Towards Efficient Exact Optimization of Language Model Alignment

Haozhe Ji[1], Cheng Lu[2], Yilin Niu[3], Pei Ke[1],

Hongning Wang[1], Jun Zhu[2], Jie Tang[4], Minlie Huang[1]

[1]CoAI Group, [2]TSAIL Group, [3]Zhipu AI, [4]KEG

Introduction

- *Aligning language models (LMs) to generate human preferred responses is crucial to the development of* ***reliable*** *AI systems.*
- *It is essential to develop* ***principled*** *and* ***scalable*** *alignment method.*
- ***Principle****: Theoretically grounded in principle.*
- ***Scalable****: Accommodate to growing scale.*

Introduction

- *The Recipe of LM alignment [**Ouyang et al., 2022**]:*
  - ***SFT stage**: Supervised Fine-Tuning*

$\boldsymbol{x}$ → → $\boldsymbol{y}$

prompt *response*

$$\mathcal{L}_{\text{sft}}(\pi_\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}^{\text{sft}}}\Big[-\log \pi_\theta(\boldsymbol{y}|\boldsymbol{x})\Big]$$

  - ***RM stage**: Reward Modeling*

$\boldsymbol{x}$ $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ → → $\boldsymbol{y}_w > \boldsymbol{y}_l$

prompt *responses* *preference*

$$\mathcal{L}_r(r_\phi) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^{\text{pref}}}\left[-\log \frac{e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_w)}}{e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_w)} + e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_l)}}\right]$$

  - ***Alignment stage**: Learning with (proxy) Human Feedback*

$$\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\Big)$$

*Reward model (from **RM stage**)*

*SFT policy (from **SFT stage**)*

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022)

Introduction

- *Reinforcement Learning from Human Feedback (RLHF)* [**Ouyang et al., 2022**]:
  - ***PPO***: *Framing as* ***KL-regularized reward maximization*** *and solved by RL.*

$$\mathcal{J}_{\text{lhf}}^{\beta}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\right)$$

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022)

Introduction

- *Reinforcement Learning from Human Feedback (RLHF)* [**Ouyang et al., 2022**]:
  - ***PPO***: *Framing as **KL-regularized reward maximization** and solved by RL.*

$$\mathcal{J}_{\text{lhf}}^{\beta}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\right)$$

$$R(\boldsymbol{x},\boldsymbol{y}) = r_\phi(\boldsymbol{x},\boldsymbol{y}) - \beta\log\frac{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$$

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022)

Introduction

- *Reinforcement Learning from Human Feedback (RLHF)* [**Ouyang et al., 2022**]:
  - ***PPO***: *Framing as **KL-regularized reward maximization** and solved by RL.*

$$\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\right)$$

$$R(\boldsymbol{x},\boldsymbol{y}) = r_\phi(\boldsymbol{x},\boldsymbol{y}) - \beta\log\frac{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$$

$$\nabla_\theta\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}},\boldsymbol{y}\sim\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}\left[R(\boldsymbol{x},\boldsymbol{y})\nabla_\theta\log\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\right]$$

*Policy gradient method, e.g., PPO [**Schulman et al., 2017**]*

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022)

Introduction

- *Reinforcement Learning from Human Feedback (RLHF)* [**Ouyang et al., 2022**]:
  - ***PPO***: *Framing as **KL-regularized reward maximization** and solved by RL.*

$$\mathcal{J}_{\text{lhf}}^{\beta}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\right)$$

$$R(\boldsymbol{x},\boldsymbol{y}) = r_\phi(\boldsymbol{x},\boldsymbol{y}) - \beta\log\frac{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$$

$$\nabla_\theta\mathcal{J}_{\text{lhf}}^{\beta}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}},\boldsymbol{y}\sim\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}\left[R(\boldsymbol{x},\boldsymbol{y})\nabla_\theta\log\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\right]$$

*Policy gradient method, e.g., PPO [**Schulman et al., 2017**]*

*RL has **high variance** in policy gradient estimation*
*RL needs to **sample in training loop***
} ***Inefficiency** of convergence*

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022)

Introduction

- Direct Preference Optimization (DPO) [**Rafailov et al., 2023**]:
  - **Key intuition**: Policy optimization as reward modeling.

$$\mathcal{J}_{\mathrm{lhf}}^{\beta}(\pi_\theta)$$

Alignment objective

→ *KKT condition* →

$$\pi_\beta^*(\boldsymbol{y}|\boldsymbol{x}) = \pi_{\mathrm{sft}}(\boldsymbol{y}|\boldsymbol{x}) \frac{e^{\frac{1}{\beta} r_\phi(\boldsymbol{x},\boldsymbol{y})}}{Z_\beta(\boldsymbol{x})}$$

Analytic solution of maximizing $\mathcal{J}_{\mathrm{lhf}}^{\beta}(\pi_\theta)$

↓ *Simple algebra*

$$r_\phi(\boldsymbol{x},\boldsymbol{y}) = \beta \log \frac{\pi_\beta^*(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\mathrm{sft}}(\boldsymbol{y}|\boldsymbol{x})} + \beta \log Z_\beta(\boldsymbol{x})$$

Reward model as a function of π_β^*

← *BT model* ←

$$\mathcal{L}_{\mathrm{dpo}}(\pi_\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^{\mathrm{pref}}}\left[-\log\sigma\left(\beta \log \frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\mathrm{sft}}(\boldsymbol{y}_w|\boldsymbol{x})} - \beta \log \frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\mathrm{sft}}(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right]$$

DPO: Optimize the policy using preference loss

Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." *Advances in Neural Information Processing Systems* 36 (2024)

