

794 A Limitations and Societal Impacts

795 Our discussion highlights an ineffective regularization with direct alignment algorithms used widely
 796 to align to human preferences. In this work we analyze and resolve these issues. However, We still
 797 assume an underlying Bradley-Terry model of human preferences, as these models might not be
 798 accurate in explaining the ways humans give preferences and do not experiment with larger models
 799 due to limited computational resources. Our work is to advance alignment algorithms that avoid
 800 over-optimization and ensure the development of models that are safe for real-world deployment.

801 B Proofs and Derivations

802 B.1 IS-DPO is an unbiased estimation of online DPO

803 **Theorem 1 Restated.** Assuming $\text{Supp}(\pi_{\text{ref}}) = \text{Supp}(\pi_{\theta})$, then objective (Eqn. (11)) is an unbiased
 804 estimation of the objective in Eqn. (10) and the gradient concerning π_{θ} of the weighted squared loss
 805 equals to the gradient of KL divergence. Proof. Online-DPO objective can be expressed as:

$$\begin{aligned} \mathcal{L}_{\text{Online-DPO}}(\pi_{\theta}, \pi_{\text{ref}}) = \\ - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, (\mathbf{y}^w, \mathbf{y}^l) \sim \pi_{\theta}(\cdot | \mathbf{x})} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w | \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^l | \mathbf{x})} \right) \right] \end{aligned}$$

806 Expanding the expectation leads:

$$\begin{aligned} &= - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\sum_{\mathbf{y}^w, \mathbf{y}^l} \pi_{\theta}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x}) \log \sigma \left(\beta \log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w | \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^l | \mathbf{x})} \right) \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\sum_{\mathbf{y}^w, \mathbf{y}^l} \pi_{\text{ref}}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x}) \frac{\pi_{\theta}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x})} \log \sigma \left(\beta \log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w | \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^l | \mathbf{x})} \right) \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\mathbb{E}_{(\mathbf{y}^w, \mathbf{y}^l) \sim \pi_{\text{ref}}(\cdot | \mathbf{x})} \left[\frac{\pi_{\theta}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x})} \log \sigma \left(\beta \log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w | \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^l | \mathbf{x})} \right) \right] \right] \end{aligned}$$

807 Where we denote $\pi(\mathbf{y}^w, \mathbf{y}^l) = \pi(\mathbf{y}^w | \mathbf{x})\pi(\mathbf{y}^l | \mathbf{x})$. This yields Eqn. (11).

808 Similarly, given a prompt \mathbf{x} , consider the gradient of KL divergence:

$$\nabla_{\theta} \text{KL}(\pi_{\theta} || \pi_{\text{ref}}) = \sum_{\mathbf{y}} \nabla_{\theta} \pi_{\theta}(\mathbf{y} | \mathbf{x}) + \sum_{\mathbf{y}} \log \left(\frac{\pi_{\theta}(\mathbf{y} | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} | \mathbf{x})} \right) \nabla_{\theta} \pi_{\theta}(\mathbf{y} | \mathbf{x})$$

809 We can drop the first term since $\sum_{\mathbf{y}} \nabla_{\theta} \pi_{\theta}(\mathbf{y} | \mathbf{x}) = \nabla_{\theta} \left(\sum_{\mathbf{y}} \pi_{\theta}(\mathbf{y} | \mathbf{x}) \right) = \nabla_{\theta}(1) = 0$. We now consider
 810 the gradient of weighted-squared loss with Importance Sampling:

$$\begin{aligned} &\frac{1}{2} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, (\mathbf{y}^w, \mathbf{y}^l) \sim \pi_{\text{ref}}(\cdot | \mathbf{x})} \left[\frac{\pi_{\theta}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w, \mathbf{y}^l | \mathbf{x})} \nabla_{\theta} \left(\log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w | \mathbf{x})} - \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^l | \mathbf{x})} \right)^2 \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, (\mathbf{y}^w, \mathbf{y}^l) \sim \pi_{\theta}(\cdot | \mathbf{x})} \left[\left(\log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w | \mathbf{x})} - \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^l | \mathbf{x})} \right) (\nabla_{\theta} \log \pi_{\theta}(\mathbf{y}^w | \mathbf{x}) - \nabla_{\theta} \log \pi_{\theta}(\mathbf{y}^l | \mathbf{x})) \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, (\mathbf{y}^w, \mathbf{y}^l) \sim \pi_{\theta}(\cdot | \mathbf{x})} \left[\log \frac{\pi_{\theta}(\mathbf{y}^w | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^w | \mathbf{x})} \nabla_{\theta} \log \pi_{\theta}(\mathbf{y}^w | \mathbf{x}) + \log \frac{\pi_{\theta}(\mathbf{y}^l | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^l | \mathbf{x})} \nabla_{\theta} \log \pi_{\theta}(\mathbf{y}^l | \mathbf{x}) \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, \mathbf{y} \sim \pi_{\theta}(\cdot | \mathbf{x})} \left[\log \frac{\pi_{\theta}(\mathbf{y} | \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} | \mathbf{x})} \nabla_{\theta} \log \pi_{\theta}(\mathbf{y} | \mathbf{x}) \right] = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\nabla_{\theta} \text{KL}(\pi_{\theta} || \pi_{\text{ref}})] \end{aligned}$$

