Note on Theoretical Understanding of Learning from Human Preferences

Tianbing Xu

December 30, 2023

1 Introduction

Reinforcement Learning from human preferences (**RLHF**[3]) relies on two approximations or assumptions:

- Pairwise preference is approximated as a pointwise Elo score, represented as a sigmoid function of the relative reward using the Bradley-Terry model.
- The reward model, trained on human preference data, can generalize to out-of-distribution data sampled by the policy.

DPO ([2]) eliminates the need for the second approximation by directly learning a policy without a reward modeling stage. However, DPO is susceptible to over-fitting due to the unbounded and unreliable nature of relative rewards and potential model drifting from the reference policy when KL-regularization is weak.

Ψ Policy Optimization (ΨPO, [1]) offers a theoretical understanding of learning from human preferences by bypassing both approximations in RLHF. Ψ Policy Optimization is a versatile learning paradigm, with DPO and RLHF being specific cases of ΨPO when pairwise preferences are modeled using Bradley-Terry modeling. This is achieved through the adoption of a log-odd functional mapping for preferences. Additionally, Identity Policy Optimization is proposed as another special case of Ψ Policy Optimization, specifically tailored for the identity mapping to human preferences.

2 Notations and Context

The Language Generation Process is conceptualized as a Context Bandit problem. For a given context $x \in \mathcal{X}$, an action $y \in \mathcal{Y}$ is generated using a policy network (or language model) π. Employing a behavior policy μ, two actions $y, y' \sim \mu(x)$ are independently generated given the context x. Human preferences are then elicited by having individuals rate these two actions, denoting a preference for y if $I(y, y'|x) = 1$.

The probability of preference is defined as:

$$p^*(y > y'|x) = \mathbf{E}\left[I(y, y'|x)\right] \tag{1}$$

The expected preference over the distribution of the behavior policy μ is given by:

$$p^*(y > \mu) = \mathbf{E}_{y' \sim \mu}\left[p^*(y > y'|x)\right]$$

The total preference of policy π over behavior policy μ is expressed as:

$$p^*(\pi > \mu) = \mathbf{E}_{x \sim \rho, y \sim \pi}\left[p^*(y > \mu|x)\right]$$

Note: In the subsequent formulas, the variable x is ignored for simplicity without loss of generalization.

3 Ψ Policy Optimization

Ψ-Preference Optimization (ΨPO) is defined by the following objective function:

$$\max_{\pi} \mathbb{E}_{y\sim\pi, y'\sim\mu} \left[\Psi(p^*(y > y'))\right] - \tau D_{KL}(\pi||\pi_0) \tag{2}$$

Here, Ψ represents a non-decreasing function $[0,1] \to \mathbb{R}$, and π_0 is the reference policy. The objective aims to strike a balance between optimizing a non-linear function of preference probability and maintaining proximity to the reference policy, with the additional inclusion of the Kullback-Leibler (KL) regularization term. The parameter τ controls the strength of the KL regularization.

3.1 The Unification of RLHF and DPO

Here, we establish that RLHF and DPO are special cases of ΨPO when employing a log-odd mapping of the reward function through Bradley-Terry modeling of human preferences.

The Bradley-Terry preference probability is defined as follows:

$$p(y > y') = \sigma(r(y) - r(y')) = \frac{\exp(r(y))}{\exp(r(y)) + \exp(r(y'))} \tag{3}$$

This assumes that, with a large sample of human references, the Bradley-Terry preference probability converges to the true human preference probability:

$$p(y > y') \to p^*(y > y')$$

Let Ψ be the log-odd function, the inverse mapping of the sigmoid function. We obtain the KL-constrained RL objective function in RLHF ([3]):

$$\max_{\pi} \mathbb{E}_{\pi} \left[r(y)\right] - \tau D_{KL}(\pi||\pi_0) \tag{4}$$

this can be derived from (Eq.2, 3 and 11),

$$\max_{\pi} \mathbb{E}_{y'\sim\mu} \left[\Psi(p^*(y > y'))\right] = \mathbb{E}_{y'\sim\mu} \left[r(y) - r(y')\right]$$

