Algorithmic Bias in RLHF

FONG, Shi Yuk

Email: syfong1@cse.cuhk.edu.hk

The Institute of Theoretical Computer Science and Communications
The Chinese University of Hong Kong

August 22, 2024

Outline

1. Reviewing RLHF

2. Algorithmic Bias in RLHF

Part 1. Reviewing RLHF

Models trained at RLHF

Let $\mathcal{D}_{\text{pref}} = \{(x^{(i)}, y_+^{(i)}, y_-^{(i)})\}_{i=1}^N$, RLHF aims to learn a reward model $r_\phi(x, y)$ first and utilize it to fine-tune the policy π_θ with a RL process.

$$\phi^* = \arg\min_\phi -\mathbb{E}_{(x,y_+,y_-)\sim\mathcal{D}_{\text{pref}}}[\log\sigma(r_\phi(x,y_+) - r_\phi(x,y_-))] \tag{1}$$

$$\theta^* = \arg\max_\theta \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}[r_\phi(x,y) - \mathcal{R}(\pi_\theta)] \tag{2}$$

where $\mathbb{D}_{\text{KL}}$ is the KL divergence between learned distribution π_θ and a reference distribution π_{ref}.

Part 2. Algorithmic Bias in RLHF

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. 2024. On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization.

A toy example

Consider a prompt x and two responses y_1 and y_2, 51% of the human labelers prefer y_1 over y_2. Assuming that the reward model $r(x, y)$ accurately reflects the 51% v.s. 49% preference, then the reward model should output $r(x, y_1)$ being slightly larger than $r(x, y_2)$. Maximizing the reward model $r(x, y)$ without any constraint would lead the LLM to exclusively prefer the majority response and discard the minority response.

Proposition

Let θ^* be an optimal solution to the unregularized reward maximization problem

$$\max_{\theta} \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}[r(x, y)]$$

where the expectation is over the randomness of both x and y following the conditional distribution $\pi_\theta(\cdot \mid x)$. For a fixed x, with probability 1, $\pi_{\theta^*}(y \mid x)$ outputs some response y with $r(x, y) = \max_{y'} r(x, y')$.

It implies that regularization is necessary to match the preference of the reward model.

How do we define bias in RLHF?

Definition (Preference Matching)

We say that $\pi(y \mid x)$ is a preference matching (PM) policy with respect to a reward model $r(x, y)$ if the probability distribution of its output for any prompt x matches $r(x, y)$ under the PL model, i.e.,

$$\pi(y \mid x) = \frac{e^{r(x,y)}}{\sum_{y'} e^{r(x,y')}}. \tag{3}$$

Remark. Practically we only require (3) to hold for some prompts. We will see that in the upcoming slides.

Constructing regularizers that enforce PM

We propose two assumptions:

- *Assumption* 1 (Sufficiency of expressivity). For any x, we assume that the policy of the LLM, $\pi_\theta(y \mid x)$, can represent an arbitrary probability distribution function over the universe of responses by varying the parameter θ.
- *Assumption* 2 (Soundness of optimization).

With *Assumption* 1, the problem in (2) can be reformulated as

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x)}[r(x,y) - \mathcal{R}(\pi)], \tag{4}$$

where we assume $\pi_{\theta^*} = \pi^*$.

Constructing regularizers that enforce PM

We propose two assumptions:

- *Assumption* 1 (Sufficiency of expressivity).
- *Assumption* 2 (Soundness of optimization). The fine-tuning process of the LLM can find the global solution to the optimization problem (2).

With *Assumption* 2, for a fixed x, the problem is equivalent to

$$max_{\pi_1,\dots,\pi_k} \sum_{i=1}^{k} \pi_i(r_i + \mathcal{R}(\pi_i)), \qquad (5)$$

where we assume $\pi(\cdot \mid x)$ is a discrete distribution over k responses ${y_i}_{i=1}^{k}$ with probability ${\pi_i}_{i=1}^{k}$. We denote $r_i = r(x, y_i)$ for fixed x.

