The Crucial Role of Samplers in Online Direct Preference Optimization

Ruizhe Shi*, Runlong zhou*, Simon S. Du

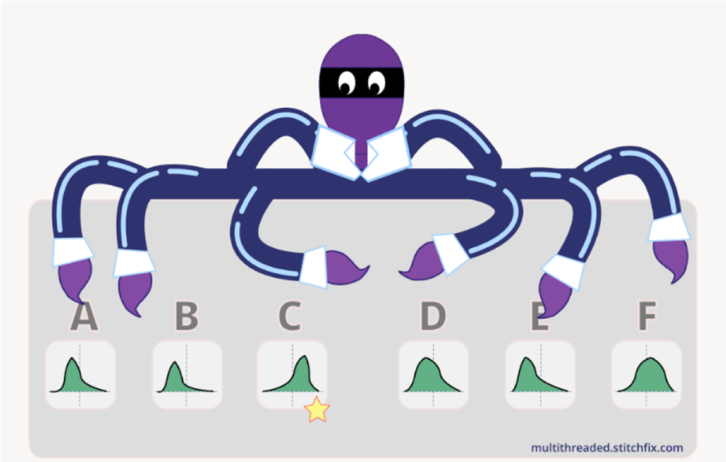
Mar 9 2025

ICLR 2025

Bandit view of language model alignment

- **Prompt (State)**: user queries (x)
- **Response (Action)**: language model generation (y)
- **Reward** function: $r(x, y) \in [0,1]$

(in this project we only study MAB)



Picture from https://multithreaded.stitchfix.com/blog/2020/08/05/bandits/

Policy

- A **tabular softmax** policy π_θ for MABs satisfies

$$\pi_\theta(y) = \frac{\mathrm{e}^{\theta_y}}{\sum_{y'} \mathrm{e}^{\theta_{y'}}}$$

Preference-based RL

- A **preference** model $p^\star(y_1 \succ y_2)$ indicating the probability that y_1 is preferred over y_2
- After choosing a pair of arms (y_1, y_2), observe a sample $p \sim \text{Bernoulli}\big(p^\star(y_1 \succ y_2)\big)$
- BT preference model

$$p^\star(y_1 \succ y_2) = \sigma\big(r(y_1) - r(y_2)\big) = \frac{\mathrm{e}^{r(y_1)}}{\mathrm{e}^{r(y_1)} + \mathrm{e}^{r(y_2)}}$$

Motivation: how fast can data help DPO converge

- Human preference dataset $\mathcal{D} = \left\{\left(y_w^{(i)}, y_l^{(i)}\right)\right\}_{i=1}^{N}$
  - In the ith sample, $y_w^{(i)}$ is preferred over $y_l^{(i)}$
- DPO (a popular alignment algorithm)
  - $\mathcal{L}_\pi(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w^{(i)}\right)}{\pi_{\mathrm{ref}}\left(y_w^{(i)}\right)} - \beta\log\frac{\pi_\theta\left(y_l^{(i)}\right)}{\pi_{\mathrm{ref}}\left(y_l^{(i)}\right)}\right)$
  - Closed-form solution: $\pi^\star(y) = \frac{1}{Z}\pi_{\mathrm{ref}}(y)\mathrm{e}^{r(y)/\beta}$
- We want to study: *how fast can DPO converge to optimality with different sampling distributions on data*?

Motivation: how fast can data help DPO converge

How fast can $\underbrace{r(y) - r(y') - \beta \log \frac{\pi_{\theta^{(t)}}(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_{\theta^{(t)}}(y')}}_{=:\,\delta(y,y';\theta^{(t)})}$ ***converge to*** 0***, for*** $\forall y, y' \in \mathcal{Y}$***?***

Ideal Case: Exact DPO

- Suppose we have two **sampling policies** π^{s1} for y_1 and π^{s2} for y_2
- Define sampling probability

Stop gradient

$$\pi^{\mathsf{s}}(y, y') := \mathsf{sg}\left(\pi^{\mathsf{s1}}(y)\pi^{\mathsf{s2}}(y') + \pi^{\mathsf{s1}}(y')\pi^{\mathsf{s2}}(y)\right)$$