Introduction

- *Direct Preference Optimization (DPO)* [**Rafailov et al., 2023**]:
  - ***Key intuition***: *Policy optimization as reward modeling.*

$$\mathcal{J}^{\beta}_{\mathrm{lhf}}(\pi_\theta)$$

Alignment objective

KKT condition →

Assume ***unlimited*** *model capacity*

$$\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x}) = \pi_{\mathrm{sft}}(\boldsymbol{y}|\boldsymbol{x}) \frac{e^{\frac{1}{\beta} r_\phi(\boldsymbol{x},\boldsymbol{y})}}{Z_\beta(\boldsymbol{x})}$$

Analytic solution of maximizing $\mathcal{J}^{\beta}_{\mathrm{lhf}}(\pi_\theta)$

↓ *Simple algebra*

$$r_\phi(\boldsymbol{x},\boldsymbol{y}) = \beta \log \frac{\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\mathrm{sft}}(\boldsymbol{y}|\boldsymbol{x})} + \beta \log Z_\beta(\boldsymbol{x})$$

Reward model as a function of π^*_β

← *BT model*

$$\mathcal{L}_{\mathrm{dpo}}(\pi_\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^{\mathrm{pref}}}\left[- \log \sigma\left(\beta \log \frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\mathrm{sft}}(\boldsymbol{y}_w|\boldsymbol{x})} - \beta \log \frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\mathrm{sft}}(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right]$$

DPO: Optimize the policy using preference loss

? ↕

  - *DPO is* ***not exactly*** *optimizing the alignment objective.*

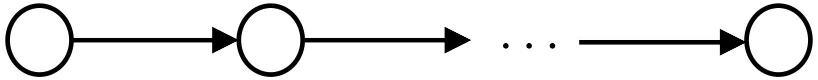
Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." *Advances in Neural Information Processing Systems* 36 (2024)

Introduction

- *Practical constraint: The expressivity gap between π_θ and π_β^**

Local-normalization

$$\pi_\theta(\boldsymbol{y}|\boldsymbol{x}) = \pi_\theta(y_1|\boldsymbol{x}) \; \pi_\theta(y_2|\boldsymbol{x}, y_1) \; \cdots \; \pi_\theta(y_n|\boldsymbol{x}, y_1, \cdots, y_{n-1})$$



Auto-Regressive Model (ARM)

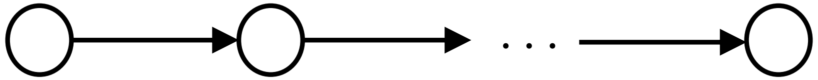
Lin, Chu-Cheng, et al. "Limitations of autoregressive models and their alternatives." *NAACL* (2021)

Introduction

- *Practical constraint: The expressivity gap between π_θ and π_β^**

Local-normalization

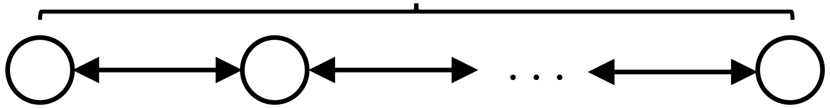
$$\pi_\theta(\boldsymbol{y}|\boldsymbol{x}) = \pi_\theta(y_1|\boldsymbol{x}) \ \pi_\theta(y_2|\boldsymbol{x}, y_1) \ \cdots \ \pi_\theta(y_n|\boldsymbol{x}, y_1, \cdots, y_{n-1})$$



Auto-Regressive Model (ARM)

Global-normalization

$$\pi_\beta^*(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left[\beta^{-1} r_\phi(\boldsymbol{x}, y_1, y_2, \cdots, y_n)\right]$$



Energy-Based Model (EBM)

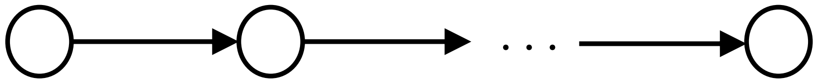
Lin, Chu-Cheng, et al. "Limitations of autoregressive models and their alternatives." *NAACL* (2021)

Introduction

- *Practical constraint: The expressivity gap between π_θ and π_β^**

Local-normalization

$$\pi_\theta(\boldsymbol{y}|\boldsymbol{x}) = \pi_\theta(y_1|\boldsymbol{x}) \;\; \pi_\theta(y_2|\boldsymbol{x}, y_1) \;\; \cdots \;\; \pi_\theta(y_n|\boldsymbol{x}, y_1, \cdots, y_{n-1})$$

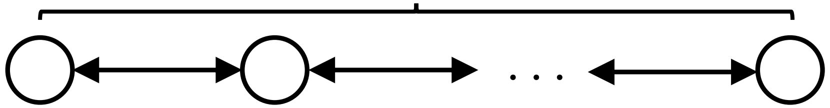


Auto-Regressive Model (ARM)

Pros: Efficient sampling in O(Poly(n)) time
Cons: Assume AR factorization of Prob(sequence)

Global-normalization

$$\pi_\beta^*(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left[\beta^{-1} r_\phi(\boldsymbol{x}, y_1, y_2, \cdots, y_n)\right]$$



Energy-Based Model (EBM)

Pros: No assumption on modeling Prob(sequence)
Cons: Inefficient sampling in O(Superpoly(n))

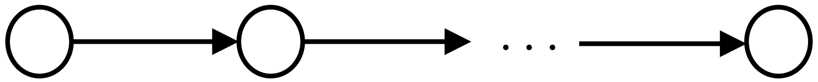
Lin, Chu-Cheng, et al. "Limitations of autoregressive models and their alternatives." *NAACL* (2021)

