

THE CRUCIAL ROLE OF SAMPLERS IN ONLINE DIRECT PREFERENCE OPTIMIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) has emerged as a stable, scalable, and efficient solution for language model alignment. Despite its empirical success, the *optimization* properties, particularly the impact of samplers on its convergence rates, remain underexplored. In this paper, we provide a rigorous analysis of DPO’s *convergence rates* with different sampling strategies under the exact gradient setting, revealing a surprising separation: uniform sampling achieves *linear* convergence, while our proposed online sampler achieves *quadratic* convergence. We further adapt the sampler to practical settings by incorporating posterior distributions and *logit mixing*, demonstrating significant improvements over previous approaches. On Safe-RLHF dataset, our method exhibits a 9.5% improvement over vanilla DPO and on-policy DPO; on Iterative-Prompt, our approach outperforms vanilla DPO and hybrid GSHF by over 9.5%. Our results not only offer insights into the theoretical standing of DPO but also pave the way for potential algorithm designs in the future.

1 INTRODUCTION

Aligning language models (LMs) to human preferences is a critical pursuit due to its great potentials to push forward artificial intelligence (AI) development, and to enable AI to serve humanity better (Ji et al., 2023b). Reinforcement learning from human feedback (RLHF) (Ziegler et al., 2019; Bai et al., 2022) has been a widely-used approach, gaining tremendous successes in aligning LMs (OpenAI, 2024). However, the multi-stage pipeline of RLHF, including reward model training and RL tuning, is sensitive to hyperparameters and costly to train. DPO (Rafailov et al., 2023) directly combines these stages and tunes LMs in an offline way, gaining popularity due to its stability and efficiency.

The empirical success of DPO has recently sparked a significant increase in interest for understanding its theoretical properties. Through modeling RLHF as a KL-regularized contextual bandit problem or Markov decision process, many works (Xiong et al., 2024; Xie et al., 2024; Liu et al., 2024b; Khaki et al., 2024; Song et al., 2024) obtain strong theoretical results and highlight the role of samplers in DPO. Specifically, they point out drawbacks of the offline sampler in vanilla DPO, and propose on-policy sampler or other samplers as better choices, as validated empirically (Dong et al., 2024; Guo et al., 2024; Tajwar et al., 2024).

However, these theoretical explanations are largely built upon traditional RL and analyze the impact of samplers from the view of data, namely sample complexity, thus involving some *impractical* assumptions, such as the access to an oracle for maximum likelihood estimation (MLE). Meanwhile, from the *optimization* perspective, the *convergence rates* of gradient descent in DPO within different sampling regimes remain an underexplored question. A particular setting of our interest is to give provable guarantees for an *online* sampler depending on the current policy.

1.1 CONTRIBUTIONS

To fill this research gap, we focus on analyzing the crucial role of samplers in DPO, from the view of optimization. Based on our theoretical findings, we can further derive a new effective approach, demonstrating advantages in empirical experiments over previous approaches. We summarize our contributions as follows:

- **Theoretical separations.** We analyze the convergence rates of DPO with various samplers under *tabular softmax parametrization*, and demonstrate theoretical advantages brought by specific samplers. Specifically, we show a separation that our proposed samplers, DPO-Mix-R and DPO-Mix-P , achieve quadratic convergence rates, while the commonly used one, DPO-Unif , can only achieve a linear convergence rate. Numerical simulations support our results. See Section 4.
- **Practical improvements.** We design a new sampler for practical DPO. Specifically, we employ *logit mixing* to align sampling distribution to our theory. LM alignment experiments show that under the same computation budget, our method demonstrates significant advantages over baselines. On Safe-RLHF dataset, our method exhibits an over 9.5% improvement over vanilla DPO and on-policy DPO. On Iterative-Prompt dataset, our method shows an 9.5% improvement compared to vanilla DPO and hybrid GSHF. See Section 5.
- **Explainability and generalizability.** We show that our theoretical framework can explain many existing DPO variants and thus provides a new perspective on their theoretical advantages, demonstrating the generalizability of this work and its potential for designing more powerful algorithms. We study the existing DPO variants, and find that after setting a posterior distribution on the response set, vanilla DPO and on-policy DPO can both be mapped to DPO-Unif , while Hybrid GSHF and Online GSHF (Xiong et al., 2024) can be viewed as approximations to DPO-Mix-P and DPO-Mix-R , respectively. This further validates the soundness of our theoretical findings. See Section 5.

2 RELATED WORK

Theoretical study of RLHF / DPO. Zhu et al. (2023) formulate RLHF as contextual bandits, and prove the convergence of the maximum likelihood estimator. Xiong et al. (2024) further consider KL-regularization and show the benefits in sample complexity of online exploration in DPO. Xie et al. (2024) study the online exploration problem from the perspective of KL-regularized Markov decision processes, and show provable guarantees in sample complexity of a exploration bonus. Liu et al. (2024c) investigate the overoptimization issue, and prove a finite-sample suboptimality gap. Song et al. (2024) show a separation of coverage conditions for offline DPO and online RLHF. These works primarily focus on the perspective of data, which is widely adopted in RL literature. For Xiong et al. (2024); Xie et al. (2024), their policy update iteration is to directly solve MLE instead of doing *gradient descent* as ours. Song et al. (2024) focus on data coverage, and have not studied the convergence rates. In contrast, this paper analyzes DPO from the perspective of *optimization*, offering a complementary while more practical viewpoint.

Variants of DPO. There are two line of works exploring the variants of DPO. 1) **Objective function.** Ψ -PO (Azar et al., 2023) changes the reward term to alternate mappings from preference pairs. RPO (Liu et al., 2024c) adds an imitation loss to mitigate the overoptimization issue. CPO (Xu et al., 2024) removes the π_{ref} term and adds an imitation loss to ensure that the policy does not deviate too much. SimPO (Meng et al., 2024) also removes the π_{ref} term for efficiency, while using length normalization for better length control. 2) **Sampler.** Liu et al. (2024b); Khaki et al. (2024) utilize rejection sampling to adjust the data distribution to the theoretically-optimal policy before training. On-policy DPO (Guo et al., 2024; Tajwar et al., 2024; Ding et al., 2024) emphasize the importance of the on-policy sampler. Iterative DPO (Xiong et al., 2024; Dong et al., 2024) introduces an iterative training scheme, where an online policy is used to generate data pairs, annotated by a gold reward model, and the DPO training is subsequently applied to update the policy. XPO (Xie et al., 2024) follows the setting of iterative DPO, and adds an optimistic term to the DPO objective. In this paper, we focus on the latter direction, and only study the original objective.

Other RLHF approaches. There is also a line of works (Munos et al., 2023; Swamy et al., 2024; Rosset et al., 2024; Zhang et al., 2024) studying RLHF from a game-theoretic perspective. Nash-MD-PG in Munos et al. (2023) uses a geometric mixture of online policy and reference policy without specifying the mixing weight. Rosset et al. (2024) re-formulates the DPO pipeline and shows theoretical guarantees for the on-policy sampler with an MLE oracle.

Convergence analysis of policy gradient methods. Mei et al. (2020); Agarwal et al. (2021); Mei et al. (2023) study the convergence rates of policy gradient methods under *tabular softmax*

108 *parametrization*, and their results have successfully shown the practical impact of analysis with exact
 109 gradient. Motivated by them, in this paper we first study DPO with access to the exact gradient.
 110

111 3 PRELIMINARIES

112 **Notations.** Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be the sigmoid function, where $\sigma(x) = 1/(1 + \exp(-x))$. For any set X ,
 113 $\Delta(X)$ represents the set of probability distributions over X . $\text{sg}(\cdot)$ is the stopping-gradient operator.
 114 Let $\mathbb{1}_k$ be a vector with 1 on the dimension corresponding to k and 0 on others (the dimension of
 115 this vector is implicitly defined from the context).
 116
 117

118 3.1 STANDARD BANDIT LEARNING

119 Firstly, we give basic concepts of standard bandit learning, which found the basis for RLHF.
 120

121 **Multi-armed bandits and contextual bandits.** A multi-armed bandit has an arm (action) space \mathcal{Y}
 122 and a reward function $r : \mathcal{Y} \rightarrow [0, 1]$. A contextual bandit has a context space \mathcal{X} , an arm space \mathcal{Y} ,
 123 and a reward function $r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$. In this work, the user prompt is viewed as a context, and
 124 the agent response is viewed as an arm. To simplify notations, our results are stated in *multi-armed*
 125 *bandits* versions. The statements and proofs can be easily extended to *contextual bandits*. **Thus, we**
 126 **will omit the prompts (contexts) and slightly abuse the notations throughout Sections 3 and 4.**
 127

128 **Policies.** A policy $\pi : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$ maps each context to a probability simplex over the arm space.
 129 For multi-armed bandits, a policy is instead a probability distribution over the arm space. We denote
 130 Π as the set of policies we study. Under *tabular softmax parametrization* which is common in
 131 previous works (Rafailov et al., 2023; Azar et al., 2023; Munos et al., 2023; Swamy et al., 2024),
 132 the policy π is parameterized by $\theta \in \mathbb{R}^{|\mathcal{Y}|}$: for any $y \in \mathcal{Y}$,

$$133 \pi_\theta(y) = \frac{\exp(\theta_y)}{\sum_{y' \in \mathcal{Y}} \exp(\theta_{y'})} .$$

134 The goal is to find the optimal policy maximizing the expected reward (with regularization).
 135
 136

137 3.2 REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)

138 Secondly, we introduce RLHF / preference-based reinforcement learning (PBRL) problem (Wirth
 139 et al., 2017; Christiano et al., 2017; Swamy et al., 2024) and current approaches.
 140

141 **Bradley-Terry (BT) model.** Given an implicit reward oracle $r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$, Bradley and Terry
 142 (1952) assume that human preference distribution $p^* : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \rightarrow \Delta(\{0, 1\})$ satisfies:

$$143 p^*(y_1 > y_2 | x) = \sigma(r(x, y_1) - r(x, y_2)) .$$

144 This means that conditioned on prompt x , response y_1 is favored over y_2 with probability $p^*(y_1 >$
 145 $y_2 | x)$ by human annotators.
 146

147 **RLHF (Ziegler et al., 2019; Bai et al., 2022).** A human preference dataset $\mathcal{D} =$
 148 $\{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$ means that in the i^{th} sample, $y_w^{(i)} > y_l^{(i)}$ conditioned on $x^{(i)}$. The reward
 149 function $r : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is learned with parameter ϕ using a negative log-likelihood loss:

$$150 \mathcal{L}_r(\phi) = -\frac{1}{N} \sum_{i=1}^N \left[\log \sigma \left(r_\phi(x^{(i)}, y_w^{(i)}) - r_\phi(x^{(i)}, y_l^{(i)}) \right) \right] . \quad (1)$$

151 Given $\pi_1, \pi_2 \in \Pi$, $\mathbb{E}_{x \sim \rho(\mathcal{X})} \text{KL}(\pi_1(\cdot | x) \| \pi_2(\cdot | x))$ where $\rho(\mathcal{X})$ is a predefined probability distribution
 152 over \mathcal{X} is abbreviated as $\text{KL}(\pi_1 \| \pi_2)$. Based on a reference policy π_{ref} , the goal of RLHF is to
 153 maximize the obtained rewards with a KL-divergence penalty:

$$154 \pi^* = \operatorname{argmax}_{\pi \in \Pi} \mathbb{E}_{x \sim \rho(\mathcal{X}), y \sim \pi(\cdot | x)} r_\phi(x, y) - \beta \text{KL}(\pi \| \pi_{\text{ref}}) , \quad (2)$$

155 where $\beta \in \mathbb{R}_+$ is the regularization coefficient. Additionally, under *tabular softmax parametrization*,
 156 we can directly write out the closed-form solution (Equation (4) in Rafailov et al. (2023)):

$$157 \pi^*(y | x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y | x) \exp \left(\frac{1}{\beta} r_\phi(x, y) \right) , \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} , \quad (3)$$

where $Z(x) = \sum_{y \in \mathcal{Y}} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r_{\phi}(x, y)\right)$ is the partition function.

