Quadratic Convergence of Online Direct Preference Optimization

Runlong Zhou

Oct 15 2024

WAIL-POD Presentation

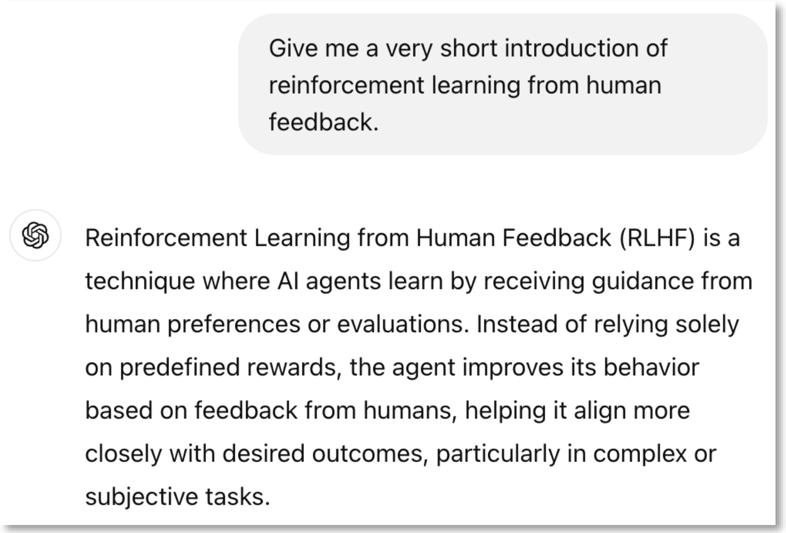
Acknowledgement

This is a joint work with

Ruizhe Shi and **Simon Du**

Language Generation

- **Prompt (State)**: user queries
- **Response (Action)**: language model generation result



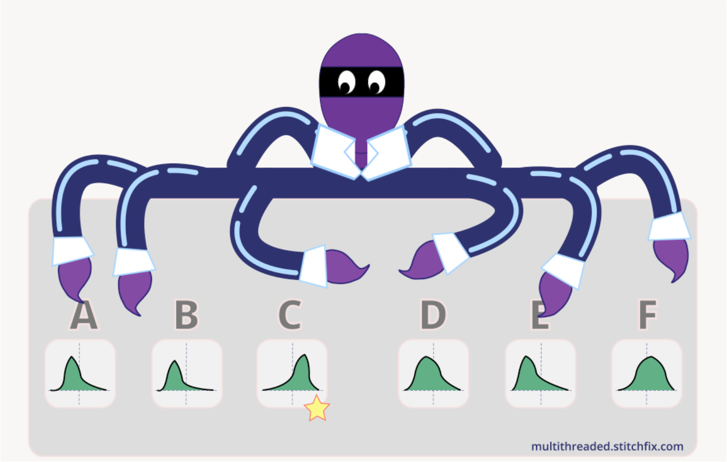
Bandits

Multi-armed bandits (MABs)

- **Arm** space $\mathcal{Y}$
- **Reward** function $r(y) \in [0,1]$

Contextual bandits (CBs)

- **Context (Prompt)** space $\mathcal{X}$
- **Arm (Response)** space $\mathcal{Y}$
- **Reward** function $r(x, y) \in [0,1]$



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

Results in this work can be easily adapted to CBs, so we focus on MABs only

Policy

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

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

Reward-based v.s. Preference-based RL

=MABs in this work

Reward-based RL
After choosing an arm y, observe a sample $r \sim R(y)$ with mean $r(y)$

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 $p \sim \text{Bernoulli}\big(p^\star(y_1 \succ y_2)\big)$

Bradley-Terry (BT) Model

- Sigmoid function

$$\sigma(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

- BT preference model

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

RL from Human Feedback (RLHF)

- 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 $y_l^{(i)}$
- **Step 1**: Learn reward function by minimizing negative log-likelihood

$$\mathcal{L}_r(\phi) = -\frac{1}{N}\sum_{i=1}^{N} \log \sigma\left(r_\phi\left(y_w^{(i)}\right) - r_\phi\left(y_l^{(i)}\right)\right)$$