- Exact DPO loss function

$$\mathcal{L}_{\text{DPO}}(\theta) := -\sum_{y, y' \in \mathcal{Y}} \pi^{\mathsf{s}}(y, y') p^{\star}(y \succ y') \log \sigma\left(\beta \log \frac{\pi_\theta(y)\pi_{\mathsf{ref}}(y')}{\pi_{\mathsf{ref}}(y)\pi_\theta(y')}\right)$$

- Policy update

$$\theta^{(t+1)} = \theta^{(t)} - \eta\alpha(\pi^{\mathsf{s1}}, \pi^{\mathsf{s2}})\nabla_\theta \mathcal{L}_{\text{DPO}}(\theta^{(t)})$$

Sampling coefficient determined by samplers

Ideal Case: Exact DPO

- Mixture of samplers

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta \left(\alpha_1 \mathcal{L}_1(\theta^{(t)}) + \alpha_2 \mathcal{L}_2(\theta^{(t)}) \right)$$

  - Central to our design

Practical Case: Empirical DPO

- No access to exact gradients

$$\theta^{(t+1)} = \theta^{(t)} - \eta G^{(t)}$$

where $G_y^{(t)}$ is a random variable that

$$\frac{1}{\beta A}\left(G_y^{(t)} - \alpha(\pi^{\mathsf{s1}}, \pi^{\mathsf{s2}})\nabla_{\theta_y}\mathcal{L}(\theta^{(t)})\right) \sim \text{sub-Gaussian}(\sigma^2)$$

- Mixture of samplers

$$\frac{1}{\beta A}\left(G_y^{(t)} - \nabla_{\theta_y}\left(\alpha_1\mathcal{L}_1(\theta^{(t)}) + \alpha_2\mathcal{L}_2(\theta^{(t)})\right)\right) \sim \text{sub-Gaussian}(\sigma^2)$$

Main results

- Uniform sampler *(vanilla)* $\pi^{\mathsf{s1}}(\cdot) = \pi^{\mathsf{s2}}(\cdot) = \mathsf{Uniform}(\mathcal{Y})$

$$\left|\delta(y, y'; \theta^{(T)})\right| \leqslant 0.588^T \ , \ \forall y, y' \in \mathcal{Y}$$

$$\max_{y, y' \in \mathcal{Y}} \left|\delta(y, y'; \theta^{(T)})\right| \geqslant \gamma^T$$

linear convergence!

determined by initialization

- Policy-difference guided sampler *(ours)*

$$① \begin{cases} \pi^{\mathsf{s1}}(\cdot) = \mathsf{Uniform}(\mathcal{Y}) \ , \\ \pi^{\mathsf{s2}}(\cdot) = \mathsf{Uniform}(\mathcal{Y}) \ , \end{cases} \quad ② \begin{cases} \pi^{\mathsf{s1}}(\cdot) \propto \mathsf{Uniform}(\mathcal{Y}) \cdot (\pi(\cdot)/\pi_{\mathsf{ref}}(\cdot))^{\beta} \\ \pi^{\mathsf{s2}}(\cdot) \propto \mathsf{Uniform}(\mathcal{Y}) \cdot (\pi_{\mathsf{ref}}(\cdot)/\pi(\cdot))^{\beta} \end{cases}$$

$$\left|\delta(y, y'; \theta^{(T)})\right| \leqslant 0.611^{2^T - 1} \ , \ \forall y, y' \in \mathcal{Y}$$

quadratic convergence!

Regime 1: Known Reward

Not practical, only for proof of idea

$$① \begin{cases} \pi^{\mathrm{s1}}(\cdot) = \mathrm{Uniform}(\mathcal{Y}) , \\ \pi^{\mathrm{s2}}(\cdot) = \mathrm{Uniform}(\mathcal{Y}) , \end{cases} \quad ② \begin{cases} \pi^{\mathrm{s1}}(\cdot) \propto \mathrm{Uniform}(\mathcal{Y}) \cdot \exp(r(\cdot)) , \\ \pi^{\mathrm{s2}}(\cdot) \propto \mathrm{Uniform}(\mathcal{Y}) \cdot \exp(-r(\cdot)) , \end{cases}$$

- Sampling coefficient $\alpha_1 = |\mathcal{Y}|^2, \alpha_2 = \sum_{y,y'} \exp\big(r(y) - r(y')\big)$
- Upper bound

Quadratic convergence!

$$\left|\delta(y, y'; \theta^{(T)})\right| \leqslant 0.5^{2^T - 1} , \ \forall y, y' \in \mathcal{Y}$$