Direct Preference Optimization (DPO, Rafailov et al. (2023)). DPO integrates reward learning with policy learning. Given the human preference dataset $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$, the DPO policy π is learned with parameter θ using a negative log-likelihood loss:

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

which can be directly derived by combining Equations (1) and (3).

In this paper, we look into the role of *samplers* in the performance of DPO. Now we formally define DPO with samplers, from the perspective of bandit algorithms. Motivated by Mei et al. (2020); Agarwal et al. (2021); Mei et al. (2023), we first consider the scenario where we know the exact loss function and its gradient with respect to the model parameter θ .

Definition 1 (Exact DPO). Given an action set \mathcal{Y} , two samplers $\pi^{s1}, \pi^{s2} \in \Pi$ for sampling the first and second action respectively, a human preference oracle $p^* : \mathcal{Y} \times \mathcal{Y} \rightarrow \Delta(\{0, 1\})$, and hyperparameters $\beta, \eta \in \mathbb{R}_+$, the sampling probability and DPO loss function are defined as

$$\begin{aligned} \pi^s(y, y') &:= \text{sg}(\pi^{s1}(y)\pi^{s2}(y') + \pi^{s1}(y')\pi^{s2}(y)), \\ \mathcal{L}_{\text{DPO}}(\theta) &:= - \sum_{y, y' \in \mathcal{Y}} \pi^s(y, y') p^*(y > y') \log \sigma \left(\beta \log \frac{\pi_{\theta}(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_{\theta}(y')} \right), \end{aligned} \quad (4)$$

and the parameter is updated by

$$\theta^{(t+1)} = \theta^{(t)} - \eta \alpha(\pi^{s1}, \pi^{s2}) \nabla_{\theta} \mathcal{L}_{\text{DPO}}(\theta^{(t)}), \quad (5)$$

where $\alpha(\pi^{s1}, \pi^{s2})$ is a sampling coefficient determined by the samplers.

Remark 1. 1) π^{s1} and π^{s2} can depend on the current parameter $\theta^{(t)}$, for example, in on-policy DPO (Guo et al., 2024; Tajwar et al., 2024). Since we stopped gradients on the sampling part, we will omit the $\theta^{(t)}$ in the future occurrences for simplicity.

2) The sampling coefficient α is for the purpose of comparing different sampling regimes with the same learning rate.

3) If the sampling regime is a mixture of ①: loss function \mathcal{L}_1 with sampling coefficient α_1 and ②: loss function \mathcal{L}_2 with sampling coefficient α_2 , the gradient update rule follows

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

Note that ① and ② can have different sets of π^{s1} and π^{s2} .

In empirical studies, we do not have access to the exact gradients. Thus, we define the scenario of empirical DPO and make mild assumptions on the gradient estimation.

Definition 2 (Empirical DPO). Given noise scale $\sigma \in \mathbb{R}_+$, DPO (σ) is defined as DPO with the gradient update in Equation (5) as

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

where $G_y^{(t)}$, i.e. the y -th entry of $G^{(t)}$, is a random variable s.t. for $\forall y \in \mathcal{Y}$,

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

Remark 2. If the samplers are mixed, e.g., ① and ② in Remark 1, then we assume

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

The closed form solution π^* in Equation (3) satisfies $r(y) - r(y') - \beta \log \frac{\pi^*(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi^*(y')} = 0$, which thus motivates us to study the convergence rate. With the update rule formally defined, now we ask:

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

217
218 We will study the convergence rates for three sampling regimes: one sampling uniformly on the
219 action space \mathcal{Y} and two with mixtures of samplers. They are defined in Definitions 3 to 5.

220 **Definition 3** (Uniform sampler). DPO-UNIF is defined as DPO with $\pi^{\text{s1}}, \pi^{\text{s2}}$ as

$$221 \pi^{\text{s1}}(\cdot) = \pi^{\text{s2}}(\cdot) = \text{Uniform}(\mathcal{Y}),$$

222 and $\alpha(\pi^{\text{s1}}, \pi^{\text{s2}}) = 2|\mathcal{Y}|^2$.

223 **Definition 4** (Reward-guided mixed sampler). DPO-MIX-R is defined as DPO with $\pi^{\text{s1}}, \pi^{\text{s2}}$ as

$$224 \textcircled{1} \begin{cases} \pi^{\text{s1}}(\cdot) = \text{Uniform}(\mathcal{Y}), \\ \pi^{\text{s2}}(\cdot) = \text{Uniform}(\mathcal{Y}), \end{cases} \quad \textcircled{2} \begin{cases} \pi^{\text{s1}}(\cdot) \propto \text{Uniform}(\mathcal{Y}) \cdot \exp(r(\cdot)), \\ \pi^{\text{s2}}(\cdot) \propto \text{Uniform}(\mathcal{Y}) \cdot \exp(-r(\cdot)), \end{cases}$$

225 and $\alpha_1 = |\mathcal{Y}|^2, \alpha_2 = \sum_{y, y' \in \mathcal{Y}} \exp(r(y) - r(y'))$.

226 **Remark 3.** When we know the reward, we intuitively want the win response distribution π^{s1} to
227 have a positive correlation with the reward (and vice versa for the lose response distribution π^{s2}).
228 Definition 4 does not define a practical sampler as r is unknown, but it is important to display our
229 idea of using a mixture of sampling policies.

230 **Definition 5** (Policy-difference-guided mixed sampler). DPO-MIX-P is defined as DPO with $\pi^{\text{s1}},$
231 π^{s2} as

$$232 \textcircled{1} \begin{cases} \pi^{\text{s1}}(\cdot) = \text{Uniform}(\mathcal{Y}), \\ \pi^{\text{s2}}(\cdot) = \text{Uniform}(\mathcal{Y}), \end{cases} \quad \textcircled{2} \begin{cases} \pi^{\text{s1}}(\cdot) \propto \text{Uniform}(\mathcal{Y}) \cdot (\pi_{\theta}(\cdot)/\pi_{\text{ref}}(\cdot))^{\beta}, \\ \pi^{\text{s2}}(\cdot) \propto \text{Uniform}(\mathcal{Y}) \cdot (\pi_{\text{ref}}(\cdot)/\pi_{\theta}(\cdot))^{\beta}, \end{cases}$$

233 and $\alpha_1 = |\mathcal{Y}|^2, \alpha_2 = \sum_{y, y' \in \mathcal{Y}} \left(\frac{\pi_{\theta}(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_{\theta}(y')} \right)^{\beta}$.

234 **Remark 4.** When we cannot know the reward, $\beta \log \frac{\pi_{\theta}(y)}{\pi_{\text{ref}}(y)}$ can work as a surrogate/approximation
235 of reward $r(y)$ (Rafailov et al., 2023). In Definition 5, $\textcircled{2}$ can also be written as $\pi^{\text{s1}} \propto \exp(\beta(\theta - \theta_{\text{ref}})),$
236 $\pi^{\text{s2}} \propto \exp(\beta(\theta_{\text{ref}} - \theta))$. Uniform(\mathcal{Y}) in Definitions 4 and 5 is for consistency with Section 5, where
237 we use a posterior distribution over \mathcal{Y} .

238 4 MAIN RESULTS

239 We show our main results on convergence rates in this section. In summary, our proposed mixed
240 samplers can provably achieve: 1) exponentially faster convergence rates (*quadratic v.s. linear*)
241 compared with the uniform sampler in the exact gradient setting, and 2) linear convergence rates
242 to the noise scale when we have only unbiased estimations of the gradient. Numerical simulations
243 corroborate these theories.

244 4.1 THEORETICAL FINDINGS

245 We present theories regarding convergence rates of different sampling regimes for exact DPO and
246 empirical DPO in this subsection, along with their proof sketches. We first define important nota-
247 tions:

$$248 \Delta(y, y'; \theta) := \sigma(r(y) - r(y')) - \sigma \left(\beta \log \frac{\pi_{\theta}(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_{\theta}(y')} \right),$$

$$249 \delta(y, y'; \theta) := r(y) - r(y') - \beta \log \frac{\pi_{\theta}(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_{\theta}(y')}.$$

250 Then we can obtain

$$251 \nabla_{\theta} \mathcal{L}(\theta) = -\beta \sum_{y, y'} \pi^{\text{s}}(y, y') \Delta(y, y'; \theta) \mathbb{1}_y$$

252 by plugging $p^*(y, y') = \sigma(r(y) - r(y'))$ and $\sigma(-x) = 1 - \sigma(x)$ into the derivative of Equation (4).
253 Hence, we can derive the iteration equation for δ following the update rule of gradient descent (5):

$$254 \delta(y, y'; \theta^{(t+1)}) = \delta(y, y'; \theta^{(t)})$$

$$- \eta \beta \alpha (\pi^{s1}, \pi^{s2}) \sum_{y''} \left(\pi^s(y, y'') \Delta(y, y''; \theta^{(t)}) - \pi^s(y', y'') \Delta(y', y''; \theta^{(t)}) \right). \quad (6)$$

We state the common condition for the upper bounds for simplicity:

Condition 1. Given an action set \mathcal{Y} , it satisfies $r(y) \in [0, 1], \forall y \in \mathcal{Y}$. $\pi_{\theta^{(0)}}$ is initialized as π_{ref} , and the regularization coefficient is $\beta \in \mathbb{R}_+$. Use the learning rate $\eta = \frac{1}{\beta^2 |\mathcal{Y}|}$.

4.1.1 FOR EXACT DPO

For DPO-Unif, we have that $\pi^s(y, y') = 2/|\mathcal{Y}|^2$, making the coefficients of each Δ on the RHS of Equation (6) identical by absolute values. To proceed, we claim a lower bound as $\sigma' \left(\log \frac{\pi_{\theta}(y) \pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y) \pi_{\theta}(y')} \right) \geq \sigma'_{\min}$, and use Lagrange interpolation, namely $\sigma'_{\min} \leq (\sigma(x) - \sigma(y))/(x - y) \leq 1/4$, to transform Δ into δ . By carefully computing the coefficients of each δ and picking learning rate, we arrive at a linear convergence. Using this linear convergence, we can turn back to bound σ'_{\min} , completing the proof. See detailed proof in Appendix A.1.1.

Theorem 1 (Upper bound of DPO-Unif). Under Condition 1, DPO-Unif satisfies

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

where $T \in \mathbb{N}$ is the number of iterations.

The construction of the lower bound is based on a simple 3-armed bandit setting. We use Taylor expansion to transform Δ into δ , and note that the quadratic remainders can be negligible when θ is close to the optimal point. And thus the linear transformation can only achieve linear convergence. See detailed proof in Appendix A.1.2.

Theorem 2 (Lower bound of DPO-Unif). Let $|\mathcal{Y}| = 3, r(y_1) = 0, r(y_2) = 1/3, r(y_3) = 1$, and $\pi_{\text{ref}} = \text{Uniform}(\mathcal{Y})$. For any $\beta \in \mathbb{R}_+$ and learning rate $\eta \in (0, \frac{2}{\beta^2 |\mathcal{Y}|}]$, there always exists small enough $\epsilon \in \mathbb{R}_+$, for any initialization $\pi_{\theta^{(0)}}$ satisfying $\max_{y, y' \in \mathcal{Y}} |\delta(y, y'; \theta^{(0)})| \leq \epsilon$ and $\min_{y, y' \in \mathcal{Y}} |\delta(y, y'; \theta^{(0)})| > 0$, DPO-Unif satisfies

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

where $T \in \mathbb{N}$ is the number of iterations and γ is a constant depending on $\theta^{(0)}$.

Next we elaborate the idea of transforming Δ into δ using Taylor expansion, and show how to eliminate the linear term using appropriate samplers and learning rate. For Theorem 3, we can apply Taylor expansion at $r(y_1) - r(y_2)$ (while for Theorem 4 we apply at $\beta \log \frac{\pi_{\theta^{(t)}}(y_1) \pi_{\text{ref}}(y_2)}{\pi_{\text{ref}}(y_1) \pi_{\theta^{(t)}}(y_2)}$), and get

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

where ξ_R is an intermediate value. If we let $\pi^s(y_1, y_2) \propto 1/\sigma'(r(y_1) - r(y_2))$ as in Definition 4, then

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

Finally we pick an appropriate η to eliminate the initial linear term in Equation (6) and thus establish a quadratic convergence. This observation motivates our design of samplers and proofs. The detailed proofs of Theorems 3 and 4 can be found in Appendices A.2 and A.3.