RL from Human Feedback (RLHF)

- **Step 2**: Learn policy by maximizing regularized value using proximal policy optimization (PPO)

$$\theta^\star_\phi = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{y\sim\pi_\theta}[r_\phi(y)] - \beta \mathrm{KL}(\pi_\theta || \pi_{\text{ref}})$$

This step is usually slow and unstable

Direct Preference Optimization (DPO)

- Under **tabular softmax parametrization**

$$\pi_\phi^\star = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{y\sim\pi}[r_\phi(y)] - \beta \mathrm{KL}(\pi || \pi_{\mathrm{ref}})$$

is equivalent to

$$\pi_\phi^\star(y) = \frac{1}{Z_\phi} \pi_{\mathrm{ref}}(y) \mathrm{e}^{r_\phi(y)/\beta}$$

where Z is the normalizing factor

Direct Preference Optimization (DPO)

- For any y,

$$r_\phi(y) = \beta\left(\log Z_\phi + \log\frac{\pi_\phi^\star(y)}{\pi_{\text{ref}}(y)}\right)$$

- Plug into reward loss and Z_ϕ cancels out!

$$\mathcal{L}_\pi(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w^{(i)}\right)}{\pi_{\text{ref}}\left(y_w^{(i)}\right)} - \log\frac{\pi_\theta\left(y_l^{(i)}\right)}{\pi_{\text{ref}}\left(y_l^{(i)}\right)}\right)$$

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

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

Focus of Study

- Recall that

$$r(y) = \beta\left(\log Z + \log\frac{\pi^\star(y)}{\pi_{\text{ref}}(y)}\right)$$

- We want to ask

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*** $\forall y, y' \in \mathcal{Y}$***?***

Results of Exact DPO

- Regime 1: Uniform Sampler
- Regime 2: Known Reward
- Regime 3: Online Sampler

Regime 1: Uniform Sampler

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

- Sampling coefficient $\alpha = 2|\mathcal{Y}|^2$
- Initialize $\pi_{\theta^{(0)}} = \pi_{\text{ref}}$
- Learning rate $\eta = \frac{1}{\beta^2|\mathcal{Y}|}$

Will be used in all regimes

- Upper bound

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

- Directly using convexity gives an $O\left(\frac{1}{T}\right)$ rate

Regime 1: Uniform Sampler

- Define and recall that

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

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

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

- Computing the gradient gives

$$\nabla_\theta \mathcal{L}(\theta) = -\beta \sum_{y,y'} \pi^{\text{s}}(y, y')\Delta(y, y'; \theta)\mathbb{1}_y$$

Holds for all regimes

Regime 1: Uniform Sampler

- Iteration equation for δ:

Holds for all regimes

$$\delta(y, y'; \theta^{(t+1)}) = \delta(y, y'; \theta^{(t)}) - \eta\beta\alpha(\pi^{\mathsf{s1}}, \pi^{\mathsf{s2}}) \sum_{y''} \Big(\pi^{\mathsf{s}}(y, y'')\Delta(y, y''; \theta^{(t)}) - \pi^{\mathsf{s}}(y', y'')\Delta(y', y''; \theta^{(t)}) \Big)$$

- Plug in $\pi^s(y, y') = 2/\,|\mathcal{Y}|^2$ makes coefficients of Δ identical
- Use $\sigma'_{\min} \le \frac{\sigma(x)-\sigma(y)}{x-y} \le \frac{1}{4}$ to convert Δ into δ by assuming that

$$\sigma'\left(\log \frac{\pi_\theta(y)\pi_{\mathsf{ref}}(y')}{\pi_{\mathsf{ref}}(y)\pi_\theta(y')}\right) \geqslant \sigma'_{\min} > \frac{1}{8}$$

Regime 1: Uniform Sampler

- We have that

$$\gamma = \max\{1 - 4\eta\beta^2 A\sigma'_{\min}, \eta\beta^2 A - 1\} + \eta\beta^2 A(1 - 4\sigma'_{\min})$$

$$\left|\delta\left(y_1, y_2; \theta^{(t+1)}\right)\right| \leq \gamma \max_{y,y'}\left|\delta\left(y, y'; \theta^{(t)}\right)\right|$$

- Plug in η gives $\gamma < 1$
- Go back and verify the assumption on $\sigma'_{\min}$ and further refine γ

Regime 2: Known Reward

Not practical, only for proof of idea

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

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

Quadratic convergence!