Intuition

$$\propto \delta(y, y''; \theta^{(t)}) - \delta(y', y''; \theta^{(t)}) + \mathcal{O}(\delta^2) = \delta(y, y'; \theta^{(t)}) + \mathcal{O}(\delta^2)$$

$$\delta(y, y'; \theta) := r(y) - r(y') - \beta \log \frac{\pi_\theta(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_\theta(y')}$$

We care about its convergence

- Recall update

$$\delta(y, y'; \theta^{(t+1)}) = \delta(y, y'; \theta^{(t)}) - \eta\beta \sum_{y'' \in \mathcal{Y}} \left[\pi^{\mathsf{s}}(y, y'')\Delta(y, y''; \theta^{(t)}) - \pi^{\mathsf{s}}(y', y'')\Delta(y', y'', \theta^{(t)})\right]$$

where $\Delta(y, y'; \theta) := \sigma(r(y) - r(y')) - \sigma\left(\beta \log \frac{\pi_\theta(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_\theta(y')}\right)$

- Taylor expansion at $r(y_1) - r(y_2)$ and setting $\pi^s(y_1, y_2) \propto 1/\sigma'(r(y_1) - r(y_2))$ gives

$$\pi^{\mathsf{s}}(y, y'')\Delta(y, y''; \theta^{(t)}) - \pi^{\mathsf{s}}(y', y'')\Delta(y', y''; \theta^{(t)}) = \textbf{constant} \cdot \delta(y, y'; \theta^{(t)}) + \textbf{quadratic term}$$

Regime 1: Known Reward

- The choice of η eliminates the linear term:

$$\delta(a, a'; \theta^{(t+1)}) = (1 - \eta\beta^2 A)\delta(a, a'; \theta^{(t)}) + \frac{\eta\beta^2}{2} \sum_{a''} \left(\frac{\sigma''(\xi_{\mathsf{R}}(a, a''; \theta^{(t)}))}{\sigma'(r(a) - r(a''))} \delta(a, a''; \theta^{(t)})^2 - \frac{\sigma''(\xi_{\mathsf{R}}(a', a''; \theta^{(t)}))}{\sigma'(r(a') - r(a''))} \delta(a', a''; \theta^{(t)})^2 \right)$$

- Bounding $\sigma'' \leq \frac{1}{6\sqrt{3}} < 0.097$ and $\sigma' \geq \sigma'(1) > 0.196$ gives

$$\left|\delta\left(y, y'; \theta^{(t+1)}\right)\right| < 0.5 \max_{a,a'} \delta\left(a, a'; \theta^{(t)}\right)^2$$

Regime 2: Online Sampler

Current policy

$$① \begin{cases} \pi^{s1}(\cdot) = \mathrm{Uniform}(\mathcal{Y}) \ , \\ \pi^{s2}(\cdot) = \mathrm{Uniform}(\mathcal{Y}) \ , \end{cases} \quad ② \begin{cases} \pi^{s1}(\cdot) \propto \mathrm{Uniform}(\mathcal{Y}) \cdot (\pi(\cdot)/\pi_{\mathrm{ref}}(\cdot))^{\beta} \\ \pi^{s2}(\cdot) \propto \mathrm{Uniform}(\mathcal{Y}) \cdot (\pi_{\mathrm{ref}}(\cdot)/\pi(\cdot))^{\beta} \end{cases}$$

- ② equivalent to $\pi^{s1} \propto \exp\big(\beta(\theta - \theta_{\mathrm{ref}})\big), \pi^{s2} \propto \exp\big(\beta(\theta_{\mathrm{ref}} - \theta)\big)$
- Sampling coefficient $\alpha_1 = |\mathcal{Y}|^2, \alpha_2 = \sum_{y,y'} \left(\frac{\pi(y)\pi_{\mathrm{ref}}(y')}{\pi_{\mathrm{ref}}(y)\pi(y')}\right)^{\beta}$
- Upper bound

Quadratic convergence!

$$\left|\delta(y, y'; \theta^{(T)})\right| \leqslant 0.611^{2^T - 1} \ , \ \forall y, y' \in \mathcal{Y}$$