Theorem 3 (Upper bound of DPO-Mix-R). Under Condition 1, DPO-Mix-R satisfies

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

where $T \in \mathbb{N}$ is the number of iterations.

Theorem 4 (Upper bound of DPO-Mix-P). Under Condition 1, DPO-Mix-P satisfies

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

where $T \in \mathbb{N}$ is the number of iterations.

Remark 5. DPO-Mix-R is not practical as the ground truth reward oracle is inaccessible. DPO-Mix-P is practical as it depends on only quantities we know, at the cost of a slightly slower convergence rate.

4.1.2 FOR EMPIRICAL DPO

As in Definition 2, exact gradients are inaccessible in practice. Here we show the guarantees of DPO-MIX-R and DPO-MIX-P with only unbiased estimation of gradients, that they can achieve linear convergence rates to the noise scale. The basic idea is to first eliminate the linear term in expectation as we do in Section 4.1.1, and then calculate the error propagation step by step. The proofs of Theorems 5 and 6 can be found in Appendix B.

Theorem 5. Under Condition 1 with the noise scale $\sigma \in (0, 1/576)$, DPO-MIX-R (σ) satisfies

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

where $T = \lceil \log \frac{1}{\sigma} \rceil$ is the number of iterations.

Theorem 6. Under Condition 1 with the noise scale $\sigma \in (0, 1/576)$, DPO-MIX-P* (σ) satisfies

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

where $T = \lceil \log \frac{1}{\sigma} \rceil$ is the number of iterations, and DPO-MIX-P* (σ) is DPO-MIX-P (σ) with a rejection sampling process: each time we get $y, y' \in \mathcal{Y}$ sampled from \mathcal{Q} , if $\psi(y, y'; \theta^{(t)}) := \left| \beta \log \frac{\pi_{\theta^{(t)}}(y)\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)\pi_{\theta^{(t)}}(y')} \right| > 1$, then reject this data pair with probability $1 - \frac{e+e^{-1}}{e^\psi + e^{-\psi}}$; and α_2 needs to be changed to $\frac{1}{2} \sum_{y, y' \in \mathcal{Y}} \min\{e^{\psi(y, y'; \theta^{(t)})} + e^{-\psi(y, y'; \theta^{(t)})}, e + e^{-1}\}$.

Remark 6. The rejection process is to modify the joint sampling distribution when the policy deviates too much, such that the coefficient of the quadratic term can still be bounded. This issue also exists in DPO-Unif (σ), and thus its convergence rate is not guaranteed.

4.2 NUMERICAL SIMULATIONS

We verify our theoretical findings with numerical simulations in *contextual bandits*. As shown in Figure 1, the two proposed samplers DPO-MIX-P and DPO-MIX-R show great improvements over DPO-Unif. The detailed configurations and more results can be found in Appendix D.1.

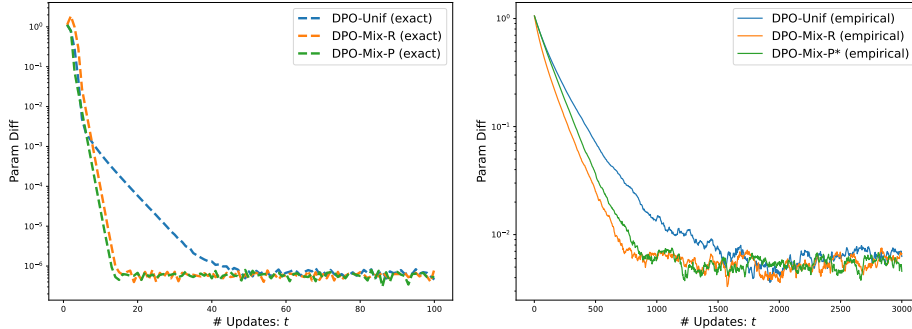


Figure 1: **Contextual bandit experiments for exact DPO and empirical DPO.** The x -axis is the number of gradient updates, and the y -axis is the total parameter difference $\sum_{y, y'} \delta(y, y'; \theta^{(t)})^2$. The left figure illustrates exact DPO, and the right figure illustrates empirical DPO. For exact DPO, the lower bound is due to the precision of floating numbers. For empirical DPO, the lower bound is due to sampling variances. The separation is clear in exact DPO, and still exists in empirical DPO.

5 IMPLICATIONS FOR PRACTICAL DPO

In this section, we show the implications of theoretical results in Section 4 for practical DPO design.

5.1 ALIGNING THEORY TO PRACTICE

Rethinking DPO. We can rewrite a policy $\pi \in \Pi$ as $\pi(y|x) \propto \pi_{\text{ref}}(y|x) e^{\varphi(x,y)/\beta}$, where $\varphi(x, y) \in \mathbb{R}_+$. Then the training objective of DPO can be rewritten as:

$$\varphi^*(x, \cdot) = \operatorname{argmin}_{\varphi(x, \cdot)} \sum_{y_1, y_2 \in \mathcal{Y}} \pi^s(y_1, y_2|x) \cdot \underbrace{(-\sigma(r(x, y_1) - r(x, y_2)) \log \sigma(\varphi(x, y_1) - \varphi(x, y_2)))}_{\text{cross entropy loss}},$$

which is learning a reward model $\varphi(x, y)$ towards $r(x, y) + C(x)$, where $C(x) \in \mathbb{R}$ is a constant. In Section 4, we have discussed the role of samplers in this implicit reward learning stage. Here we introduce a lemma (for multi-armed bandits) to connect it with the final performance.

Lemma 1 (Performance difference lemma). *For any θ , define its value as*

$$V^\theta := \mathbb{E}_{y \sim \pi_\theta} r(y) - \beta \text{KL}(\pi_\theta \| \pi_{\text{ref}}),$$

and let V^* be the value of the optimal policy π^* in Equation (3), then we have

$$\begin{aligned} V^* - V^\theta &= \sum_{y, y' \in \mathcal{Y}} \pi^*(y) \pi_\theta(y') \left(r(y) - r(y') - \beta \log \frac{\pi_\theta(y) \pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y) \pi_\theta(y')} \right) - \beta \text{KL}(\pi^* \| \pi_\theta) \\ &\leq \sum_{y, y' \in \mathcal{Y}} \pi^*(y) \pi_\theta(y') \left(r(y) - r(y') - \beta \log \frac{\pi_\theta(y) \pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y) \pi_\theta(y')} \right) \\ &= \mathbb{E}_{y \sim \pi^*, y' \sim \pi_\theta} \delta(y, y'; \theta). \end{aligned} \quad (7)$$

Setting the posterior. Lemma 1 indicates that reward learning should concentrate on responses with high probabilities for π^* and π_θ , and thus motivates us to change the distribution over \mathcal{Y} to a posterior distribution close to π^* or π_θ in practical implementation. This perspective provides an alternate explanation for Liu et al. (2024b), which uses rejection sampling to align the sampling distribution to π^* . Considering the fact that π^* is usually inaccessible, we propose to let $\pi_\theta^{2\beta}$ (after normalization) be the posterior distribution. **Setting the sampling temperature as 2β** , we can thus derive our new practical algorithm following Definition 5:

$$\textcircled{1} \begin{cases} \pi^{s1}(\cdot|x) = \pi_\theta(\cdot|x), \\ \pi^{s2}(\cdot|x) = \pi_\theta(\cdot|x), \end{cases} \quad \textcircled{2} \begin{cases} \pi^{s1}(\cdot|x) \propto \pi_\theta^{3/2}(\cdot|x) \pi_{\text{ref}}^{-1/2}(\cdot|x), \\ \pi^{s2}(\cdot|x) \propto \pi_\theta^{1/2}(\cdot|x) \pi_{\text{ref}}^{1/2}(\cdot|x), \end{cases}$$

and with a reward margin $r_{\text{max}} \in \mathbb{R}_+$ the mixing ratio can be roughly approximated as

$$\textcircled{1} : \textcircled{2} = 2 : (\exp(r_{\text{max}}) + \exp(-r_{\text{max}})). \quad (8)$$

Logit mixing. The proposed samplers involve a hybridization between two policies, and a common approach to approximate hybrid distributions is *logit mixing* (Shi et al., 2024; Liu et al., 2024a). Here we show how to understand this point in a theoretically sound way. Given $\pi_1, \pi_2 \in \Pi$, $w_1, w_2 \in \mathbb{R}$, we consider a new logit as $\zeta := w_1 \zeta_1 + w_2 \zeta_2$, where ζ_1, ζ_2 represent the per-token logits of policies π_1, π_2 , namely $\zeta_k(y_t|x, y_{<t}) = \log \pi_k(y_t|x, y_{<t})$. Note that

$$\begin{aligned} \operatorname{argmax}_{y \in \mathcal{Y}} \pi_1^{w_1}(y|x) \pi_2^{w_2}(y|x) &= \operatorname{argmax}_{y \in \mathcal{Y}} w_1 \log \pi_1(y|x) + w_2 \log \pi_2(y|x) \\ &= \operatorname{argmax}_{y \in \mathcal{Y}} \sum_{t=0}^{|y|} w_1 \zeta_1(y_t|x, y_{<t}) + w_2 \zeta_2(y_t|x, y_{<t}) \\ &= \operatorname{argmax}_{y \in \mathcal{Y}} \sum_{t=0}^{|y|} \zeta(y_t|x, y_{<t}). \end{aligned}$$

This indicates that, greedy decoding from $\pi \propto \pi_1^{w_1} \pi_2^{w_2}$ is equivalent to greedy decoding from $w_1 \zeta_1 + w_2 \zeta_2$. Thus, our proposed samplers can be implemented through mixing the logits of π_{ref} and π_θ .

Understanding existing approaches. Vanilla DPO (Rafailov et al., 2023) and its online variant (Xiong et al., 2024) can be incorporated into our theoretical framework. As shown in Table 1, vanilla DPO, which assumes that pair-comparison data are sampled from π_{ref} (see Section 4 of Rafailov et al. (2023)), can be viewed as DPO-UNIF; On-policy DPO (Guo et al., 2024; Tajwar et al., 2024) proposes to sample response pairs using π_{θ} , and is thus equivalent to DPO-UNIF; Hybrid GSHF (Option I in Xiong et al. (2024)) sets $\pi^{\text{s1}} = \pi_{\theta}$ and $\pi^{\text{s2}} = \pi_{\text{ref}}$, equivalent to DPO-MIX-P (②); and Online GSHF (Option II in Xiong et al. (2024)) adopts the best/worst-of- K response generated by π_{θ} , which can be approximately viewed as generating from $\pi_{\theta}(\cdot) \exp(r(\cdot)/\beta)$ and $\pi_{\theta}(\cdot) \exp(-r(\cdot)/\beta)$, *i.e.* DPO-MIX-R (②). Notably, the ① part is often omitted in DPO variants, and it can be attributed to the infinitely large reward margin in the implementation (Xiong et al., 2024; Dong et al., 2024), making the mixing ratio $\rightarrow 0 : 1$ in Equation (8) (see more details in Section 4 of Rosset et al. (2024) and Appendix C).

Table 1: **Comparison with existing approaches.** We find that many baselines can be mapped to components of our proposed samplers, offering an alternative explanation for their advantages.