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

Regime 2: Known Reward

- Taylor expansion at $r(y_1) - r(y_2)$:

$$\Delta(y_1, y_2; \theta^{(t)}) = \sigma'(r(y_1) - r(y_2))\delta(y_1, y_2; \theta^{(t)}) + \frac{\sigma''(\xi_{\mathrm{R}})}{2}\delta(y_1, y_2; \theta^{(t)})^2$$

- Recall update

$$\delta(y, y'; \theta^{(t+1)}) = \delta(y, y'; \theta^{(t)}) - \eta\beta\alpha(\pi^{\mathsf{s1}}, \pi^{\mathsf{s2}}) \sum_{y''} \Big(\pi^{\mathsf{s}}(y, y'')\Delta(y, y''; \theta^{(t)}) - \pi^{\mathsf{s}}(y', y'')\Delta(y', y''; \theta^{(t)})\Big)$$

- Setting $\pi^s(y_1, y_2) \propto 1/\sigma'\big(r(y_1) - r(y_2)\big)$ 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 2: 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 3: Online Sampler

Current policy

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

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

Quadratic convergence!

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

Regime 3: Online Sampler

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

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

- Like Regime 1, assume $\sigma'\left(\beta(\theta_a - \theta_{a'})\right) \geq \sigma'_{\min}$ and verify in the end

Empirical DPO

- (For **Regime 2**) Same equation:

$$\mathbb{E}[(G_a - G_{a'})^{(t)}] = -\beta A \delta(a, a'; \theta^{(t)}) - \frac{\beta}{2} \underbrace{\sum_{a''} \left(\frac{\sigma''(\xi_{\mathsf{R}}(a, a''; \theta^{(t)}))}{\sigma'(r(a) - r(a''))} \delta(a, a''; \theta^{(t)})^2 - \frac{\sigma''(\xi_{\mathsf{R}}(a', a''; \theta^{(t)}))}{\sigma'(r(a') - r(a''))} \delta(a', a''; \theta^{(t)})^2 \right)}_{=: N_t(a, a')}$$

- When operating under expectation:
  - $\mathbb{E}[\delta(;\theta^{(t+1)})]$ needs $\mathbb{E}\left[\delta(;\theta^{(t)})^2\right]$
  - $\mathbb{E}\left[\delta(;\theta^{(t)})^2\right]$ needs $\mathbb{E}\left[\delta(;\theta^{(t-1)})^4\right]$
  - …
  - $\mathbb{E}[\delta(;\theta^{(T)})]$ needs $\mathbb{E}\left[\delta(;\theta^{(t)})^n\right]$ for any t, n such that $2^t \cdot n \leq 2^T$

Bounding Moments

- With some manipulation, we have

Noise

$$\mathbb{E}[\delta(a,a';\theta^{(t+1)})^{2n}] \leqslant \sum_{k=0}^{2n} \binom{2n}{k} (6\sigma\sqrt{n})^k \cdot \frac{1}{2^{2n-k}} \max_{a_1,a_2} \mathbb{E}[\delta(a_1,a_2;\theta^{(t)})^{4n-2k}]$$

- Take $T = \log 1/\sigma$, then with sufficiently small σ and any $2^t \cdot n \leq 2^T$,

$$\mathbb{E}[\delta(a,a';\theta^{(t)})^{2n}] \leqslant \left(12\sqrt{n}\sigma + \frac{1}{2^t}\right)^{2n}$$

- This implies

$$\sqrt{\mathbb{E}\left[\delta(y,y';\theta^{(T)})^2\right]} \leqslant 14\sigma \;,\; \forall y,y' \in \mathcal{Y}$$

Regime 3?