Introduction

- *Practical constraint: The expressivity gap between π_θ and π^*_β*

Local-normalization

$$\pi_\theta(\boldsymbol{y}|\boldsymbol{x}) = \pi_\theta(y_1|\boldsymbol{x}) \;\; \pi_\theta(y_2|\boldsymbol{x}, y_1) \quad \cdots \quad \pi_\theta(y_n|\boldsymbol{x}, y_1, \cdots, y_{n-1})$$

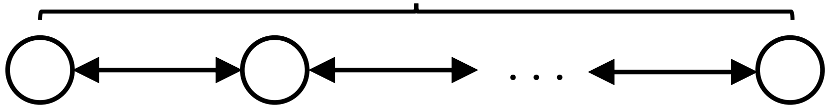


Auto-Regressive Model (ARM)

Pros: Efficient sampling in O(Poly(n)) time
Cons: Assume AR factorization of Prob(sequence)

Global-normalization

$$\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left[\beta^{-1} r_\phi(\boldsymbol{x}, y_1, y_2, \cdots, y_n)\right]$$



Energy-Based Model (EBM)

Pros: No assumption on modeling Prob(sequence)
Cons: Inefficient sampling in O(Superpoly(n))

- *Theoretical justification [**Lin et al., 2021**]:*
  - *There are some "hard" sequences whose unnormalized scores are easy to compute, yet the conditional local probabilities are **intractable**.*
  - *ARMs **cannot perfectly** capture all EBM distributions with O(Poly(n))-sized parameters.*

Lin, Chu-Cheng, et al. "Limitations of autoregressive models and their alternatives." *NAACL* (2021)

Introduction

- *What does the solution of RLHF look like under this practical constraint?*
  - *KL-regularized RL as probability matching [**Korbak et al., 2021**].*

equivalent

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})]-\beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\right) \Longleftrightarrow \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left[\mathbb{D}_{\text{KL}}(\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}))\right]$$

Maximize reward with KL penalty — *Minimize reverse KL divergence*

Korbak, Tomasz, et al. "RL with KL penalties is better viewed as Bayesian inference." *arXiv preprint arXiv:2205.11275* (2022)

Introduction

- *What does the solution of RLHF look like under this practical constraint?*
  - *KL-regularized RL as probability matching [**Korbak et al., 2021**].*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\right) \quad \overset{\text{equivalent}}{\Longleftrightarrow} \quad \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left[\mathbb{D}_{\text{KL}}(\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}))\right]$$

Maximize reward with KL penalty — *Minimize reverse KL divergence*

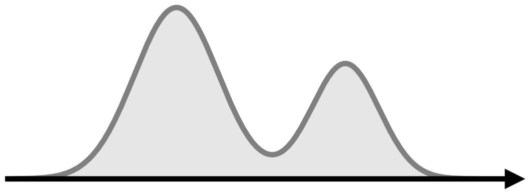
  - *The asymmetry of KL divergence:*
    - *Estimate the density of* p

Forward KL

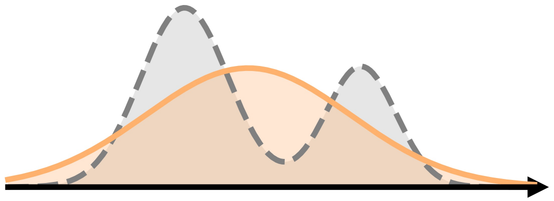
$$\mathbb{D}_{\text{KL}}(p\|\hat{p}) = \mathbb{E}_{x\sim p}\left[\log\frac{p(x)}{\hat{p}(x)}\right]$$

Reverse KL

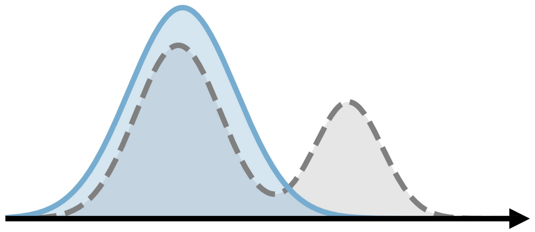
$$\mathbb{D}_{\text{KL}}(\hat{p}\|p) = \mathbb{E}_{x\sim \hat{p}}\left[\log\frac{\hat{p}(x)}{p(x)}\right]$$



Target distribution $p(x)$



Mean-seeking solution



Mode-seeking solution

Korbak, Tomasz, et al. "RL with KL penalties is better viewed as Bayesian inference." *arXiv preprint arXiv:2205.11275* (2022)

Method

- ***Key motivation***: *Policy optimization as probability matching.*
- *Without loss of generality, consider the generalized alignment objective:*

$$\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta_r \mathbb{D}_{\text{KL}}[\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \| \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\Big)$$

Method

- ***Key motivation***: *Policy optimization as probability matching.*
- *Without loss of generality, consider the generalized alignment objective:*

$$\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta_r \mathbb{D}_{\text{KL}}[\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \| \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\Big)$$

  - $\pi_\theta^{\beta_\pi}$ *is the geometric mean of* π_θ *and* π_{sft}

$$\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_\theta(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{1-\beta_\pi}$$

Method

- ***Key motivation***: *Policy optimization as probability matching.*
- *Without loss of generality, consider the generalized alignment objective:*

$$\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left(\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta_r \mathbb{D}_{\text{KL}}[\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\right)$$

- $\pi_\theta^{\beta_\pi}$ *is the geometric mean of* π_θ *and* π_{sft}

$$\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_\theta(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{1-\beta_\pi}$$

- *Decompose the KL regularization*

$$\beta = \beta_r \cdot \beta_\pi$$

regularize reward (β_r) *regularize policy* (β_π)

