding AR	Perceptivity: Interference TS	Perceptivity: Interference CC	Adaptability: Rephrasing CR	Adaptability: Rephrasing FR	Adaptability: Reflection SC	Adaptability: Reflection SA	Adaptability: Reasoning MR	Adaptability: Reasoning GR	Interactivity: Questioning IC	Interactivity: Questioning PI
Base (Mistral-7B-Instruct)	6.223	7.202	7.141	7.477	7.839	8.294	6.526	6.480	4.123	4.836	4.455	5.061	5.818	5.641
DPO (iter=1)	6.361	7.889	6.483	7.699	8.149	8.973	7.098	7.423	3.448	**6.123**	3.421	4.492	5.639	5.858
DPO (iter=2)	6.327	7.611	6.206	8.106	8.052	9.111	6.670	7.153	3.494	5.884	3.360	4.691	5.837	6.078
DPO (iter=3)	5.391	6.019	4.521	6.890	6.631	8.177	5.437	5.723	3.448	5.295	3.142	4.015	5.256	5.529
SPPO (iter=1)	6.475	7.432	7.464	7.714	8.353	8.580	6.917	6.714	4.136	5.055	4.403	5.400	6.036	5.966
SPPO (iter=2)	6.541	7.516	7.496	7.808	8.313	8.731	7.077	6.867	4.136	5.281	4.488	5.477	6.098	5.751
SPPO (iter=3)	6.577	7.575	7.547	7.944	8.365	8.797	7.040	6.865	4.442	5.185	4.346	5.394	6.092	5.906
Step-DPO (iter=1)	6.433	7.463	7.054	7.790	8.157	8.593	6.827	6.748	4.234	4.849	4.236	5.519	5.982	6.171
Step-DPO (iter=2)	6.553	7.616	7.043	7.925	8.147	8.662	6.790	6.878	4.331	5.048	4.366	**5.734**	**6.391**	6.254
Step-DPO (iter=3)	6.442	7.665	7.023	7.767	8.016	8.589	6.723	6.581	4.305	5.014	4.153	5.453	6.202	**6.257**
`MPO`⋆ (iter=1)	6.630	7.624	**7.846**	8.085	8.398	8.947	7.105	7.286	4.208	4.993	4.377	5.264	6.179	5.873
`MPO`⋆ (iter=2)	6.735	7.838	7.723	8.196	**8.590**	9.027	7.347	7.209	4.240	5.137	4.469	5.531	6.181	6.061
`MPO`⋆ (iter=3)	6.733	**7.868**	7.686	**8.289**	8.510	9.078	7.330	7.529	**4.461**	4.829	4.225	5.366	6.198	6.155
`OMPO`⋆ (iter=2)	6.736	7.733	7.723	8.257	8.478	9.122	7.300	7.421	4.123	5.288	**4.506**	5.513	6.179	5.923
`OMPO`⋆ (iter=3)	**6.776**	7.649	7.792	8.281	8.578	**9.136**	**7.424**	**7.635**	4.377	5.308	4.312	5.455	6.187	5.954

References

[1] Wang et al. "Is rlhf more difficult than standard rl? a theoretical perspective". In: *NeurIPS*. 2023.
[2] Swamy et al. "A Minimaximalist Approach to Reinforcement Learning from Human Feedback". In: *ICML*. 2024.
[3] Shani et al. "Multi-turn Reinforcement Learning from Preference Human Feedback". In: *NeurIPS*. 2024.
[4] Alacaoglu et al. "A natural actor-critic framework for zero-sum Markov games". In: *ICML*. 2022.

Acknowledgement

HASLERSTIFTUNG
FNSNF Swiss National Science Foundation
erc European Research Council Established by the European Commission
ZEISS