Algorithm	Practical π^{s1}	Practical π^{s2}	Equivalent Sampler	Posterior Distribution
Vanilla DPO	π_{ref}	π_{ref}	DPO-UNIF	$\pi_{\text{ref}}^{2\beta}$
On-policy DPO	π_{θ}	π_{θ}	DPO-UNIF	$\pi_{\theta}^{2\beta}$
Hybrid GSHF	π_{θ}	π_{ref}	DPO-MIX-P (②)	$\pi_{\theta}^{\beta} \pi_{\text{ref}}^{\beta}$
Online GSHF	π_{θ} (best-of- K)	π_{θ} (worst-of- K)	DPO-MIX-R (②)	$\pi_{\theta}^{2\beta}$
Ours	$\pi_{\theta}^{3/2} \pi_{\text{ref}}^{-1/2}$	$\pi_{\theta}^{1/2} \pi_{\text{ref}}^{1/2}$	DPO-MIX-P	$\pi_{\theta}^{2\beta}$

Table 2: **Results on Safe-RLHF.** The average reward is scored by the gold reward model on train set and test set, and win-rate is against the reference model. Each algorithm is trained for 3 iterations, and in the final iteration, ours shows advantages over baselines across all metrics. **We repeat training each algorithm using 3 seeds, and show the mean and standard-variance.**

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%

Table 3: **Results on Iterative-Prompt.** The average reward is scored by the gold reward model on train set and test set, and win-rate is against the reference model. Each algorithm is trained for 3 iterations, and in the final iteration, ours shows advantages over baselines across all metrics. **We repeat training each algorithm using 3 seeds, and show the mean and standard-variance.**

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%

5.2 ALIGNMENT EXPERIMENTS

Experiment setup. We conduct experiments on two datasets, Safe-RLHF (Ji et al., 2023a) and Iterative-Prompt (Xiong et al., 2024; Dong et al., 2024). Our pipeline is mainly borrowed from Dong et al. (2024). For each iteration, responses are generated for a fixed set of prompts. Specifically, given prompt x , we generate $y_1 \sim \pi^{\text{S1}}(\cdot|x)$ and $y_2 \sim \pi^{\text{S2}}(\cdot|x)$. Each generated pair is annotated by a gold reward model (Dong et al., 2023) as (r_1, r_2) , and the corresponding loss is

$$\mathcal{L}_{(y_1, y_2)}(\theta) = -\sigma(r_{\max} \cdot (r_1 - r_2)) \log \sigma \left(\beta \log \frac{\pi_{\theta}(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_{\text{ref}}(y_1|x)\pi_{\theta}(y_2|x)} \right) - \sigma(r_{\max} \cdot (r_2 - r_1)) \log \sigma \left(\beta \log \frac{\pi_{\theta}(y_2|x)\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y_2|x)\pi_{\theta}(y_1|x)} \right),$$

where $r_{\max} \in \mathbb{R}_+$ is the reward margin. See more details in Appendix C.

Results. Experimental results on LM alignment are provided in Tables 2 and 3. In addition to the reward model, we employ GPT4o-mini to assess the win-rates on test sets. See implementations in Appendix C for details. On Safe-RLHF dataset, our method is 11% better than vanilla DPO and 9.5% better than on-policy DPO. On Iterative-Prompt dataset, ours improves by 13% compared to vanilla DPO and by 9.5% compared to hybrid GSHF. We also show the reward-KL curves in Figure 2, to indicate that the tuned models do not deviate much from π_{ref} .

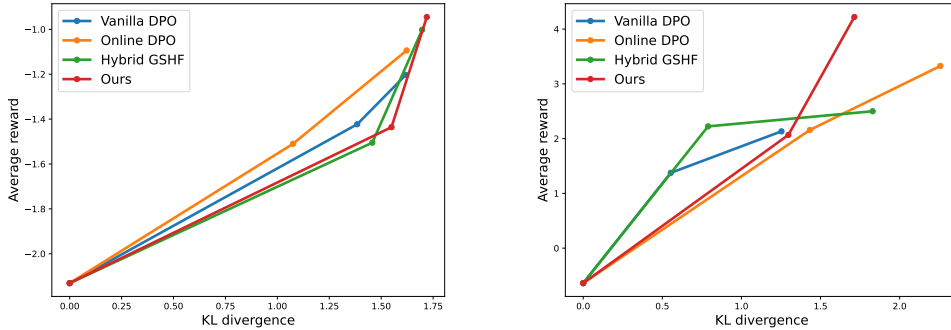


Figure 2: **The Reward-KL curves.** The left figure illustrates results on Safe-RLHF, and the right one illustrates results on Iterative-Prompt. The KL-divergence is measured on a subset of prompts in the test set. The results indicate that the KL-divergence of trained models does not deviate much from the reference model, and our method performs best in balancing reward and KL-divergence.

Clarification on evaluations. It is not enough to only show the results scored by reward models, since DPO algorithm is not explicitly learning the reward rankings (Meng et al., 2024; Chen et al., 2024). Due to restricted resources, we have not evaluated on open-benchmarks (Zheng et al., 2023; Dubois et al., 2024). Our work has demonstrated the potential to train models more effectively with minimal changes to the existing DPO pipeline. We hope this will inspire the community, especially those with rich computational resources, to conduct more systematic experiments.

6 CONCLUSION

This paper studies the convergence rates of DPO with different samplers. We demonstrate that DPO-Mix-R and DPO-Mix-P offer quadratic convergence rates, outperforming the linear rate of DPO-Unif. Our theoretical findings are supported by numerical simulations and LM alignment experiments.

It is also important to acknowledge our limitations. 1) The selection of the posterior distribution is not unique, and thus many useful samplers have yet to be developed from our framework and need further experiments. 2) The convergence analysis is based on *tabular softmax parametrization*, and a future direction would be exploring more practical settings such as *log-linear parametrization* and *function approximation*.

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

REFERENCES

- Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *ArXiv*, abs/2310.12036, 2023.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022.
- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444.
- Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. *ArXiv*, abs/2405.19534, 2024.
- Paul Francis Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *ArXiv*, abs/1706.03741, 2017.
- Muong Ding, Souradip Chakraborty, Vibhu Agrawal, Zora Che, Alec Koppel, Mengdi Wang, A. S. Bedi, and Furong Huang. Sail: Self-improving efficient online alignment of large language models. *ArXiv*, abs/2406.15567, 2024.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.
- Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf, 2024.
- Yann Dubois, Bal’azs Galambosi, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpaca-eval: A simple way to debias automatic evaluators. *ArXiv*, abs/2404.04475, 2024.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.
- Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *ArXiv*, abs/2307.04657, 2023a.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O’Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen Marcus McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey. *ArXiv*, abs/2310.19852, 2023b.
- Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. Rs-dpo: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. *ArXiv*, abs/2402.10038, 2024.
- Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares-López, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models. In *Forty-first International Conference on Machine Learning*, 2024a.

- 594 Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu
595 Liu. Statistical rejection sampling improves preference optimization. In *The Twelfth International*
596 *Conference on Learning Representations*, 2024b.
- 597 Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and
598 Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adver-
599 sarial regularizer. *ArXiv*, abs/2405.16436, 2024c.
- 600 Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence
601 rates of softmax policy gradient methods. In *International conference on machine learning*, pages
602 6820–6829. PMLR, 2020.
- 603 Jincheng Mei, Bo Dai, Alekh Agarwal, Mohammad Ghavamzadeh, Csaba Szepesvari, and Dale
604 Schuurmans. Ordering-based conditions for global convergence of policy gradient methods. In
605 *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- 606 Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a
607 reference-free reward. *ArXiv*, abs/2405.14734, 2024.
- 608 Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland,
609 Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash
610 learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.
- 611 OpenAI. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- 612 Philippe Rigollet. Lecture Notes for High-Dimensional Statistics
613 - 18.S997, Spring 2015. [https://ocw.mit.edu/courses/
614 18-s997-high-dimensional-statistics-spring-2015/
615 619e4ae252f1b26cbe0f7a29d5932978_MIT18_S997S15_CourseNotes.pdf](https://ocw.mit.edu/courses/18-s997-high-dimensional-statistics-spring-2015/619e4ae252f1b26cbe0f7a29d5932978_MIT18_S997S15_CourseNotes.pdf),
616 2015. Accessed: 2024-08-28.
- 617 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
618 Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-*
619 *seventh Conference on Neural Information Processing Systems*, 2023.
- 620 Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacrose, Ahmed Awadallah, and
621 Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general
622 preferences. *ArXiv*, abs/2404.03715, 2024.
- 623 Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A. Smith, and Simon S.
624 Du. Decoding-time language model alignment with multiple objectives. *ArXiv*, abs/2406.18853,
625 2024.
- 626 Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. The importance of online
627 data: Understanding preference fine-tuning via coverage, 2024.
- 628 Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A
629 minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint*
630 *arXiv:2401.04056*, 2024.
- 631 Fahim Tajwar, Anika Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Ste-
632 fano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage
633 suboptimal, on-policy data. *ArXiv*, abs/2404.14367, 2024.
- 634 Christian Wirth, Riad Akrou, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-
635 based reinforcement learning methods. *J. Mach. Learn. Res.*, 18:136:1–136:46, 2017.
- 636 Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and
637 Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation
638 for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.
- 639 Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang.
640 Iterative preference learning from human feedback: Bridging theory and practice for RLHF under
641 KL-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

648 Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton
649 Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm
650 performance in machine translation. *ArXiv*, abs/2401.08417, 2024.
651
652 Natalia Zhang, Xinqi Wang, Qiwen Cui, Runlong Zhou, Sham M. Kakade, and Simon S. Du. Multi-
653 agent reinforcement learning from human feedback: Data coverage and algorithmic techniques.
654 2024.
655
656 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
657 Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica.
658 Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023.
659
660 Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human
661 feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun
662 Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th
663 International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning
664 Research*, pages 43037–43067. PMLR, 23–29 Jul 2023.
665
666 Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul
667 Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv*,
668 abs/1909.08593, 2019.
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701

Appendix

Table of Contents

A Proofs of Convergence Rates of Exact DPO	15
A.1 Theorems 1 and 2: Linear Convergence of Exact DPO-Unif	15
A.2 Theorem 3: Quadratic Convergence of Exact DPO-Mix-R	19
A.3 Theorem 4: Quadratic Convergence of Exact DPO-Mix-P	20
B Proof of Convergence Rates of Empirical DPO	21
B.1 Technical Lemma	21
B.2 Theorem 5: Convergence of Empirical DPO-Mix-R	21
B.3 Theorem 6: Convergence of Empirical DPO-Mix-P*	23
C Implementation Details	25
D Supplementary Results	25
D.1 More Numerical Simulations	25
D.2 Example Generations	30
E Further Discussions	31

702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

A PROOFS OF CONVERGENCE RATES OF EXACT DPO

Without loss of generality, we assume π_{ref} to be uniform distribution throughout this section. In the main text, we use \mathcal{Y} to represent the action space and y to represent an action for compatibility with other LM papers. From here, we turn back to \mathcal{A} for action space, a for an action, and A for the size of \mathcal{A} since all the proofs are conducted in bandit environments. And for notational ease, we make the following definitions:

$$\begin{aligned}\Delta(a, a'; \theta) &:= \sigma(r(a) - r(a')) - \sigma(\beta(\theta_a - \theta_{a'})) , \\ \delta(a, a'; \theta) &:= r(a) - r(a') - \beta(\theta_a - \theta_{a'}) .\end{aligned}$$

A.1 THEOREMS 1 AND 2: LINEAR CONVERGENCE OF EXACT DPO-UNIF

A.1.1 PROOF OF UPPER BOUND

For DPO with uniform sampler on action pairs, we first claim that for any θ appearing in the optimization process,

$$\max_{a, a'} \{\beta(\theta_a - \theta_{a'})\} \leq R_{\max} ,$$

where R_{\max} will be bounded later, and let $\sigma'_{\min} := \sigma'(R_{\max}) = \sigma(R_{\max})\sigma(-R_{\max})$. Then we have

$$\sigma'_{\min} \leq \frac{\sigma(x) - \sigma(y)}{x - y} \leq \frac{1}{4} \text{ when } |x|, |y| \leq R_{\max} \text{ and } x \neq y , \quad (9)$$

$$\mathcal{L}(\theta) = -\frac{2}{A^2} \sum_{a, a'} p^*(a > a') \log \sigma \left(\beta \log \frac{\pi_{\theta}(a)}{\pi_{\theta}(a')} \right) , \quad (10)$$

$$\nabla_{\theta} \mathcal{L}(\theta) = -\frac{2\beta}{A^2} \sum_{a, a'} \Delta(a, a'; \theta) \mathbb{1}_a . \quad (11)$$

Equation (11) reduces to

$$\nabla_{\theta_a} \mathcal{L}(\theta) = -\frac{2\beta}{A^2} \sum_{a'} \Delta(a, a'; \theta) .$$