- *Analytic solution is also* π_β^**.*
- *Unify the regularization setting of PPO* $(\beta_\pi = 1, \beta_r = \beta)$ *and DPO* $(\beta_\pi = \beta, \beta_r = 1)$

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\left[\log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})}\right]$$

  - *Calculating reverse KL requires sampling from* $\pi_\theta^{\beta_\pi}$ *, which prohibits straightforward back propagation.*

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\left[\log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})}\right]$$

Importance Sampling (IS)
π_{sft} *as the proposal distribution*

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})} \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})}\right]$$

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\left[\log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})}\right]$$

Importance Sampling (IS)
π_{sft} *as the proposal distribution*

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})} \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})}\right]$$

Define $f_\theta(\boldsymbol{x}, \boldsymbol{y}) = \log \pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) - \log \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$
as the log policy ratio

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})} \log \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})} e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}}\right]$$

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})} \log \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})} e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}}\right]$$

  - *The partition function* $Z_{\beta_r}(\boldsymbol{x})$ *is intractable.*

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})} \log \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})} e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}} \right]$$

  - *The partition function* $Z_{\beta_r}(\boldsymbol{x})$ *is intractable.*
  - *Inspiration from Self-Normalized Importance Sampling (SNIS)*
    - *Estimate* $\mathbb{E}_{x\sim p}[f(x)]$ *where we can only compute the* ***unnormalized*** $P(x)$

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})} \log \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})} e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}} \right]$$

  - *The partition function* $Z_{\beta_r}(\boldsymbol{x})$ *is intractable.*
  - *Inspiration from Self-Normalized Importance Sampling (SNIS)*
    - *Estimate* $\mathbb{E}_{x\sim p}[f(x)]$ *where we can only compute the* ***unnormalized*** $P(x)$

$$\mathbb{E}_{x\sim p}[f(x)] = \sum_x p(x) f(x)$$

$$p(x) = \frac{P(x)}{\sum_x P(x)}$$

$$\mathbb{E}_{x\sim p}[f(x)] = \frac{\sum_x P(x) f(x)}{\sum_x P(x)}$$

$$\frac{\sum_x P(x) f(x)}{\sum_x P(x)} = \frac{\mathbb{E}_q[\frac{P(x)}{q(x)} f(x)]}{\mathbb{E}_q[\frac{P(x)}{q(x)}]}$$

$$\mathbb{E}_{x\sim p}[f(x)] = \lim_{N\to\infty} \frac{\sum_{i=1}^N \frac{P(x_i)}{q(x_i)} f(x_i)}{\sum_{i=1}^N \frac{P(x_i)}{q(x_i)}}$$

where $x_1, \cdots, x_N \sim q$ *are i.i.d. samples*

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})} \log \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})} e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}} \right]$$

$$Z_{\beta_r}(\boldsymbol{x}) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}[\exp(\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r})]$$

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})} \log \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})} e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}} \right]$$

$$Z_{\beta_r}(\boldsymbol{x}) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}[\exp(\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r})]$$

- *Sample K i.i.d. continuations* $\boldsymbol{y}_{1:K} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K\}$ *from* $\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \lim_{K\to\infty} \sum_{k=1}^{K} \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_k)}}{\sum_{j=1}^{K} e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_j)}} \log \frac{\frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_k)}}{\sum_{j=1}^{K} e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_j)}}}{\frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_k)}}{\sum_{j=1}^{K} \frac{1}{\beta_r} e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}}$$

$p_{f_\theta}(i|\boldsymbol{y}_{1:K}, \boldsymbol{x})$

Distribution of log policy ratio

$p_{r_\phi}(i|\boldsymbol{y}_{1:K}, \boldsymbol{x})$

Distribution of reward model

Method

- *Deriving the probability matching objective of* $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})} \log \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})} e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}} \right]$$

$$Z_{\beta_r}(\boldsymbol{x}) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}[\exp(\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r})]$$

  - *Sample K i.i.d. continuations* $\boldsymbol{y}_{1:K} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K\}$ *from* $\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi} \| \pi_{\beta_r}^*) = \lim_{K\to\infty} \sum_{k=1}^{K} \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_k)}}{\sum_{j=1}^K e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_j)}} \log \frac{\frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_k)}}{\sum_{j=1}^K e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_j)}}}{\frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_k)}}{\sum_{j=1}^K \frac{1}{\beta_r} e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}}$$

Reverse KL $\mathbb{D}_{\text{KL}}(p_{f_\theta} || p_{r_\phi})$ *of* p_{f_θ} *and* p_{r_ϕ}

Method

- *Introduce the Efficient Exact Optimization (**EXO**) objective of alignment*
  - ***Learning from the reward model***

$$\mathcal{L}_{\text{exo}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[\mathbb{D}_{\text{KL}}\left(p_{f_\theta}(\cdot|\boldsymbol{y}_{1:K},\boldsymbol{x})\|p_{r_\phi}(\cdot|\boldsymbol{y}_{1:K},\boldsymbol{x})\right)\right]$$

    - *Where we define:* *regularize policy* *regularize reward*

$$p_{f_\theta}(i|\boldsymbol{y}_{1:K},\boldsymbol{x}) = \frac{e^{\beta_\pi \log\frac{\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^K e^{\beta_\pi \log\frac{\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}} \qquad p_{r_\phi}(i|\boldsymbol{y}_{1:K},\boldsymbol{x}) = \frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^K e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}$$