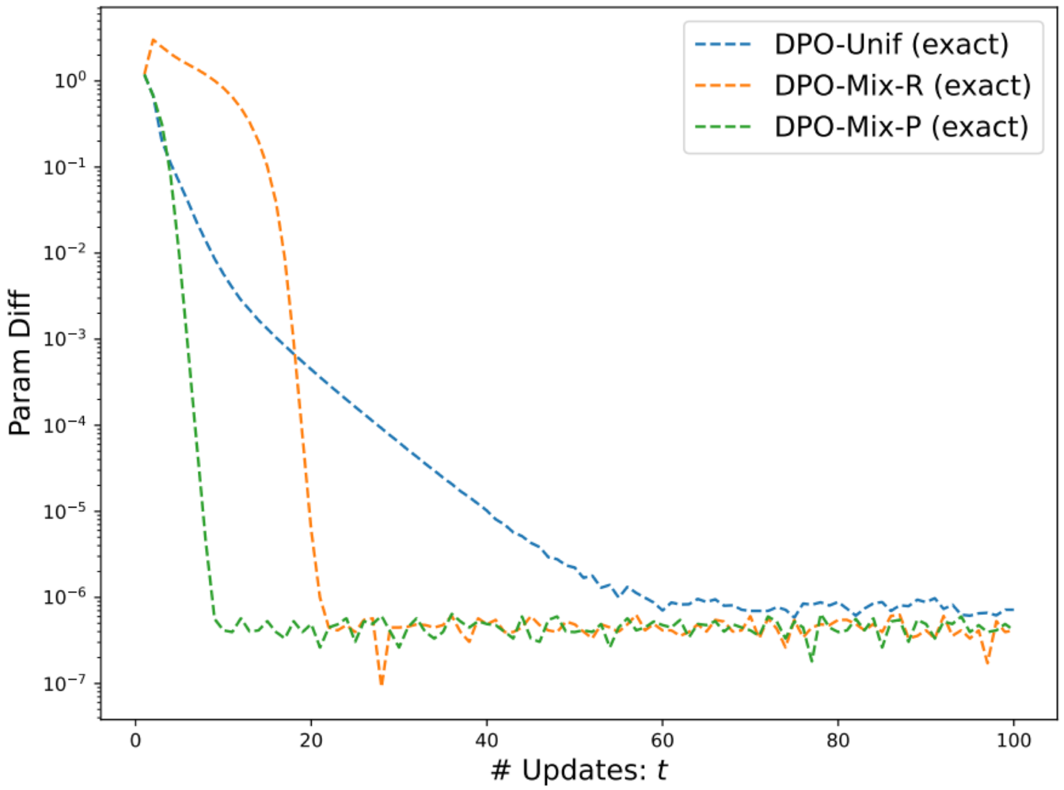
Regime 2: Online Sampler

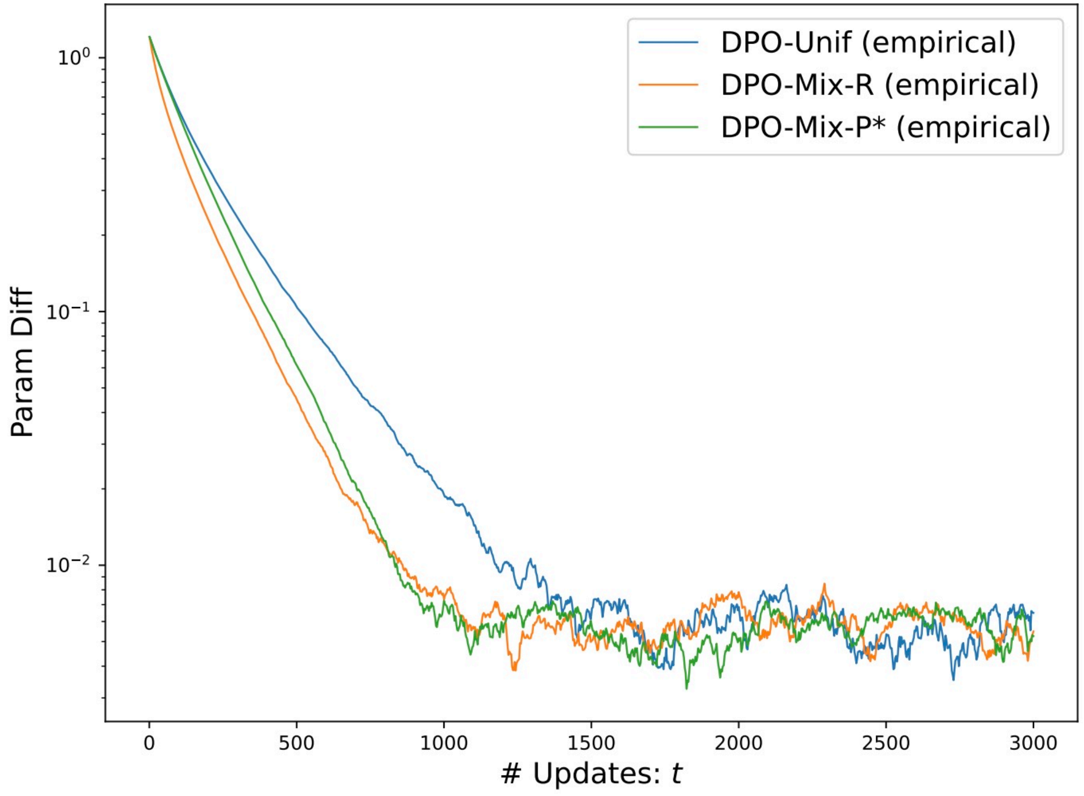
- Taylor expansion at $\beta \log \frac{\pi(y)\pi_{\mathrm{ref}}(y')}{\pi_{\mathrm{ref}}(y)\pi(y')}$