Thus for any action pair (a, a') ,

$$\begin{aligned}(\theta_a - \theta_{a'})^{(t+1)} &= (\theta_a - \theta_{a'})^{(t)} + \frac{2\eta\beta\alpha(\pi^{s1}, \pi^{s2})}{A^2} \sum_{a''} \left(\Delta(a, a''; \theta^{(t)}) - \Delta(a', a''; \theta^{(t)}) \right) \\ &= (\theta_a - \theta_{a'})^{(t)} + 4\eta\beta \sum_{a''} \left(\Delta(a, a''; \theta^{(t)}) - \Delta(a', a''; \theta^{(t)}) \right) .\end{aligned}$$

At time t , sort the actions in the order that $r(a_i) - \beta\theta_{a_i}^{(t)} \leq r(a_{i+1}) - \beta\theta_{a_{i+1}}^{(t)}$. Then we have $\Delta(a_i, a_j; \theta^{(t)}) \geq 0$ if $i > j$. Note that it is possible that the order of actions at time $t+1$ is different, and in the following proof for any index i , a_i is from the order at time t . Let $l < r$, then

$$\begin{aligned}&\delta(a_r, a_l; \theta^{(t+1)}) \\ &= \delta(a_r, a_l; \theta^{(t)}) - 4\eta\beta^2 \sum_{i=1}^A \left(\Delta(a_r, a_i; \theta^{(t)}) - \Delta(a_l, a_i; \theta^{(t)}) \right) \\ &\stackrel{(i)}{\leq} \delta(a_r, a_l; \theta^{(t)}) - 4\eta\beta^2 \sum_{i=1}^{l-1} \left(\sigma'_{\min} \delta(a_r, a_i; \theta^{(t)}) - \frac{1}{4} \delta(a_l, a_i; \theta^{(t)}) \right) \\ &\quad - 4\eta\beta^2 \sum_{i=l}^r \left(\sigma'_{\min} \delta(a_r, a_i; \theta^{(t)}) - \sigma'_{\min} \delta(a_l, a_i; \theta^{(t)}) \right) - 4\eta\beta^2 \sum_{i=r+1}^A \left(\frac{1}{4} \delta(a_r, a_i; \theta^{(t)}) - \sigma'_{\min} \delta(a_l, a_i; \theta^{(t)}) \right) \\ &= \delta(a_r, a_l; \theta^{(t)}) - 4\eta\beta^2 \left[\sigma'_{\min} (l-1) \delta(a_r, a_l; \theta^{(t)}) - \left(\frac{1}{4} - \sigma'_{\min} \right) \sum_{i=1}^{l-1} \delta(a_l, a_i; \theta^{(t)}) \right]\end{aligned}$$

$$\begin{aligned}
& -4\eta\beta^2\sigma'_{\min}(r-l+1)\delta(a_r, a_l; \theta^{(t)}) - 4\eta\beta^2 \left[\sigma'_{\min}(A-r)\delta(a_r, a_l; \theta^{(t)}) - \left(\frac{1}{4} - \sigma'_{\min}\right) \sum_{i=r+1}^A \delta(a_i, a_r; \theta^{(t)}) \right] \\
& = (1 - 4\eta\beta^2 A\sigma'_{\min}) \delta(a_r, a_l; \theta^{(t)}) + 4\eta\beta^2 \left(\frac{1}{4} - \sigma'_{\min}\right) \left(\sum_{i=1}^{l-1} \delta(a_l, a_i; \theta^{(t)}) + \sum_{i=r+1}^A \delta(a_i, a_r; \theta^{(t)}) \right),
\end{aligned}$$

where (i) is by using Equation (9) for different cases of x and y and whether $x - y > 0$. Similarly, for the lower bound:

$$\begin{aligned}
& -\delta(a_r, a_l; \theta^{(t+1)}) \\
& = 4\eta\beta^2 \sum_{i=1}^A \left(\Delta(a_r, a_i; \theta^{(t)}) - \Delta(a_l, a_i; \theta^{(t)}) \right) - \delta(a_r, a_l; \theta^{(t)}) \\
& \leq 4\eta\beta^2 \sum_{i=1}^{l-1} \left(\frac{1}{4}\delta(a_r, a_i; \theta^{(t)}) - \sigma'_{\min}\delta(a_l, a_i; \theta^{(t)}) \right) + 4\eta\beta^2 \sum_{i=l}^r \left(\frac{1}{4}\delta(a_r, a_i; \theta^{(t)}) - \frac{1}{4}\delta(a_l, a_i; \theta^{(t)}) \right) \\
& \quad + 4\eta\beta^2 \sum_{i=r+1}^A \left(\sigma'_{\min}\delta(a_r, a_i; \theta^{(t)}) - \frac{1}{4}\delta(a_l, a_i; \theta^{(t)}) \right) - \delta(a_r, a_l; \theta^{(t)}) \\
& = 4\eta\beta^2 \left[\frac{1}{4}(l-1)\delta(a_r, a_l; \theta^{(t)}) + \left(\frac{1}{4} - \sigma'_{\min}\right) \sum_{i=1}^{l-1} \delta(a_l, a_i; \theta^{(t)}) \right] + 4\eta\beta^2 \cdot \frac{1}{4}(r-l+1)\delta(a_r, a_l; \theta^{(t)}) \\
& \quad + 4\eta\beta^2 \left[\frac{1}{4}(A-r)\delta(a_r, a_l; \theta^{(t)}) + \left(\frac{1}{4} - \sigma'_{\min}\right) \sum_{i=r+1}^A \delta(a_i, a_r; \theta^{(t)}) \right] - \delta(a_r, a_l; \theta^{(t)}) \\
& = (\eta\beta^2 A - 1) \delta(a_r, a_l; \theta^{(t)}) + 4\eta\beta^2 \left(\frac{1}{4} - \sigma'_{\min}\right) \left(\sum_{i=1}^{l-1} \delta(a_l, a_i; \theta^{(t)}) + \sum_{i=r+1}^A \delta(a_i, a_r; \theta^{(t)}) \right).
\end{aligned}$$

Now taking $\eta = \frac{1}{\beta^2 A}$, then we have

$$\begin{aligned}
\delta(a_r, a_l; \theta^{(t+1)}) & \leq (2 - 8\sigma'_{\min}) \max_{a, a'} \delta(a, a'; \theta^{(t)}), \\
-\delta(a_r, a_l; \theta^{(t+1)}) & \leq (1 - 4\sigma'_{\min}) \max_{a, a'} \delta(a, a'; \theta^{(t)}).
\end{aligned}$$

Define

$$\gamma := 2 - 8\sigma'_{\min}$$

as the contraction factor, then

$$\left| \delta(a_r, a_l; \theta^{(t+1)}) \right| \leq \gamma \max_{a, a'} \left| \delta(a, a'; \theta^{(t)}) \right|. \quad (12)$$

Recall that we initialize $\theta^{(0)} = \vec{0}$. Next we use induction to verify that throughout the process ($t \geq 0$),

$$\left| \delta(a_r, a_l; \theta^{(t+1)}) \right| \leq 0.214\gamma^t, \quad \text{and} \quad \left| \beta(\theta_a - \theta_{a'})^{(t+1)} \right| < 1.214. \quad (13)$$

For time $t = 0$, we have special versions: $r(a_1) \leq r(a_2) \leq \dots \leq r(a_A)$.

$$\begin{aligned}
\delta(a_r, a_l; \theta^{(1)}) & = r(a_r) - r(a_l) - 4\eta\beta^2 \sum_{i=1}^A (\Delta(a_r, a_i; \theta^{(0)}) - \Delta(a_l, a_i; \theta^{(0)})) \\
& \stackrel{(i)}{=} r(a_r) - r(a_l) - 4\eta\beta^2 \sum_{i=1}^A (\sigma(r(a_r) - r(a_i)) - \sigma(r(a_l) - r(a_i))) \\
& \stackrel{(ii)}{\leq} r(a_r) - r(a_l) - 4\eta\beta^2 \sum_{i=1}^A \sigma'(1) [r(a_r) - r(a_i) - (r(a_l) - r(a_i))] \\
& = (1 - 4\eta\beta^2 A\sigma'(1)) (r(a_r) - r(a_l))
\end{aligned}$$

$$\begin{aligned}
& \stackrel{\text{(iii)}}{\leq} 0.214 ; \\
-\delta(a_r, a_l; \theta^{(1)}) &= 4\eta\beta^2 \sum_{i=1}^A (\Delta(a_r, a_i; \theta^{(0)}) - \Delta(a_l, a_i; \theta^{(0)})) - (r(a_r) - r(a_l)) \\
&= 4\eta\beta^2 \sum_{i=1}^A (\sigma(r(a_r) - r(a_i)) - \sigma(r(a_l) - r(a_i))) - (r(a_r) - r(a_l)) \\
&\leq 4\eta\beta^2 \sum_{i=1}^A \frac{1}{4} [r(a_r) - r(a_i) - (r(a_l) - r(a_i))] - (r(a_r) - r(a_l)) \\
&= (\eta\beta^2 A - 1) (r(a_r) - r(a_l)) \\
&= 0 ,
\end{aligned}$$

where (i) is by $\theta^{(0)} = \vec{0}$; (ii) is by Equation (9) and $r(a_r) - r(a_i) \geq r(a_l) - r(a_i)$; (iii) is by $r(a_r) - r(a_l) \leq 1$. So $|\delta(a_r, a_l; \theta^{(1)})| \leq 0.214$, and $|\beta(\theta_a - \theta_{a'})_1| \leq |r(a) - r(a')| + |\delta(a_r, a_l; \theta^{(1)})| \leq 1.214$. Suppose for time $t - 1$, Equation (13) holds, then Equation (12) holds. So for time t ,

$$|\delta(a_r, a_l; \theta^{(t+1)})| \leq \gamma \max_{a, a'} |\delta(a_r, a_l; \theta^{(t)})| \leq 0.214\gamma^t \leq 0.214 ,$$

and

$$|\beta(\theta_a - \theta_{a'})^{(t+1)}| \leq |r(a) - r(a')| + |\delta(a_r, a_l; \theta^{(t+1)})| \leq 1.214 .$$

Thus we have

$$\begin{aligned}
\gamma &= 2 - 8\sigma'_{\min} \leq 2 - 8\sigma'(1.214) < 0.588 , \\
|\delta(a_r, a_l; \theta^{(T)})| &\leq 0.588^T .
\end{aligned}$$

A.1.2 CONSTRUCTION OF LOWER BOUND

Consider a three-armed bandit setting with rewards $r(a_1) = 0, r(a_2) = 1/3, r(a_3) = 1$ and any regularization coefficient $\beta \in \mathbb{R}_+$. The update rule satisfies:

$$\delta(a_2, a_1; \theta^{(t+1)}) = \delta(a_2, a_1; \theta^{(t)}) - 4\eta\beta^2 \left(2\Delta(a_2, a_1; \theta^{(t)}) + \Delta(a_3, a_1; \theta^{(t)}) - \Delta(a_3, a_2; \theta^{(t)}) \right) , \quad (14)$$

$$\delta(a_3, a_2; \theta^{(t+1)}) = \delta(a_3, a_2; \theta^{(t)}) - 4\eta\beta^2 \left(2\Delta(a_3, a_2; \theta^{(t)}) + \Delta(a_3, a_1; \theta^{(t)}) - \Delta(a_2, a_1; \theta^{(t)}) \right) , \quad (15)$$

$$\delta(a_3, a_1; \theta^{(t+1)}) = \delta(a_3, a_1; \theta^{(t)}) - 4\eta\beta^2 \left(2\Delta(a_3, a_1; \theta^{(t)}) + \Delta(a_3, a_2; \theta^{(t)}) + \Delta(a_2, a_1; \theta^{(t)}) \right) .$$

Define $x_t := \delta(a_2, a_1; \theta^{(t)})$, and $y_t := \delta(a_3, a_2; \theta^{(t)})$. Clearly we have $\delta(a_3, a_1; \theta^{(t)}) = x_t + y_t$. We can perform Taylor expansion on Equations (14) and (15) and get

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 - 4\eta\beta^2(2\sigma'(1/3) + \sigma'(1)) & 4\eta(\sigma'(2/3) - \sigma'(1)) \\ 4\eta\beta^2(\sigma'(1/3) - \sigma'(1)) & 1 - 4\eta\beta^2(2\sigma'(2/3) + \sigma'(1)) \end{pmatrix}}_{:=B} \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \eta\beta^2 \begin{pmatrix} u_t \\ v_t \end{pmatrix} , \quad (16)$$

where

$$|u_t| \leq \frac{4x_t^2 + 3y_t^2}{3\sqrt{3}} \leq x_t^2 + y_t^2 , \quad |v_t| \leq \frac{3x_t^2 + 4y_t^2}{3\sqrt{3}} \leq x_t^2 + y_t^2 . \quad (17)$$

Now we analyze the eigenvalues of B under three scenarios.