  - ***Learning from the preference data** (K=2)*

$$\mathcal{L}_{\text{exo-pref}}(\pi_\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^{\text{pref}}}\left[\mathbb{D}_{\text{KL}}\left(p_{f_\theta}(\cdot|\boldsymbol{y}_w,\boldsymbol{y}_l,\boldsymbol{x})\|p_{r_h}(\cdot|\boldsymbol{y}_w,\boldsymbol{y}_l,\boldsymbol{x})\right)\right]$$

    - *Where the preference probability $p_{r_h}(\cdot\,|\boldsymbol{y}_w,\boldsymbol{y}_l,\boldsymbol{x})$ is a label-smoothed one-hot distribution.*

Method

- *Justification of exactness*
  - *The gradient of EXO aligns with the gradient of the generalized alignment objective and the reverse KL asymptotically for policy with* ***arbitrary*** θ *when* $K \to \infty$**.**

$$\nabla_\theta \mathcal{L}_{\text{exo}}(\pi_\theta) = \nabla_\theta \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\big[\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \| \pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}))\big]$$
$$= -\frac{1}{\beta_r}\nabla_\theta \mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}).$$

  - *EXO reaches the same* ***mode-seeking*** *solution as RLHF.*
  - *In practice, EXO converges effectively and efficiently with finite K (will be shown later empirically).*

Comparison with DPO

- *Generalizing DPO:*
  - *Sample K completions* $\boldsymbol{y}_{1:K} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K\}$ *from* $\pi_{\text{sft}}(y|x)$
  - *Substitute hard human preference with soft distribution defined by reward model*

$$\mathcal{L}_{\text{dpo-rw}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[-\sum_{i=1}^{K}\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}\log\frac{e^{\beta_\pi\log\frac{\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^{K}e^{\beta_\pi\log\frac{\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}}\right]$$

Forward KL $\mathbb{D}_{\text{KL}}(p_{f_\theta}||p_{r_\phi})$ *of* p_{f_θ} *and* p_{r_ϕ} *(up to a constant)*

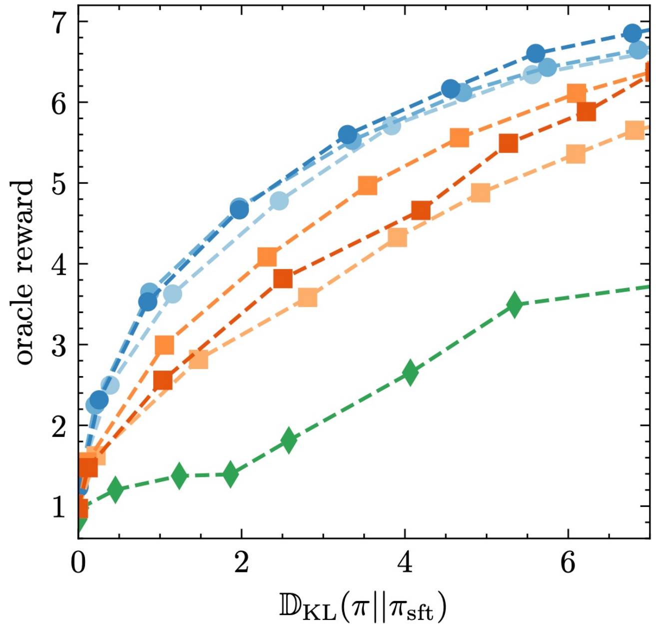
  - *The gradient of DPO-rw aligns with the gradient of the forward KL asymptotically for policy with* ***arbitrary*** θ *when* $K \to \infty$.

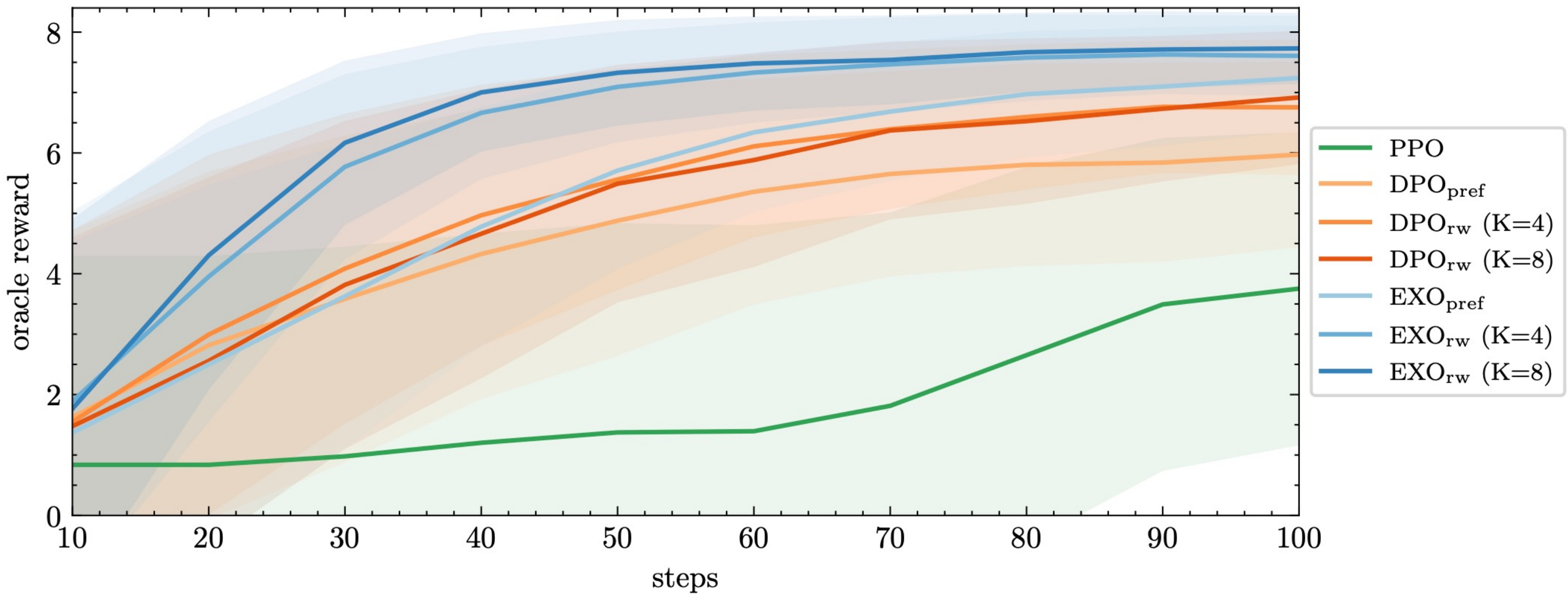
$$\nabla_\theta\mathcal{L}_{\text{dpo-rw}}(\pi_\theta) = \nabla_\theta\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left[\mathbb{D}_{\text{KL}}(\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x})\|\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}))\right]$$