811 Which concludes the proof.

812 B.2 Detailed derivation of regularization effect in DAAs

$$\begin{aligned}
& \mathbb{E}_{(\mathbf{x}, \mathbf{y}^w, \mathbf{y}^l) \sim (D, \pi_{\text{ref}})} \left[\nabla_{\theta} \frac{1}{2} \rho_{\theta}^2 \right] \\
&= \mathbb{E}_{(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) \sim (D, \pi_{\text{ref}})} \left[\nabla_{\theta} \frac{1}{2} \rho_{\theta}^2 \right] \\
&= 2 \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim (D, \pi_{\text{ref}})} \left[\log \frac{\pi_{\theta}(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})} \nabla \log \pi_{\theta}(\mathbf{y}|\mathbf{x}) \right] - \\
&\quad 2 \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim (D, \pi_{\text{ref}})} \left[\log \frac{\pi_{\theta}(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})} \right] \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim (D, \pi_{\text{ref}})} [\nabla \log \pi_{\theta}(\mathbf{y}|\mathbf{x})] \\
&= 2 \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim (D, \pi_{\text{ref}})} \left[\log \frac{\pi_{\theta}(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})} \nabla \log \pi_{\theta}(\mathbf{y}|\mathbf{x}) \right] + \\
&\quad 2 D_{KL} [\pi_{\text{ref}}(\cdot|\mathbf{x}) | \pi_{\theta}(\cdot|\mathbf{x})] \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim (D, \pi_{\text{ref}})} [\nabla \log \pi_{\theta}(\mathbf{y}|\mathbf{x})]
\end{aligned}$$

813 C Training Details

For the following dialogue history to a chatbot, which response is more helpful?

Dialogue history:
<dialogue history>

Response A:
<Response A>

Response B: <Response B>

FIRST provide a one-sentence comparison of the two responses and explain which you feel is more helpful. SECOND, on a new line, state only "A" or "B" to indicate which response is more helpful. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
More helpful: <"A" or "B">

Table 2: Prompt for GPT-4 evaluation on the dialogue generation task. Texts in blue are placeholders to be substituted by the real data.

814 **SFT Training** For the summarization task, we use the SFT split of Reddit TL;DR summarization.
815 For Anthropic-HH we use the chosen responses from the preference dataset for SFT stage. We pool
816 together both datasets into a single SFT dataset.

817 **Preference Training** For TL;DR summarization dataset, we train all methods for 2 epochs. To
818 evaluate the efficiency of addressing the over-optimization problem, we vary the regularization
819 strength $\beta \in \{0.01, 0.05, 0.1\}$.

820 Across all SFT and Preference training, we use a global batch size of 64 (with 4 gradient accumulation
821 steps), and AdamW optimizer with a learning rate of 1×10^{-6} (cosine learning rate scheduler warm-
822 up for 100 steps) and a max length of 640.

823 **Golden Reward Training details** We follow the synthetic setup where the *golden* reward model
824 serves as human evaluation and provides preference labels [Gao et al., 2022, Tang et al., 2024a].

825 We first initialize the golden reward model with a SFT version of Llama-3.1-8B on the pooled
826 SFT data. We then train the golden reward model on the combined preference of the TL;DR and
827 Anthropic-HH dataset. Specifically, we use a batch size of 128, with a learning rate of 5×10^{-6} , we
828 train for one epoch with a cosine learning rate schedule.

829 The golden reward model achieves high validation accuracies with 75.2% validation accuracy,
830 showing high correlation with human preferences.

Details of KL Divergence Estimation In this paper, we construct an unbiased estimate of $KL(\pi_\theta || \pi_{\text{ref}})$ by sampling. More specifically, we first sample N prompts $\{x^i\}$ from the evaluation set, for each prompt x^i , we sample a response $y \sim \pi_\theta(\cdot|x^i)$ from the learned policy π_θ . Following [Schulman, 2020], The KL divergence is estimated as follows:

$$\frac{1}{N} \sum_{i=1}^N \log \frac{\pi_\theta(y^i|x^i)}{\pi_{\text{ref}}(y^i|x^i)} + \left(\frac{\pi_\theta(y^i|x^i)}{\pi_{\text{ref}}(y^i|x^i)} - 1 \right)$$

831 **Compute Resources Specification** We train and evaluate our models using NVIDIA 4xH100 GPUs.
832 All evaluations are computed with "gpt-4o-mini" as judge, with random positional flips to avoid
833 position bias.

834 D Evaluation Details

835 **Golden reward Evaluation** we sample 2 completions per prompt from the learned policy with 512
836 prompts from the evaluation set. We sample with temperature $\tau = 0.7$ and top- p sampling with
837 $p = 0.95$. to evaluate its performance. To calculate the winrate, we consider all combinations of
838 pairs between the completions generated by the learned policy and the reference completions from
839 the preference dataset, and then compare the scores from the golden reward model on the pair of
840 generations to calculate win-rate.

841 **GPT-4 Evaluation** For GPT-4 evaluation, we sample 256 prompts from the evaluation set. For each
842 prompt, we sample 1 completion from the learned policy. To evaluate the dialogue generation task,
843 we use the prompt shown in Table 2, similar to [Ji et al., 2024] with random position flipping to avoid
844 position bias.