By employing relative reward (Eq.10) without learning a preference model, we arrive at DPO ([2]):

$$\min_{\pi} \mathbb{E}_{(y,y')\sim\mathcal{D}} \left[-\log\sigma\left(\tau\log\frac{\pi(y)}{\pi(y')} - \tau\log\frac{\pi_0(y)}{\pi_0(y')}\right)\right] \tag{5}$$

3.2 Identity Policy Optimization (IPO)

Let Ψ be the identity mapping, and using Eq. (2), we obtain the direct regularized optimization of total preference:

$$\max_{\pi} p^*(\pi, \mu) - \tau D_{KL}(\pi||\pi_0) \tag{6}$$

In contrast to DPO, the reward is bounded within the range of $[0,1]$.

This can be further derived into a computationally efficient method to optimize the loss function $L(\pi)$:

$$\min_{\pi} L(\pi) = \mathbb{E}_{y\sim\pi, y'\sim\mu} \left[\left(h_\pi(y,y') - \frac{p^*(y > \mu) - p^*(y' > \mu)}{\tau}\right)\right] \tag{7}$$

Here, $h_\pi(y,y')$ represents the measure of closeness to the reference policy of the relative preference and is derived from the relative reward with a closed from (Eq. (10)):

$$h_\pi(y,y') = \log\frac{\pi(y)}{\pi(y')} - \log\frac{\pi_0(y)}{\pi_0(y')}$$

3.3 Sampled Loss for IPO

With an unbiased estimator of $\mathcal{L}(\pi)$, we obtain the Population IPO loss:

$$\min_{\pi} \mathbb{E}_{y,y' \sim \mu} \left[\left(h_\pi(y, y') - \frac{I(y, y')}{\tau} \right)^2 \right] \tag{8}$$

Given samples $I(y, y') = (y_{w,i}, y_{l,i}, 1)$ from offline Data $\mathcal{D}$, the goal is to minimize the loss function:

$$\min_{\pi} \mathbb{E}_{(y_w, y_l) \sim \mathcal{D}} \left[\left(h_\pi(y, y') - \frac{1}{2\tau} \right)^2 \right] \tag{9}$$

With strong regularization τ, the learned policy π is expected to closely align with the reference policy π_0, helping mitigate over-fitting.

4 Appendix

For the Bradley-Terry model, the equality is given by:

$$p(y > y') = \sigma(r(y) - r(y')) = \frac{\exp(r(y))}{\exp(r(y)) + \exp(r(y'))}$$

The closed-form expression for the optimal policy with KL-constrained RL (Eq 4) is:

$$\pi(y) = \frac{1}{Z(x)} \pi_0(y) \exp(\frac{1}{\tau} r(y))$$

This yields the corresponding reward:

$$r(y) = \tau \log \frac{\pi(y)}{\pi_0(y)} + \tau \log Z(x)$$

Notably, the relative reward between two responses, as defined in Eq (10), is given by:

$$\delta_r(y, y') = r(y) - r(y') = \tau \left(\log \frac{\pi(y)}{\pi_0(y)} - \log \frac{\pi(y)}{\pi_0(y')} \right) = \tau \left(\log \frac{\pi(y)}{\pi(y')} - \log \frac{\pi_0(y)}{\pi_0(y')} \right) \tag{10}$$

The sigmoid function mapping from $\mathbb{R}$ to $[0, 1]$ is defined as:

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

The logit (log-odds) function mapping from $[0, 1]$ to $\mathbb{R}$ is defined as:

$$\phi(p) = \log \frac{p}{1 - p}$$

The identities include: $\phi(p) = \sigma^{-1}(p)$
And, consequently:

$$\phi(\sigma(x)) = x \tag{11}$$

References

[1] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, Rémi Munos *A General Theoretical Paradigm to Understand Learning from Human Preferences, 2023*

[2] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn, *Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023*

[3] L. Ouyang, et. al OpenAI, *Training language models to follow instructions with human feedback, NIPS 2022*