- ***Inexactness****: DPO minimizes the forward KL, while RLHF, e.g., PPO minimizes the reverse KL.*

Experiments

- *Synthetic experiment: Generate IMDB review with positive sentiment*
  - *Oracle reward (Human labeler): Classifier trained on IMDB review classification dataset*





Oracle reward** vs **KL

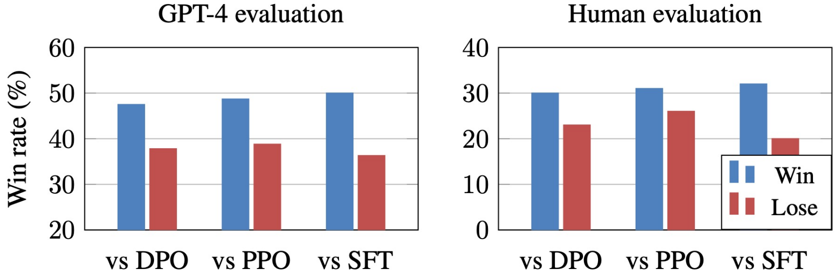
Oracle reward** vs **Training steps

Experiments

- *Alignment on real human preferences:*
  - *Text summarization: TL;DR preference dataset*
  - *Dialogue generation: Anthropic-HH dataset (helpfulness subset)*
  - *Instruction following: Filtered real user query from an online API*

Method	Reward Model (%) vs SFT	Reward Model (%) vs Chosen	GPT-4 (%) vs SFT	GPT-4 (%) vs Chosen
w/ Preferences				
DPO_{pref}	68.3	23.7	57.0	30.5
EXO_{pref}	**92.5**	**60.1**	**83.0**	**55.0**
w/ Reward Model				
Best-of-N	99.3	75.8	83.5	60.0
PPO	93.2	58.3	77.0	52.0
DPO_{rw}	82.7	39.8	70.0	41.0
EXO_{rw}	**97.3**	**76.4**	**88.5**	**64.0**

Method	Reward Model (%) vs SFT	Reward Model (%) vs Chosen	GPT-4 (%) vs SFT	GPT-4 (%) vs Chosen
w/ Preferences				
DPO_{pref}	66.3	65.1	58.0	37.0
EXO_{pref}	**76.4**	**76.7**	**73.0**	**51.0**
w/ Reward Model				
Best-of-N	94.6	98.2	86.0	63.0
PPO	75.0	74.0	66.5	52.0
DPO_{rw}	79.9	81.3	75.5	49.0
EXO_{rw}	**85.6**	**87.2**	**83.5**	**60.0**



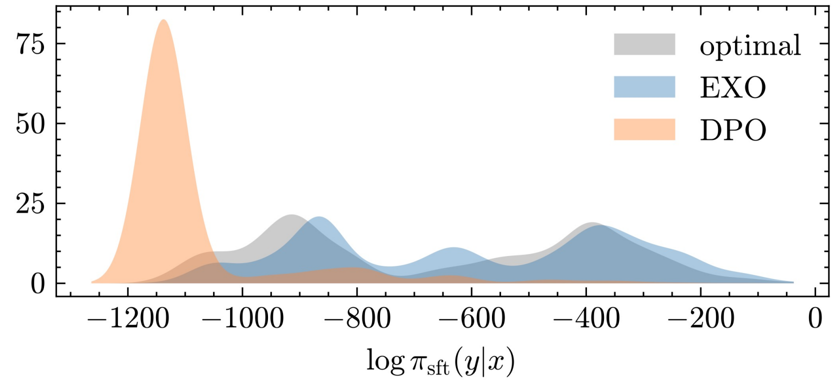
  - *Outperforms DPO and PPO in both settings of learning from preferences & reward model.*
  - *On par with Best-of-N (N=128) but much more computationally efficient in inference.*
  - *Scaling to realistic instruction-following dataset with consistent improvement.*

Experiments

- *Visualization: Compare the density of DPO and EXO with the optimal policy*
  - *Given a test prompt "**This Fox spectacle was a big hit when released in** "*
  - *Estimate the empirical policy distribution of π_θ and π^*_β by SNIS:*

$$\hat{\pi}_\theta(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\sum_{j=1}^{M}\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})/\pi_{\mathrm{sft}}(\boldsymbol{y}_j|\boldsymbol{x})} \qquad \hat{\pi}^*_\beta(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_{\mathrm{sft}}(\boldsymbol{y}_i|\boldsymbol{x})\exp(r(\boldsymbol{x},\boldsymbol{y}_i)/\beta)}{\sum_{j=1}^{M}\exp(r(\boldsymbol{x},\boldsymbol{y}_j)/\beta)}$$