$$\delta(a,a';\theta^{(t+1)}) = (1-\eta\beta^2 A)\delta(a,a';\theta^{(t)}) - \frac{\eta\beta^2}{2}\sum_{a''}\left(\frac{\sigma''(\xi_{\mathsf{P}}(a,a'';\theta^{(t)}))}{\sigma'(\beta(\theta_a-\theta_{a''})^{(t)})}\delta(a,a'';\theta^{(t)})^2 - \frac{\sigma''(\xi_{\mathsf{P}}(a',a'';\theta^{(t)}))}{\sigma'(\beta(\theta_{a'}-\theta_{a''})^{(t)})}\delta(a',a'';\theta^{(t)})^2\right)$$

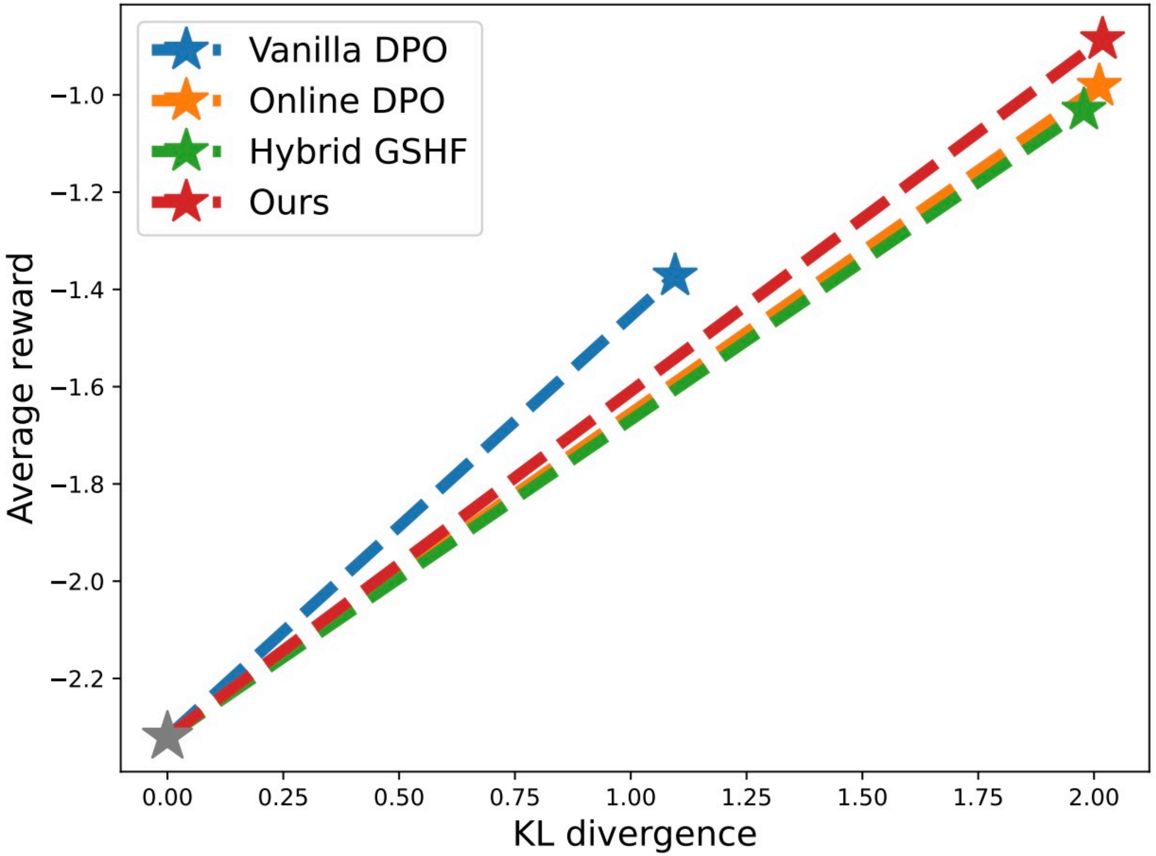
- Then same as regime 1.