1. If

$$0 < \eta\beta^2 < \frac{1}{4(2\sigma'(1/3) + \sigma'(1))} \approx 0.366 ,$$

then we have

$$\begin{aligned}
\det(\lambda I - B) &= \lambda^2 - (B_{11} + B_{22})\lambda + \underbrace{B_{11}B_{22}}_{\leq (B_{11}+B_{22})^2/4} - \underbrace{B_{12}B_{21}}_{>0} .
\end{aligned}$$

918 2. If

$$919 \frac{1}{4(2\sigma'(1/3) + \sigma'(1))} \leq \eta\beta^2 \leq \frac{1}{4(2\sigma'(2/3) + \sigma'(1))} \approx 0.388 ,$$

922 then we have

$$923 \det(\lambda I - B) = \lambda^2 - (B_{11} + B_{22})\lambda + \underbrace{B_{11}B_{22}}_{\leq 0} - \underbrace{B_{12}B_{21}}_{>0} .$$

927 3. If

$$928 \frac{1}{4(2\sigma'(2/3) + \sigma'(1))} < \eta\beta^2 < \frac{1}{2(2\sigma'(1/3) + \sigma'(2/3))} \approx 0.704 ,$$

932 then we have

$$933 \det(\lambda I - B) = \lambda^2 - (B_{11} + B_{22})\lambda + \underbrace{B_{11}B_{22}}_{\leq (B_{11}+B_{22})^2/4} - \underbrace{B_{12}B_{21}}_{>0} .$$

936 Therefore B has two different eigenvalues $\lambda_1, \lambda_2 \in (-1, 0) \cup (0, 1)$, with normalized eigenvectors w_1, w_2 . Clearly $w_{ij} \in (-1, 0) \cup (0, 1)$, $\forall i, j \in \{1, 2\}$. Then we define $\lambda_{\max} := \max(|\lambda_1|, |\lambda_2|)$, $\lambda_{\min} := \min(|\lambda_1|, |\lambda_2|)$, Now perform basis transformation with new basis (w_1, w_2) . Thus Equation (16) can be rewritten as

$$941 \begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix} + \begin{pmatrix} u'_t \\ v'_t \end{pmatrix} ,$$

944 Let w'_1, w'_2 be the inverse basis, and define $\alpha := \max_{i,j \in \{1,2\}} |w'_{ij}|$, and $\epsilon :=$
 945 $\min(\lambda_{\min}, 1 - \lambda_{\max}) / (64\alpha^2)$. Now initialize $|x_0|, |y_0| \in (0, \epsilon)$. Then we have
 946 $\max_{i,j \in \{1,2,3\}} |\delta(a_i, a_j; \theta^{(0)})| \leq 2\epsilon$. Therefore

$$949 |p_0|, |q_0| \stackrel{(i)}{\leq} 2\alpha\epsilon ,$$

951 and

$$952 |u'_t|, |v'_t| \stackrel{(ii)}{\leq} 2\alpha(x_t^2 + y_t^2)$$

$$953 \stackrel{(iii)}{\leq} 4\alpha(p_t^2 + q_t^2) .$$

956 (i) and (ii) comes from the fact that $p_t = w'_{11}x_t + w'_{21}y_t$ and $q_t = w'_{12}x_t + w'_{22}y_t$, and Equation (17);
 957 (iii) is from the fact that $x_t = w_{11}p_t + w_{21}q_t$ and $y_t = w_{12}p_t + w_{22}q_t$, and Cauchy-Schwarz
 958 inequality. Now we have

$$960 |p_{t+1}| + |q_{t+1}| \leq [\lambda_{\max} + 8\alpha(|p_t| + |q_t|)] (|p_t| + |q_t|)$$

$$961 \stackrel{(iv)}{\leq} (\lambda_{\max} + 32\alpha^2\epsilon) (|p_t| + |q_t|)$$

$$962 \leq \frac{1 + \lambda_{\max}}{2} (|p_t| + |q_t|) .$$

$$963 |p_{t+1}| + |q_{t+1}| \geq [\lambda_{\min} - 8\alpha(|p_t| + |q_t|)] (|p_t| + |q_t|)$$

$$964 \stackrel{(v)}{\geq} (\lambda_{\min} - 32\alpha^2\epsilon) (|p_t| + |q_t|)$$

$$965 \geq \frac{\lambda_{\min}}{2} (|p_t| + |q_t|) ,$$

968 where (iv) and (v) are based on simple induction that $|p_t| + |q_t|$ will not increase. And it thus
 969 indicates that $\max(|x_t|, |y_t|)$ can at most be linear convergence when $\eta\beta^2 \leq \frac{2}{A} \approx 0.667$.

972 A.2 THEOREM 3: QUADRATIC CONVERGENCE OF EXACT DPO-MIX-R
973

974 We study DPO with a mixture of fixed samplers: $Z^+ Z^- \cdot \pi^{s1} \times \pi^{s2} + A^2 \cdot \text{Uniform}(\mathcal{A}) \times \text{Uniform}(\mathcal{A})$,
975 where $Z^+ = \sum_a \exp(r(a))$, $\pi^{s1}(a) = \exp(r(a))/Z^+$ and $Z^- = \sum_a \exp(-r(a))$, $\pi^{s2}(A) =$
976 $\exp(-r(a))/Z^-$. We have

977
$$\alpha_1 \mathcal{L}_1(\theta) + \alpha_2 \mathcal{L}_2(\theta)$$

978
979
$$= - \sum_{a, a'} \left(A^2 \cdot \frac{1}{A^2} + Z^+ Z^- \cdot \pi^{s1}(a) \pi^{s2}(a') \right) \left[p^*(a > a') \log \sigma \left(\beta \log \frac{\pi_\theta(a)}{\pi_\theta(a')} \right) + p^*(a' > a) \log \sigma \left(\beta \log \frac{\pi_\theta(a')}{\pi_\theta(a)} \right) \right]$$

980
981
982
$$= - \sum_{a, a'} (\exp(r(a) - r(a')) + 1) \left[p^*(a > a') \log \sigma \left(\beta \log \frac{\pi_\theta(a)}{\pi_\theta(a')} \right) + p^*(a' > a) \log \sigma \left(\beta \log \frac{\pi_\theta(a')}{\pi_\theta(a)} \right) \right],$$

983
984
$$\nabla_\theta (\alpha_1 \mathcal{L}_1(\theta) + \alpha_2 \mathcal{L}_2(\theta))$$

985
$$= -\beta \sum_{a, a'} (\exp(r(a) - r(a')) + 1) \Delta(a, a'; \theta) (\mathbb{1}_a - \mathbb{1}_{a'})$$

986
987
$$= -\beta \sum_{a, a'} (\exp(r(a) - r(a')) + \exp(r(a') - r(a)) + 2) \Delta(a, a'; \theta) \mathbb{1}_a$$

988
989
990
$$= -\beta \sum_{a, a'} \frac{\Delta(a, a'; \theta)}{\sigma'(r(a) - r(a'))} \mathbb{1}_a. \tag{18}$$

991
992

993 Equation (18) reduces to

994
995
$$\nabla_{\theta_a} (\alpha_1 \mathcal{L}_1(\theta) + \alpha_2 \mathcal{L}_2(\theta)) = -\beta \sum_{a'} \frac{\Delta(a, a'; \theta)}{\sigma'(r(a) - r(a'))}.$$

996

997 Fix parameter θ . For any action pair a, a' , through Taylor expansion we have that

998
999
$$\Delta(a, a'; \theta) = \sigma'(r(a) - r(a')) \delta(a, a'; \theta) - \frac{\sigma''(\xi_{\mathbb{R}}(a, a'; \theta))}{2} \delta(a, a'; \theta)^2,$$

1000

1001 where $\xi_{\mathbb{R}}(a, a'; \theta)$ is between $r(a) - r(a')$ and $\beta(\theta_a - \theta_{a'})$. We have that at time step t , for any
1002 action pair (a, a') ,

1003
1004
$$\delta(a, a'; \theta^{(t+1)}) = \delta(a, a'; \theta^{(t)}) - \eta \beta^2 \sum_{a''} \left(\frac{\Delta(a, a''; \theta^{(t)})}{\sigma'(r(a) - r(a''))} - \frac{\Delta(a', a''; \theta^{(t)})}{\sigma'(r(a') - r(a''))} \right)$$

1005
1006
$$= \delta(a, a'; \theta^{(t)}) - \eta \beta^2 \sum_{a''} (\delta(a, a''; \theta^{(t)}) - \delta(a', a''; \theta^{(t)}))$$

1007
1008
$$+ \frac{\eta \beta^2}{2} \sum_{a''} \left(\frac{\sigma''(\xi_{\mathbb{R}}(a, a''; \theta^{(t)}))}{\sigma'(r(a) - r(a''))} \delta(a, a''; \theta^{(t)})^2 - \frac{\sigma''(\xi_{\mathbb{R}}(a', a''; \theta^{(t)}))}{\sigma'(r(a') - r(a''))} \delta(a', a''; \theta^{(t)})^2 \right)$$

1009
1010
$$= (1 - \eta \beta^2 A) \delta(a, a'; \theta^{(t)})$$

1011
1012
$$+ \frac{\eta \beta^2}{2} \sum_{a''} \left(\frac{\sigma''(\xi_{\mathbb{R}}(a, a''; \theta^{(t)}))}{\sigma'(r(a) - r(a''))} \delta(a, a''; \theta^{(t)})^2 - \frac{\sigma''(\xi_{\mathbb{R}}(a', a''; \theta^{(t)}))}{\sigma'(r(a') - r(a''))} \delta(a', a''; \theta^{(t)})^2 \right).$$

1013
1014
1015

1016 From the range of r , we know that $\sigma'(r(a) - r(a')) \geq \sigma'(1) > 0.196$. We have
1017 $|\sigma''(\xi_{\mathbb{R}}(a, a''; \theta^{(t)}))| \leq \sigma''_{\max} := \sup_{0 \leq x \leq 1} x(1-x)(1-2x) = 1/(6\sqrt{3}) < 0.097$. Set

1018
1019
$$\eta = \frac{1}{\beta^2 A},$$

1020

1021 then

1022
$$\left| \delta(a, a'; \theta^{(t+1)}) \right| \leq \frac{1}{2A} \sum_{a''} \left(\frac{\sigma''_{\max}}{\sigma'(1)} \delta(a, a''; \theta^{(t)})^2 + \frac{\sigma''_{\max}}{\sigma'(1)} \delta(a', a''; \theta^{(t)})^2 \right)$$

1023
1024
$$\leq \frac{\sigma''_{\max}}{\sigma'(1)} \max_{a, a'} \delta(a, a'; \theta^{(t)})^2$$

1025

$$1026 \quad < \frac{1}{2} \max_{a,a'} \delta(a, a'; \theta^{(t)})^2 .$$

1027
1028 Since $\max_{a,a'} |\delta(a, a'; \theta^{(0)})| \leq 1$, we can show a quadratic convergence for this regime:

$$1029 \quad \left| \delta(a, a'; \theta^{(t)}) \right| \leq 0.5^{2^t - 1} .$$

1030 A.3 THEOREM 4: QUADRATIC CONVERGENCE OF EXACT DPO-MIX-P

1031
1032 We study DPO with a mixture of on-policy samplers (with gradient stopped) and uniform samplers:
1033 $Z^+ Z^- \cdot \pi^{s_1} \times \pi^{s_2} + A^2 \cdot \text{Uniform}(\mathcal{A}) \times \text{Uniform}(\mathcal{A})$, where $Z^+ = \sum_a \exp(\beta \theta_a)$, $\pi^{s_1}(a) =$
1034 $\exp(\beta \theta_a)/Z^+$ and $Z^- = \sum_a \exp(-\beta \theta_a)$, $\pi^{s_2}(a) = \exp(-\beta \theta_a)/Z^-$. Samely we have

$$1035 \quad \nabla_{\theta_a} (\alpha_1 \mathcal{L}_1(\theta) + \alpha_2 \mathcal{L}_2(\theta)) = -\beta \sum_{a'} \frac{\Delta(a, a'; \theta)}{\sigma'(\beta(\theta_a - \theta_{a'}))} .$$