- $\sigma'\left(\beta(\theta_a - \theta_{a'})\right)$ hard to bound under estimation scheme
- If we use Taylor expansion at any point $z(a, a')$:

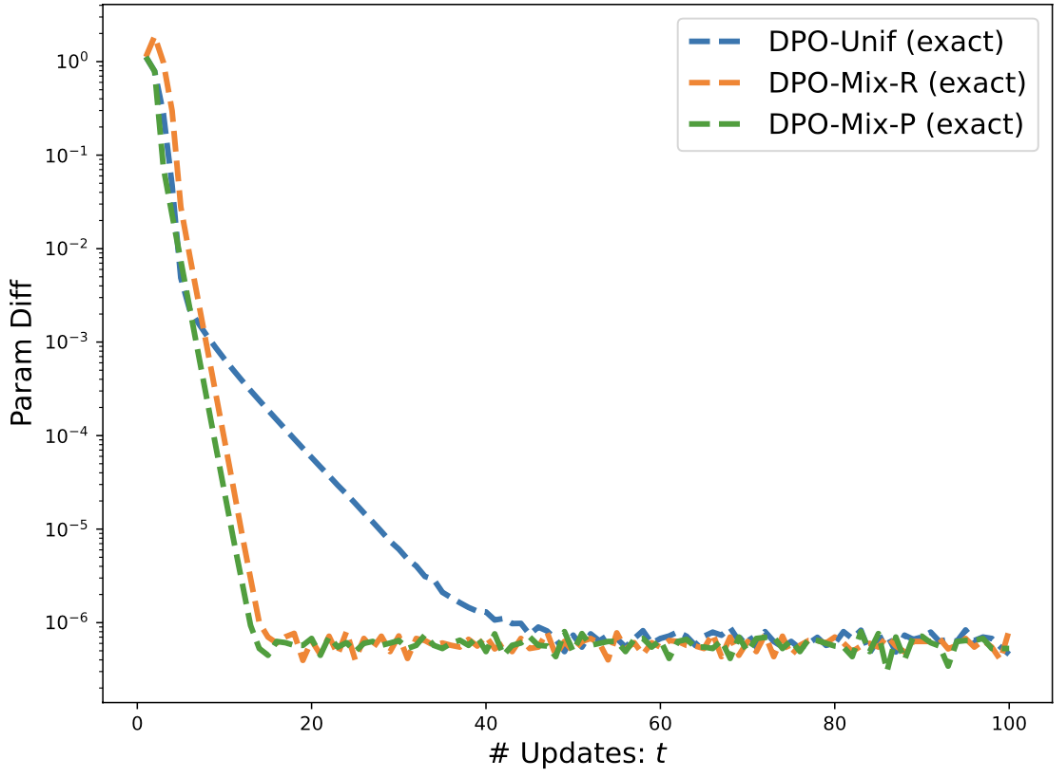
$$\Delta(a, a'; \theta) = \sigma'(z(a, a'))\delta(a, a'; \theta) + \frac{\sigma''(\xi_1(a, a'; \theta))}{2}(r(a) - r(a') - z(a, a'))^2$$
$$- \frac{\sigma''(\xi_2(a, a'; \theta))}{2}[\beta(\theta_a - \theta_{a'}) - z(a, a')]^2 ,$$

