The Crucial Role of Samplers in Online Direct Preference Optimization

Ruizhe Shi*   Runlong Zhou*   Simon S. Du

* *indicates equal contribution.*

清華大学 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

Multi-armed bandits (MABs)

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

Results can be easily adapted to contextual bandits, so we focus on MABs only

Preference-based RL

- After choosing a **pair** (y_1, y_2), we observe a sample $p \sim \text{Bernoulli}\left(p^\star(y_1 \succ y_2)\right)$ **(Preference model)**
- **Bradley-Terry** model:

Sigmoid function

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

Tabular softmax parameterization

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

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

DPO loss:

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

Closed-form solution:

$$\pi^\star(y) = \frac{1}{Z}\pi_{\text{ref}}(y)e^{r(y)/\beta}$$

Question we study:

*How **fast** can **DPO w. different sampling strategies** converge to the **closed-form solution**?*

For sampling, here we mean how we sample (y_1, y_2).

Formulation

π^{s1} for y_1 and π^{s2} for y_2. Joint probability

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

Stop gradient

Exact DPO loss and policy update:

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

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

Mixture of samplers:

Sampling coefficients

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

Convergence quantities:

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

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

How fast can $\delta(y, y'; \theta^{(t)})$ converge to 0?

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

Taylor expansion: $\Delta \to \delta$

Uniform Sampler (**Unif**)

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

Reward-guided Sampler (**Mix-R**)

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

Policy-guided Sampler (**Mix-P**)

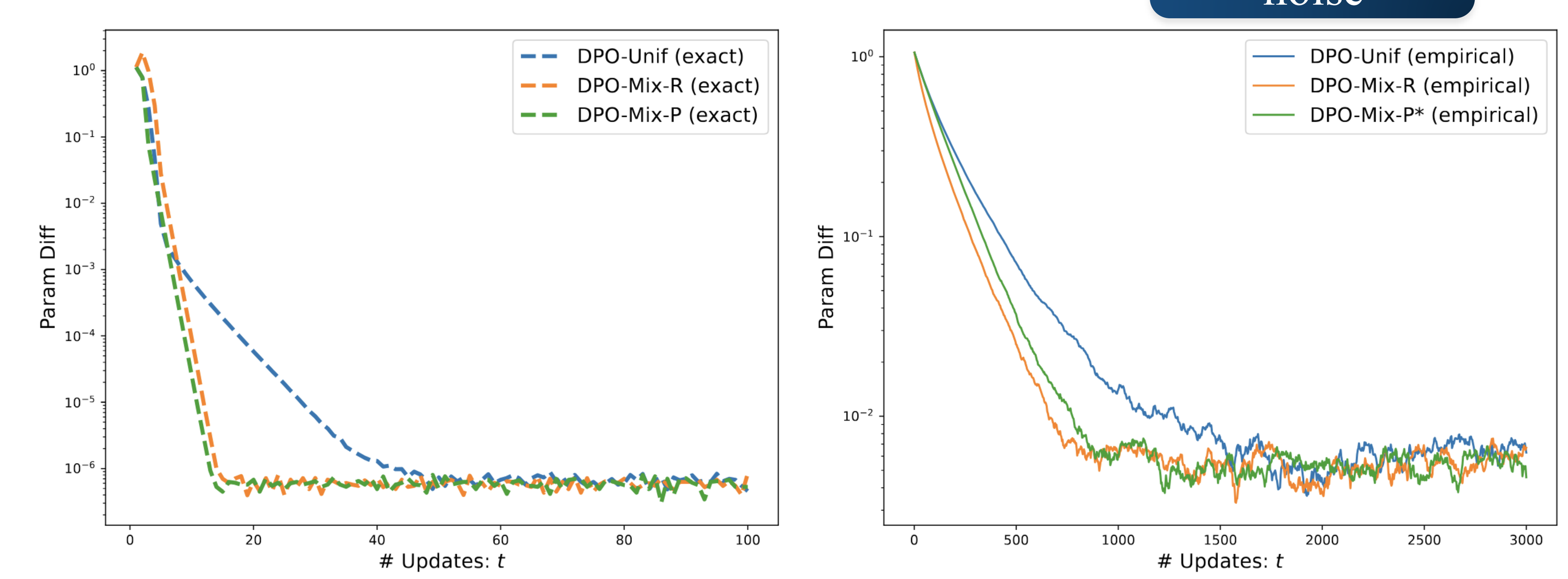
$$① \begin{cases} \pi^{s1}(\cdot) = \text{Uniform}(\mathcal{Y}), \\ \pi^{s2}(\cdot) = \text{Uniform}(\mathcal{Y}), \end{cases} \quad ② \begin{cases} \pi^{s1}(\cdot) \propto \text{Uniform}(\mathcal{Y}) \cdot (\pi(\cdot)/\pi_{\text{ref}}(\cdot))^\beta \\ \pi^{s2}(\cdot) \propto \text{Uniform}(\mathcal{Y})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

Convergence rate:

	Unif	Mix-R	Mix-P
Exact	0.588^T	0.5^{2^T-1}	0.611^{2^T-1}
Empirical	unknown	linear to $O(\sigma)$	linear to $O(\sigma)$

Variance of noise

Bandit simulation:



Benchmarks:

Safe-RLHF

Algorithm	Iters	Reward (train)	Win-rate (train)	Reward (test)	Win-rate (GPT4o-mini)
Vanilla DPO	2	-1.438(±0.092)	68.1(±0.8)%	-1.391(±0.076)	-
	3	-1.238(±0.085)	71.3(±1.1)%	-1.242(±0.045)	71.5%
On-policy DPO	2	-1.328(±0.258)	69.4(±3.1)%	-1.362(±0.235)	-
	3	-1.003(±0.118)	74.2(±1.3)%	-1.004(±0.100)	73.0%
Hybrid GSHF	2	-1.349(±0.295)	70.5(±3.2)%	-1.335(±0.302)	-
	3	-1.007(±0.149)	75.2(±0.9)%	-0.946(±0.138)	81.0%
Ours	2	-1.323(±0.242)	69.8(±2.6)%	-1.295(±0.226)	-
	3	**-0.894**(±0.043)	**75.6**(±0.3)%	**-0.923**(±0.086)	**82.5%**

Iterative-Prompt

Algorithm	Iters	Reward (train)	Win-rate (train)	Reward (test)	Win-rate (GPT4o-mini)
Vanilla DPO	2	1.460(±0.035)	71.5(±0.2)%	1.418(±0.038)	-
	3	2.146(±0.108)	79.6(±1.0)%	2.166(±0.042)	76.5%
On-policy DPO	2	2.135(±0.029)	78.8(±0.5)%	2.132(±0.023)	-
	3	3.712(±0.507)	85.1(±2.4)%	3.704(±0.331)	88.0%
Hybrid GSHF	2	2.138(±0.020)	79.4(±0.2)%	2.136(±0.077)	-
	3	2.481(±0.088)	81.7(±0.8)%	2.497(±0.052)	80.0%
Ours	2	2.060(±0.030)	78.2(±0.2)%	2.067(±0.008)	-
	3	**4.249**(±0.365)	**87.1**(±3.0)%	**4.248**(±0.388)	**89.5%**

Check out our paper to see how to implement these regimes in practical DPO!