1036
1037 Fix parameter θ . For any action pair a, a' , through Taylor expansion we have that

$$1038 \quad \Delta(a, a'; \theta) = \sigma'(\beta(\theta_a - \theta_{a'})) \delta(a, a'; \theta) + \frac{\sigma''(\xi_P(a, a'; \theta))}{2} \delta(a, a'; \theta)^2 ,$$

1039
1040 where $\xi_P(a, a'; \theta)$ is between $r(a) - r(a')$ and $\beta(\theta_a - \theta_{a'})$. We have that at time step t , for any
1041 action pair (a, a') ,

$$1042 \quad \begin{aligned} 1043 \delta(a, a'; \theta^{(t+1)}) &= \delta(a, a'; \theta^{(t)}) - \eta \beta^2 \sum_{a''} \left(\frac{\Delta(a, a''; \theta^{(t)})}{\sigma'(\beta(\theta_a - \theta_{a''})^{(t)})} - \frac{\Delta(a', a''; \theta^{(t)})}{\sigma'(\beta(\theta_{a'} - \theta_{a''})^{(t)})} \right) \\ 1044 &= \delta(a, a'; \theta^{(t)}) - \eta \beta^2 \sum_{a''} (\delta(a, a''; \theta^{(t)}) - \delta(a', a''; \theta^{(t)})) \\ 1045 &\quad - \frac{\eta \beta^2}{2} \sum_{a''} \left(\frac{\sigma''(\xi_P(a, a''; \theta^{(t)}))}{\sigma'(\beta(\theta_a - \theta_{a''})^{(t)})} \delta(a, a''; \theta^{(t)})^2 - \frac{\sigma''(\xi_P(a', a''; \theta^{(t)}))}{\sigma'(\beta(\theta_{a'} - \theta_{a''})^{(t)})} \delta(a', a''; \theta^{(t)})^2 \right) \\ 1046 &= (1 - \eta \beta^2 A) \delta(a, a'; \theta^{(t)}) \\ 1047 &\quad - \frac{\eta \beta^2}{2} \sum_{a''} \left(\frac{\sigma''(\xi_P(a, a''; \theta^{(t)}))}{\sigma'(\beta(\theta_a - \theta_{a''})^{(t)})} \delta(a, a''; \theta^{(t)})^2 - \frac{\sigma''(\xi_P(a', a''; \theta^{(t)}))}{\sigma'(\beta(\theta_{a'} - \theta_{a''})^{(t)})} \delta(a', a''; \theta^{(t)})^2 \right) . \end{aligned}$$

1048
1049 We still first claim that $\sigma'(\beta(\theta_a - \theta_{a'})_t) \geq \sigma'_{\min}$, and will bound it later. We have
1050 $|\sigma''(\xi_P(a, a''; \theta^{(t)}))| \leq \sigma''_{\max} < 0.097$. Set

$$1051 \quad \eta = \frac{1}{\beta^2 A} ,$$

1052 then

$$1053 \quad \begin{aligned} 1054 \left| \delta(a, a'; \theta^{(t+1)}) \right| &\leq \frac{\sigma''_{\max}}{2A\sigma'_{\min}} \sum_{a''} (\delta(a, a''; \theta^{(t)})^2 + \delta(a', a''; \theta^{(t)})^2) \\ 1055 &\leq \frac{\sigma''_{\max}}{\sigma'_{\min}} \max_{a,a'} \delta(a, a'; \theta^{(t)})^2 . \end{aligned}$$

1056
1057 At time step $t = 0$ we have $\sigma'(\beta(\theta_a - \theta_{a'})^{(0)}) = \sigma'(0) = 0.25$ and $\max_{a,a'} |\delta(a, a'; \theta^{(0)})| \leq 1$, so

$$1058 \quad \max_{a,a'} \left| \delta(a, a'; \theta^{(1)}) \right| < 0.388 .$$

1059
1060 By simple induction, we have that

$$1061 \quad \sigma'_{\min} \geq \sigma'(1 + \max_{a,a'} |\delta(a, a'; \theta^{(t)})|) \geq \sigma'(1.388) > 0.159 ,$$

$$1062 \quad \max_{a,a'} \left| \delta(a, a'; \theta^{(t+1)}) \right| \leq \frac{0.097}{0.159} \max_{a,a'} \delta(a, a'; \theta^{(t)})^2 < 0.611 \max_{a,a'} \delta(a, a'; \theta^{(t)})^2 .$$

1063
1064 which is a quadratic convergence:

$$1065 \quad \left| \delta(a, a'; \theta^{(t)}) \right| \leq 0.611^{2^t - 1} .$$

B PROOF OF CONVERGENCE RATES OF EMPIRICAL DPO

For notational ease, we make the following definitions throughout this section:

$$\begin{aligned}\Delta(a, a'; \theta) &:= \sigma(r(a) - r(a')) - \sigma(\beta(\theta_a - \theta_{a'})), \\ \delta(a, a'; \theta) &:= r(a) - r(a') - \beta(\theta_a - \theta_{a'}).\end{aligned}$$

This section conforms to Definition 2. Denote the filtration \mathcal{F}_t as all the samples on and before time step t .

B.1 TECHNICAL LEMMA

Lemma 2 (Lemma 1.4 in [Philippe Rigollet \(2015\)](#)). *Let X be a random variable such that*

$$\mathbb{P}[|X| > t] \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

then for any positive integer $k \geq 2$,

$$\mathbb{E}[|X|^k] \leq (\sigma e^{1/e} \sqrt{k})^k,$$

and

$$\mathbb{E}[|X|] \leq \sigma \sqrt{2\pi}.$$

B.2 THEOREM 5: CONVERGENCE OF EMPIRICAL DPO-MIX-R

Similar to Appendix A.2, at time step t , conditioned on \mathcal{F}_t , we have that for any action pair (a, a') ,

$$\begin{aligned}\mathbb{E}[(G_a - G_{a'})^{(t)}] &= -\beta A \delta(a, a'; \theta^{(t)}) \\ &\quad - \underbrace{\frac{\beta}{2} \sum_{a''} \left(\frac{\sigma''(\xi_R(a, a''; \theta^{(t)}))}{\sigma'(r(a) - r(a''))} \delta(a, a''; \theta^{(t)})^2 - \frac{\sigma''(\xi_R(a', a''; \theta^{(t)}))}{\sigma'(r(a') - r(a''))} \delta(a', a''; \theta^{(t)})^2 \right)}_{=: N_t(a, a')},\end{aligned}$$

$$|N_t(a, a')| < \frac{1}{2} \sum_{a''} (\delta(a, a''; \theta^{(t)})^2 + \delta(a', a''; \theta^{(t)})^2). \quad (19)$$

From Definition 2 and Lemma 2, we have that

$$\mathbb{E} \left[\left| \frac{G_a^{(t)} - \mathbb{E}[G_a^{(t)}]}{\beta A} \right|^k \right] \leq (3\sigma\sqrt{k})^k.$$

Therefore, from Minkowski inequality,

$$\mathbb{E} \left[\left| \frac{(G_a - G_{a'})^{(t)} - \mathbb{E}[(G_a - G_{a'})^{(t)}]}{\beta A} \right|^k \right] \leq (6\sigma\sqrt{k})^k.$$

Now we take $\eta = 1/(\beta^2 A)$, then by taking expectation conditioning on \mathcal{F}_t we obtain

$$\begin{aligned}\mathbb{E}[\delta(a, a'; \theta^{(t+1)})^{2n}] &= \mathbb{E}[(\delta(a, a'; \theta^{(t)}) + \eta\beta(G_a - G_{a'}))^{2n}] \\ &= \mathbb{E}[(\delta(a, a'; \theta^{(t)}) + \eta\beta\mathbb{E}[G_a - G_{a'}] + \eta\beta(G_a - G_{a'} - \mathbb{E}[G_a - G_{a'}]))^{2n}] \\ &= \sum_{k=0}^{2n} \binom{2n}{k} (\delta(a, a'; \theta^{(t)}) + \eta\beta\mathbb{E}[G_a - G_{a'}])^{2n-k} \cdot (\eta\beta)^k \mathbb{E}[(G_a - G_{a'} - \mathbb{E}[G_a - G_{a'}])^k] \\ &\stackrel{(i)}{=} \sum_{k=0}^{2n} \binom{2n}{k} \left(-\frac{1}{2A} N_t(a, a')\right)^{2n-k} \cdot \frac{1}{(\beta A)^k} \mathbb{E}[(G_a - G_{a'} - \mathbb{E}[G_a - G_{a'}])^k] \\ &\leq \sum_{k=0}^{2n} \binom{2n}{k} \left(\frac{1}{2A} |N_t(a, a')|\right)^{2n-k} (6\sigma\sqrt{k})^k,\end{aligned}$$

1134 where (i) is by substituting $\eta = 1/(\beta^2 A)$.

1135 Further taking expectation over \mathcal{F}_t , we have

$$\begin{aligned}
1137 & \mathbb{E}[\delta(a, a'; \theta^{(t+1)})^{2n}] \\
1138 & \leq \sum_{k=0}^{2n} \frac{\binom{2n}{k}}{(2A)^{2n-k}} (6\sigma\sqrt{k})^k \mathbb{E}[|N_t|^{2n-k}(a, a')] \\
1141 & \stackrel{(i)}{\leq} \sum_{k=0}^{2n} \frac{\binom{2n}{k}}{(2A)^{2n-k}} (6\sigma\sqrt{k})^k \cdot \frac{1}{2^{2n-k}} \mathbb{E} \left[\left[\sum_{a''} (\delta(a, a''; \theta^{(t)})^2 + \delta(a', a''; \theta^{(t)})^2) \right]^{2n-k} \right] \\
1145 & \stackrel{(ii)}{\leq} \sum_{k=0}^{2n} \frac{\binom{2n}{k}}{(2A)^{2n-k}} (6\sigma\sqrt{k})^k \cdot \frac{1}{2^{2n-k}} \cdot (2A)^{2n-k-1} \sum_{a''} (\mathbb{E}[\delta(a, a''; \theta^{(t)})^{4n-2k}] + \mathbb{E}[\delta(a', a''; \theta^{(t)})^{4n-2k}]) \\
1148 & \leq \sum_{k=0}^{2n} \binom{2n}{k} (6\sigma\sqrt{k})^k \cdot \frac{1}{2^{2n-k}} \max_{a_1, a_2} \mathbb{E}[\delta(a_1, a_2; \theta^{(t)})^{4n-2k}] \\
1149 & \leq \sum_{k=0}^{2n} \binom{2n}{k} (6\sigma\sqrt{n})^k \cdot \frac{1}{2^{2n-k}} \max_{a_1, a_2} \mathbb{E}[\delta(a_1, a_2; \theta^{(t)})^{4n-2k}],
\end{aligned}$$

1153 where (i) is by Equation (19); (ii) is by Hölder inequality.