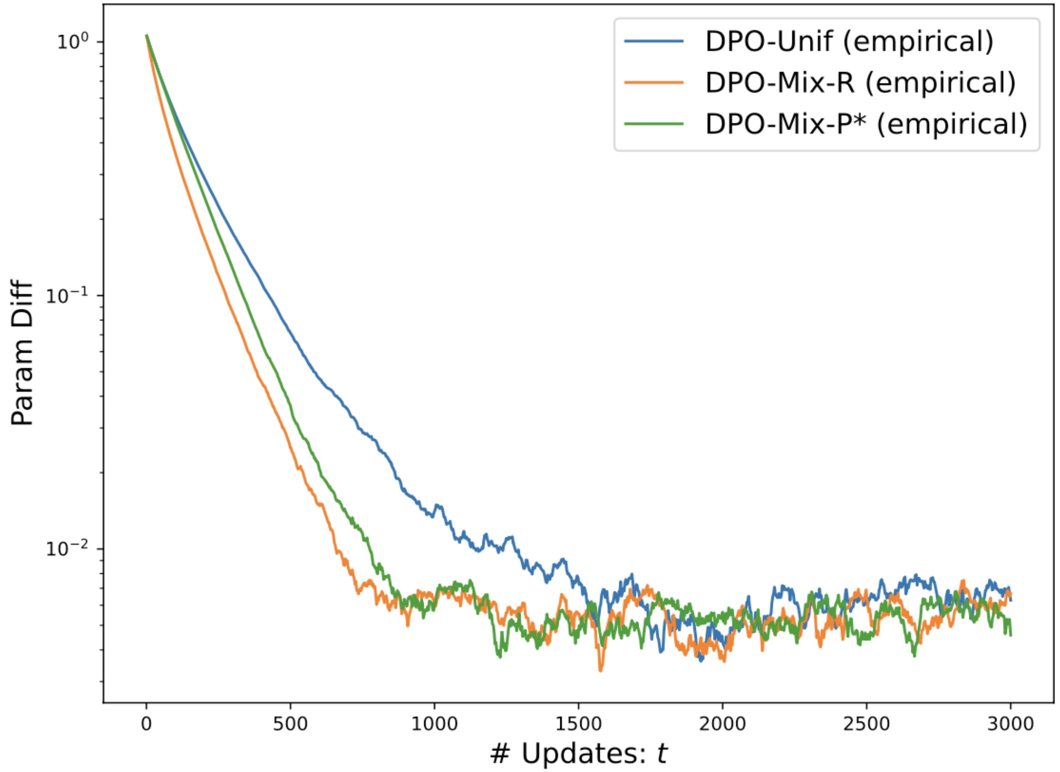
- Set $\pi^s(y_1, y_2) \propto 1/\sigma'(z(y_1, y_2))$, try to make
  - $\sigma'\left(z(y_1, y_2)\right)$ bounded
  - $[r(a) - r(a') - z(a, a')]^2 + [\beta(\theta_a - \theta_{a'}) - z(a, a')]^2$ not far from δ^2

Regime 3?

- Take $z(y_1, y_2) = \mathrm{clip}\left(\beta\left(\theta_{y_1} - \theta_{y_2}\right), [-1,1]\right)$
- Algorithm changes accordingly with a rejection sampling step
- Proof reduces to Regime 2, results are the same
- Can be applied to the exact gradient case for a faster convergence

Numerical Simulations





Safe-RLHF

Algorithm	Iters	Average reward (train)	Win-rate (train)	Average reward (test)	Win-rate (test)
Vanilla DPO	2	-1.486	67.6%	-1.423	68.7%
	3	-1.144	72.5%	-1.203	71.7%
On-policy DPO	2	-1.478	67.6%	-1.510	65.8%
	3	-1.082	73.2%	-1.094	73.2%
Hybrid GSHF	2	-1.517	68.5%	-1.505	66.9%
	3	-1.079	74.8%	-1.002	75.9%
Ours	2	-1.457	68.1%	-1.436	67.6%
	3	-0.908	75.6%	-0.945	76.2%

Iterative-Prompt

Algorithm	Iters	Average reward (train)	Win-rate (train)	Average reward (test)	Win-rate (test)
Vanilla DPO	2	1.427	71.4%	1.375	70.0%
	3	2.023	78.4%	2.133	78.8%
On-policy DPO	2	2.106	79.2%	2.157	78.7%
	3	3.131	82.4%	3.327	82.9%
Hybrid GSHF	2	2.116	79.6%	2.224	80.0%
	3	2.386	81.9%	2.500	82.8%
Ours	2	2.026	78.3%	2.068	77.3%
	3	4.149	86.6%	4.221	87.1%

Thank You