  - *Use Kernel Density Estimation to estimate the density and plot the ratio* $\rho_{\hat{\pi}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\hat{\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\mathrm{sft}}(\boldsymbol{y}|\boldsymbol{x})}$

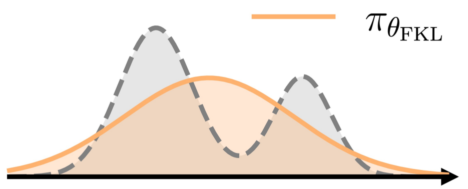


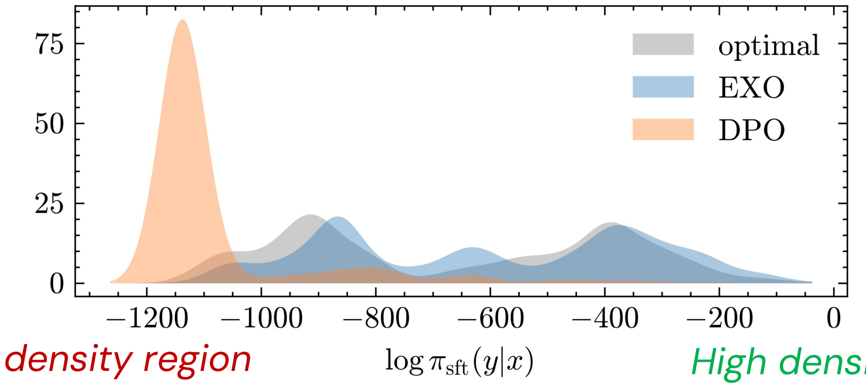
Experiments

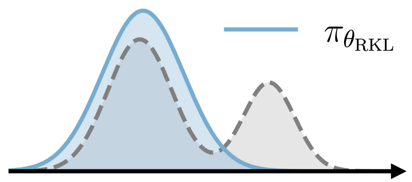
- *Visualization: Compare the density of DPO and EXO with the optimal policy*
  - *Given a test prompt "**This Fox spectacle was a big hit when released in** "*
  - *Estimate the empirical policy distribution of π_θ and π^*_β by SNIS:*

$$\hat{\pi}_\theta(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\sum_{j=1}^{M}\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})/\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})} \qquad \hat{\pi}^*_\beta(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})\exp(r(\boldsymbol{x},\boldsymbol{y}_i)/\beta)}{\sum_{j=1}^{M}\exp(r(\boldsymbol{x},\boldsymbol{y}_j)/\beta)}$$

  - *Use Kernel Density Estimation to estimate the density and plot the ratio* $\rho_{\hat{\pi}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\hat{\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$







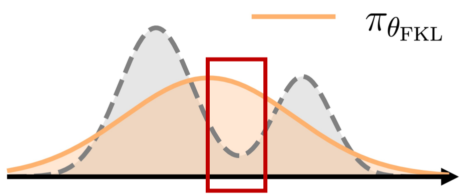
⇠ *Low density region* *High density region* ⇢

Experiments

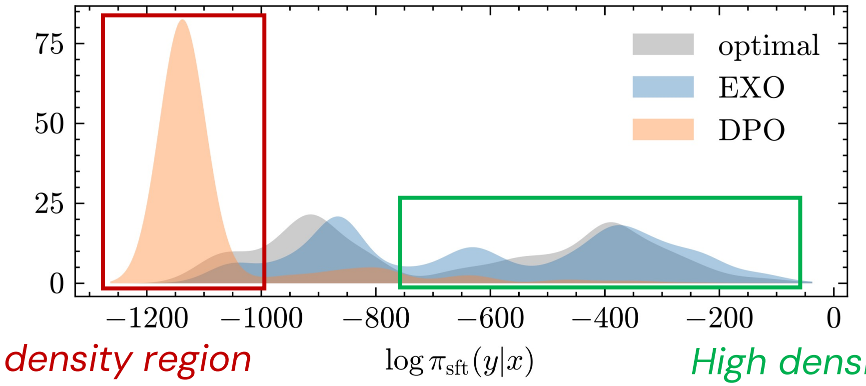
- *Visualization: Compare the density of DPO and EXO with the optimal policy*
  - *Given a test prompt "**This Fox spectacle was a big hit when released in** "*
  - *Estimate the empirical policy distribution of π_θ and π_β^* by SNIS:*

$$\hat{\pi}_\theta(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\sum_{j=1}^{M}\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})/\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})} \qquad \hat{\pi}_\beta^*(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})\exp(r(\boldsymbol{x},\boldsymbol{y}_i)/\beta)}{\sum_{j=1}^{M}\exp(r(\boldsymbol{x},\boldsymbol{y}_j)/\beta)}$$

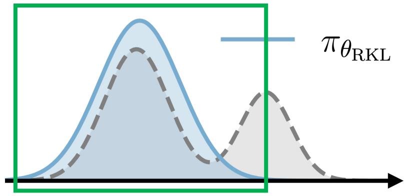
  - *Use Kernel Density Estimation to estimate the density and plot the ratio* $\rho_{\hat{\pi}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\hat{\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$