Numerical Simulations

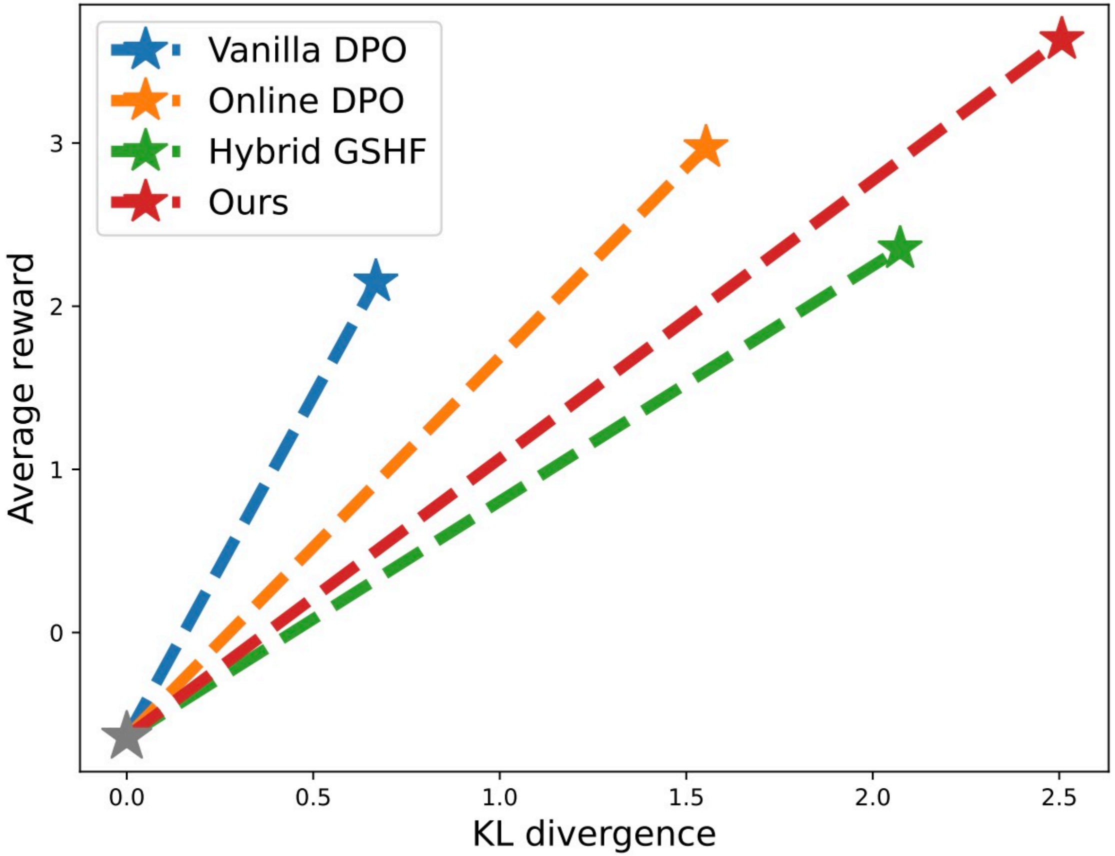




LM Implementation



Safe-RLHF



Iterative Prompt

Thank You