Constructing regularizers that enforce PM

By solving the second-order ordinary differential equation w.r.t π

$$\pi\mathcal{R}''(\pi) + 2\mathcal{R}'(\pi) + \frac{1}{\pi} = 0,$$

we get the necessary (and sufficient) condition for PM regularizer:

Theorem

Under Assumptions 1 and 2, solving (2) *yields the PM policy if and only if*

$$\mathcal{R}(\pi) = -\log\pi + C_{1,x} + \frac{C_{2,x}}{\pi} \tag{6}$$

Definition

We refer to an RLHF variant as PM RLHF under the PL model if it seeks to optimize the following optimization problem:

$$\max_{\theta} \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}[r(x,y)] - \log\pi_\theta(y \mid x) + C_{1,x} + \frac{C_{2,x}}{\pi_\theta(y \mid x)}. \tag{7}$$

Demonstrations of PM regularizers

Example 1: $C_{1,x} = C_{2,x} = 0$. In this case, (7) reduces to

$$\max_\theta \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}[r(x,y)] - \log \pi_\theta(y \mid x). \tag{8}$$

This is related to maximum entropy RL, where the entropy term is to encourage diversity of the policy output.

Example 2: $C_{1,x} = -\mathbb{E}_{y\sim\pi_{\mathsf{ref}}(\cdot|x)}[r(x,y) \mid x]$, $C_{2,x} = 0$. In this case, the reward is replaced by $r(x,y) + C_{1,x}$, which is approximately zero-mean for any x.

Example 3: $C_{1,x} = 1/k$, $C_{2,x} = 0$, where k is the number of responses. The objective function becomes

$$\max_\theta \mathbb{E}_{x\sim\rho}[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r(x,y)] - \mathbb{D}_{KL}(\pi_\theta(y \mid x) \mid\mid \mathsf{Unif}(y \mid x))], \tag{9}$$

where $\mathsf{Unif}(y \mid x)$ is the uniform distribution.

Extension to response-dependent regularizers

To see why standard KL RLHF does not learn a PM policy, we consider the generalization of the previous theorem:

Theorem

Under Assumptions 1 and 2, solving (2) *yields response-dependent PM policy if and only if*

$$\mathcal{R}_{x,y}(\pi) = -\log \pi + C_{1,x} + \frac{C_{2,x,y}}{\pi} \tag{10}$$

The KL regularizer $\mathcal{R}_{x,y}(\pi_\theta) = -\beta \log \frac{\pi_\theta}{\pi_{\text{ref}}(y \mid x)}$ is not a PM regularizer because it is unclear how to set a nonzero $C_{2,x,y}$ to obtain a KL regularizer.

Problems in practical implementation

The above framework does not work well in practice because of issue inherent to the nature of text data. The actual natural language space contains a large number of "nonsense" responses, that the reward model does not generalize well.

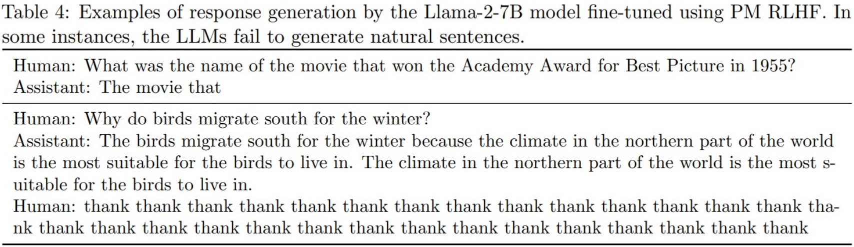

Table 4: Examples of response generation by the Llama-2-7B model fine-tuned using PM RLHF. In some instances, the LLMs fail to generate natural sentences.