Overestimate the long tail



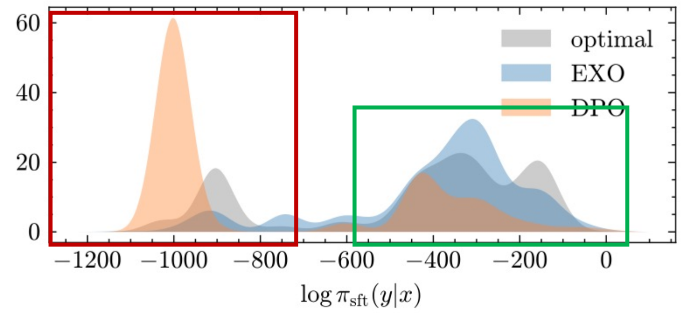
⇠ Low density region *High density region ⇢*



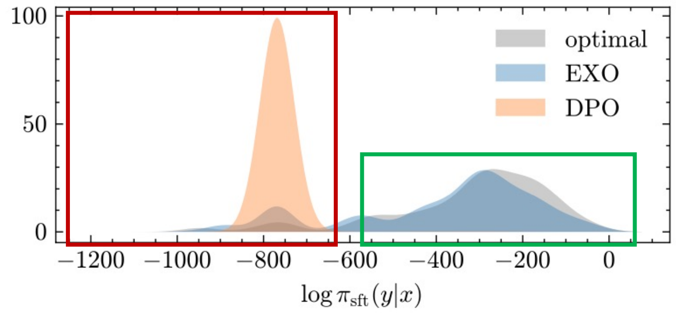
Concentrating on the mode

Experiments

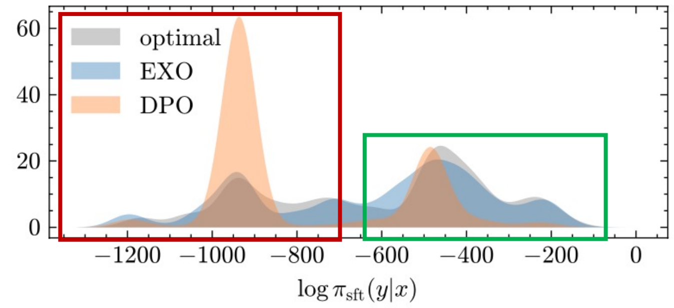
- *More visualization cases: (prevailing phenomenon, no cherry-picking)*



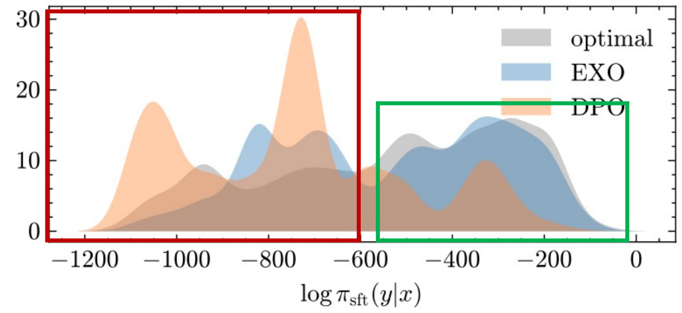
Estimated density ratio of the EXO, DPO and optimal policy given the prompt "*Is this supposed to be serious? I hope not*".



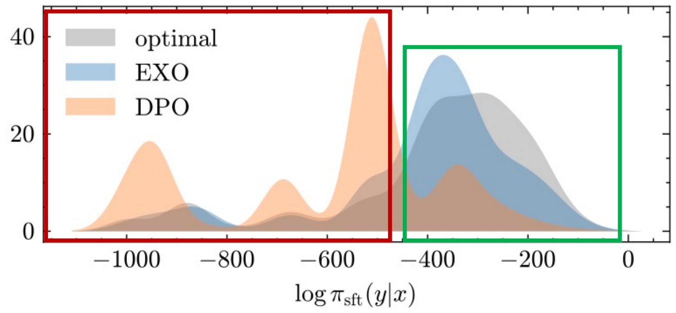
Estimated density ratio of the EXO, DPO and optimal policy given the prompt "*Great book, great movie, great soundtrack. Frank*".



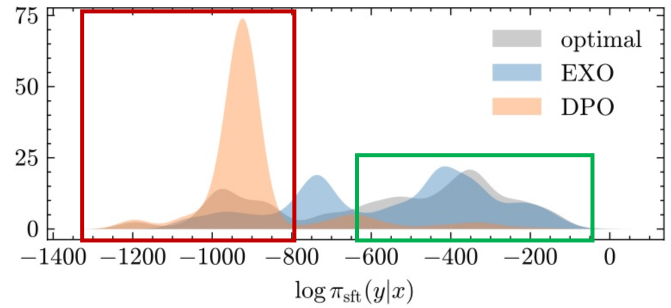
Estimated density ratio of the EXO, DPO and optimal policy given the prompt "*What we have here the standard Disney direct to DVD*".



Estimated density ratio of the EXO, DPO and optimal policy given the prompt "*This is indeed the film that popularized kung*".



Estimated density ratio of the EXO, DPO and optimal policy given the prompt "*This movie is about a group of people who are*".



Estimated density ratio of the EXO, DPO and optimal policy given the prompt "*Once the slow beginning gets underway, the film kicks*".

Conclusion

- *We unify PPO and DPO under the framework of density estimation, and examine that PPO is actually minimizing the **reverse KL** to the optimal policy; while DPO is minimizing the **forward KL** to the optimal policy.*
- *We propose efficient exact optimization (EXO) for language model alignment problem. Specifically, EXO **exactly** optimizes the alignment objective in RLHF, while being **efficient** in optimization by formulating as probability matching.*

Q & A

Homepage: https://haozheji.github.io

GitHub repo: https://github.com/haozheji/exact-optimization

Conversational AI Group of Tsinghua University: http://coai.cs.tsinghua.edu.cn/