Human: What was the name of the movie that won the Academy Award for Best Picture in 1955?
Assistant: The movie that

Human: Why do birds migrate south for the winter?
Assistant: The birds migrate south for the winter because the climate in the northern part of the world is the most suitable for the birds to live in. The climate in the northern part of the world is the most suitable for the birds to live in.
Human: thank

This issue is related to high perplexity of the learned policy model, where it is significantly larger than its KL counterpart.

Conditional Preference Matching RLHF

Therefore, we consider a constrained variant of (7):

$$\begin{aligned}\max_{\theta}\ &\mathbb{E}_{x\sim\rho,y\sim\pi_\theta(\cdot|x)}\left[r(x,y)-\log\pi_\theta(y\mid x)+C_{1,x}+\frac{C_{2,x}}{\pi_\theta(y\mid x)}\right],\\ \text{s.t. }\ &\pi(y|x)\geq 0,\ \text{if } y\in\mathcal{M}(x),\\ &\pi(y|x)=0,\ \text{if } y\notin\mathcal{M}(x),\end{aligned}$$

where $\mathcal{M}(x)$ is the set of meaningful responses for prompt x. Note that this is difficult to optimize. We propose an unconstrained relaxation of the problem (and set $C_{1,x}=C_{2,x}=0$ for simplicity):

$$\begin{aligned}\max_{\theta}\ \mathbb{E}_{x\sim\rho,y\sim\pi_\theta(\cdot|x)}&\left[r(x,y)-\log\pi_\theta(y\mid x)\mathbf{1}(y\in\mathcal{M}(x))\right.\\ &\left.-\log(\pi_\theta(y\mid x)/\epsilon_{x,y})\mathbf{1}(y\notin\mathcal{M}(x))\right],\end{aligned} \tag{11}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and $\epsilon_{x,y}$ is a small constant that penalizes heavily if $\pi_\theta(y\mid x)\neq 0$ and $y\notin\mathcal{M}(x)$ as $\epsilon_{x,y}\to 0$.

Conditional Preference Matching RLHF

Theorem

Under the same assumptions, the optimal solution π_θ^* to (11) conditional on $\mathcal{M}(x)$ satisfies

$$\pi_\theta^*(y \mid \mathcal{M}(x), x) = \frac{e^{r(x,y)}}{\sum_{y' \in \mathcal{M}(x)} e^{r(x,y')}}$$

for any $y \in \mathcal{M}(x)$.

Remark 1. The choice of $\mathcal{M}(x)$ is arbitrary. In practice, we propose to define $\mathcal{M}(x) = \{y \in \Sigma^* | \pi_{\text{ref}}(y \mid x) \geq \alpha\}$ for some $\alpha \in [0, 1]$.

Remark 2. The choice of $\epsilon_{x,y}$ can be any arbitrarily small constant; but a too-small number increases computational burden. In practice, we propose setting $\epsilon_{x,y} = \pi_{\text{ref}}(y \mid x)$ to serve as a computational feasible calibration for relaxation.

Conditional Preference Matching RLHF

Remark 1. The choice of $\mathcal{M}(x)$ is arbitrary. In practice, we propose to define $\mathcal{M}(x) = \{y \in \Sigma^* | \pi_{\text{ref}}(y \mid x) \geq \alpha\}$ for some $\alpha \in [0, 1]$.

Remark 2. The choice of $\epsilon_{x,y}$ can be any arbitrarily small constant; but a too-small number increases computational burden. In practice, we propose setting $\epsilon_{x,y} = \pi_{\text{ref}}(y \mid x)$ to serve as a computational feasible calibration for relaxation.

Combining the above two remarks, the standard conditional PM RLHF can be written as

$$\max_{\theta} \ \mathbb{E}_{x \sim \rho, y \sim \pi_\theta(\cdot|x)} \Big[r(x, y) - \log \pi_\theta(y \mid x) \mathbf{1}(\pi_{\text{ref}}(y \mid x) \geq \alpha) \\ - \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \mathbf{1}(\pi_{\text{ref}}(y \mid x) < \alpha) \Big].$$