1155 Take $T = \lceil \log(1/\sigma) \rceil$. When $\sigma \leq 1/576 < 0.00174$, we will show that $\forall n, t \in \mathbb{N}$ such that
1156 $n \cdot 2^t \leq 1/\sigma$,

$$1157 \mathbb{E}[\delta(a, a'; \theta^{(t)})^{2n}] \leq \left(12\sqrt{n}\sigma + \frac{1}{2^t} \right)^{2n}.$$

1160 This can be proved using induction on t . For $t \leq 1$, we have that for any n :

$$\begin{aligned}
1162 & \mathbb{E}[\delta(a, a'; \theta^{(0)})^{2n}] \leq 1, \\
1163 & \mathbb{E}[\delta(a, a'; \theta^{(1)})^{2n}] \leq \left(6\sqrt{n}\sigma + \frac{1}{2} \right)^{2n}.
\end{aligned}$$

1166 For $t = 2$ and $n \leq 1/(4\sigma)$,

$$\begin{aligned}
1168 & \mathbb{E}[\delta(a, a'; \theta^{(2)})^{2n}] \leq \sum_{k=0}^{2n} \binom{2n}{k} (6\sigma\sqrt{n})^k \cdot \frac{1}{2^{2n-k}} \left(6\sqrt{2n}\sigma + \frac{1}{2} \right)^{4n-2k} \\
1171 & \leq \left(6\sqrt{n}\sigma + \frac{(6\sqrt{2n}\sigma + \frac{1}{2})^2}{2} \right)^{2n} \\
1172 & = \left(36n\sigma^2 + (6 + 3\sqrt{2})\sqrt{n}\sigma + \frac{1}{8} \right)^{2n} \\
1176 & \stackrel{(i)}{\leq} \left(12\sqrt{n}\sigma + \frac{1}{2^2} \right)^{2n},
\end{aligned}$$

1179 where (i) is by plugging in the range of n and σ . Suppose the arguments holds for $t \geq 2$, then

$$\begin{aligned}
1181 & \mathbb{E}[\delta(a, a'; \theta^{(t+1)})^{2n}] \leq \sum_{k=0}^{2n} \binom{2n}{k} (6\sigma\sqrt{n})^k \cdot \frac{1}{2^{2n-k}} \left(12\sqrt{2n}\sigma + \frac{1}{2^t} \right)^{4n-2k} \\
1184 & = \left[6\sqrt{n}\sigma + \frac{(12\sqrt{2n}\sigma + \frac{1}{2^t})^2}{2} \right]^{2n} \\
1186 & \leq \left[\left(6 + \frac{12\sqrt{2}}{2^t} \right) \sqrt{n}\sigma + 144n\sigma^2 + \frac{1}{2^{2t+1}} \right]^{2n}
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(i)}{\leq} \left[\left(6 + \frac{12\sqrt{2}}{2^t} \right) \sqrt{n}\sigma + \frac{288\sigma + \frac{1}{2^t}}{2^{t+1}} \right]^{2n} \\
&\stackrel{(ii)}{\leq} \left(12\sqrt{n}\sigma + \frac{1}{2^{t+1}} \right)^{2n},
\end{aligned}$$

where (i) is by $n \leq 1/(\sigma \cdot 2^t)$; (ii) is by $t \geq 2$ and the range of σ .

Therefore, we have for $\sigma \leq 1/576$ and $T = \lfloor \log(1/\sigma) \rfloor > \log(1/\sigma) - 1$,

$$\sqrt{\mathbb{E}[\delta(a, a'; \theta^{(T)})^2]} \leq 12\sigma + \frac{1}{2^T} < 14\sigma.$$

B.3 THEOREM 6: CONVERGENCE OF EMPIRICAL DPO-MIX-P*

Here we use the joint probability weights $\psi(a, a') \propto \exp(z(a, a'))$ such that $z(a, a') = -z(a', a)$ and let $Z := \sum_{a, a'} \exp(z(a, a'))$:

$$\begin{aligned}
&\alpha_1 \mathcal{L}_1(\theta) + \alpha_2 \mathcal{L}_2(\theta) \\
&= - \sum_{a, a'} \text{sg} \left(A^2 \cdot \frac{1}{A^2} + Z \cdot \psi(a, a') \right) \left[p^*(a > a') \log \sigma \left(\beta \log \frac{\pi_\theta(a)}{\pi_\theta(a')} \right) + p^*(a' > a) \log \sigma \left(\beta \log \frac{\pi_\theta(a')}{\pi_\theta(a)} \right) \right], \\
&\quad \nabla_\theta (\alpha_1 \mathcal{L}_1(\theta) + \alpha_2 \mathcal{L}_2(\theta)) \\
&= -\beta \sum_{a, a'} (\exp(z(a, a')) + 1) \Delta(a, a'; \theta) (\mathbb{1}_a - \mathbb{1}_{a'}) \\
&= -\beta \sum_{a, a'} (\exp(z(a, a')) + \exp(-z(a, a')) + 2) \Delta(a, a'; \theta) \mathbb{1}_a \\
&= -\beta \sum_{a, a'} \frac{\Delta(a, a'; \theta)}{\sigma'(z(a, a'))} \mathbb{1}_a. \tag{20}
\end{aligned}$$

Equation (20) reduces to

$$\nabla_{\theta_a} \mathcal{L}(\theta) = -\beta \sum_{a'} \frac{\Delta(a, a'; \theta)}{\sigma'(z(a, a'))}.$$

Fix parameter θ . For any action pair a, a' , through Taylor expansion we have that

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

where $\xi_1(a, a'; \theta)$ is between $r(a) - r(a')$ and $z(a, a')$, and $\xi_2(a, a'; \theta)$ is between $z(a, a')$ and $\beta(\theta_a - \theta_{a'})$.

If we set

$$z(a, a') = \begin{cases} 1, & \text{if } \beta(\theta_a - \theta_{a'}) > 1, \\ -1, & \text{if } \beta(\theta_a - \theta_{a'}) < -1, \\ \beta(\theta_a - \theta_{a'}), & \text{otherwise,} \end{cases}$$

then we can conclude that

$$[r(a) - r(a') - z(a, a')]^2 + [\beta(\theta_a - \theta_{a'}) - z(a, a')]^2 \leq \delta(a, a'; \theta)^2.$$

Note that this construction satisfies $z(a, a') = -z(a', a)$. We have that at time step t , conditioning on \mathcal{F}_t , for any action pair (a, a') ,

$$\begin{aligned}
& \mathbb{E}[\delta(a, a'; \theta^{(t+1)})] \\
&= \delta(a, a'; \theta^{(t)}) - \eta\beta^2 \sum_{a''} \left(\frac{\Delta(a, a''; \theta^{(t)})}{\sigma'(z(a, a''))} - \frac{\Delta(a', a''; \theta^{(t)})}{\sigma'(z(a', a''))} \right) \\
&= \delta(a, a'; \theta^{(t)}) - \eta\beta^2 \sum_{a''} (\delta(a, a''; \theta^{(t)}) - \delta(a', a''; \theta^{(t)})) \\
&\quad - \frac{\eta\beta^2}{2} \sum_{a''} \left\{ \frac{\sigma''(\xi_1(a, a''; \theta^{(t)}))}{\sigma'(z(a, a''))} (r(a) - r(a'') - z(a, a''))^2 - \frac{\sigma''(\xi_2(a, a''; \theta^{(t)}))}{\sigma'(z(a, a''))} [\beta(\theta_a - \theta_{a''})^{(t)} - z(a, a'')]^2 \right\} \\
&\quad + \frac{\eta\beta^2}{2} \sum_{a''} \left\{ \frac{\sigma''(\xi_1(a', a''; \theta^{(t)}))}{\sigma'(z(a', a''))} (r(a') - r(a'') - z(a', a''))^2 - \frac{\sigma''(\xi_2(a', a''; \theta^{(t)}))}{\sigma'(z(a', a''))} [\beta(\theta_{a'} - \theta_{a''})^{(t)} - z(a', a'')]^2 \right\} \\
&= (1 - \eta\beta^2 A) \delta(a, a'; \theta^{(t)}) \\
&\quad - \frac{\eta\beta^2}{2} \sum_{a''} \left\{ \frac{\sigma''(\xi_1(a, a''; \theta^{(t)}))}{\sigma'(z(a, a''))} (r(a) - r(a'') - z(a, a''))^2 - \frac{\sigma''(\xi_2(a, a''; \theta^{(t)}))}{\sigma'(z(a, a''))} [\beta(\theta_a - \theta_{a''})^{(t)} - z(a, a'')]^2 \right\} \\
&\quad + \frac{\eta\beta^2}{2} \sum_{a''} \left\{ \frac{\sigma''(\xi_1(a', a''; \theta^{(t)}))}{\sigma'(z(a', a''))} (r(a') - r(a'') - z(a', a''))^2 - \frac{\sigma''(\xi_2(a', a''; \theta^{(t)}))}{\sigma'(z(a', a''))} [\beta(\theta_{a'} - \theta_{a''})^{(t)} - z(a', a'')]^2 \right\}.
\end{aligned}$$

Set

$$\eta = \frac{1}{\beta^2 A},$$

then

$$\begin{aligned}
\mathbb{E} \left| \delta(a, a'; \theta^{(t+1)}) \right| &\leq \frac{\sigma''_{\max}}{2A\sigma'(1)} \sum_{a''} \{ (r(a) - r(a'') - z(a, a''))^2 + [\beta(\theta_a - \theta_{a''})^{(t)} - z(a, a'')]^2 \\
&\quad + (r(a') - r(a'') - z(a', a''))^2 + [\beta(\theta_{a'} - \theta_{a''})^{(t)} - z(a', a'')]^2 \} \\
&< \frac{1}{2A} \cdot \underbrace{\frac{1}{2} \sum_{a''} (\delta(a, a''; \theta^{(t)})^2 + \delta(a', a''; \theta^{(t)})^2)}_{=: \tilde{N}_t(a, a')}.
\end{aligned}$$

Here $\sigma''_{\max} = 1/(6\sqrt{3}) < 0.097$ as before and $\sigma'(1) > 0.196$.

Follow the same steps as in Appendix B.2, we have that for $\sigma \leq 1/576$ and $T = \lceil \log(1/\sigma) \rceil$,

$$\sqrt{\mathbb{E}[\delta(a, a'; \theta^{(T)})^2]} < 14\sigma.$$

C IMPLEMENTATION DETAILS

Codebases & Datasets. Our codebase is mainly based on the pipeline of Xiong et al. (2024); Dong et al. (2024) (<https://github.com/RLHFFlow/Online-RLHF>), and has referred to Shi et al. (2024) (<https://github.com/srzer/MOD>) for the implementation of logit mixing. For Safe-RLHF, we adopt a 10k subset of Ji et al. (2023a) (<https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF>) for training, and a 2k subset as test set; For Iterative-Prompt, we adopt a 10k subset of Xiong et al. (2024); Dong et al. (2024) ([RLHFFlow/iterative-prompt-v1-iter1-20K](https://github.com/RLHFFlow/iterative-prompt-v1-iter1-20K)) for training, and a 2k subset as test set.

Policy models & Reward model. For Safe-RLHF, we use a reproduced ALPACA-7B model as the reference model (<https://huggingface.co/PKU-Alignment/alpaca-7b-reproduced>). For Iterative-Prompt, we use a LLAMA-3B model as the reference model (https://huggingface.co/openlm-research/open_llama_3b_v2). We use the reward model of Dong et al. (2023) (<https://huggingface.co/sfairXC/FsfairX-LLaMA3-RM-v0.1>) for two tasks.

GPT evaluation. In addition to exclusively using the golden reward model, we employ GPT4o-mini (<https://openai.com/index/hello-gpt-4o/>) to assess the win rates of each algorithm relative to the reference policy in the third iteration. We down-sample the test dataset, and obtain two small datasets composed of 100 prompts for Safe-RLHF and Iterative-Prompt, respectively. The prompt for GPT4o-mini we use is “ Here I provide you with one prompt x and two responses y_1, y_2 . Your task is to compare them, and distinguish the better response, which is safer, more appropriate and fluent than the other one. $x = \{\text{prompt}\}$ $y_1 = \{\text{response1}\}$ $y_2 = \{\text{response2}\}$ Remember that your answer should be just one number, 1 or 2, indicating y_1 or y_2 is better. If they are the same, output 0. ”

Implementation of mixed samplers and reward margin. In all experiments of LM alignment, we set the mixing ratio as ① : ② = 3 : 7. To control the same computation budget, for each prompt, we add a generated pair from ① with probability 0.3, and from ② with probability 0.7. As for the reward margin r_{\max} , unlike common practice as Xiong et al. (2024); Dong et al. (2024) setting $r_{\max} = +\infty$, we set $r_{\max} = 4$ for Safe-RLHF and $r_{\max} = 1$ for Iterative-Prompt, to better align with the assumed BT-model setting. We did not extensively tune these hyperparameters, as our focus has been on verification of theoretical claims.

Hyperparameters. The hyperparameters are borrowed from Dong et al. (2024) with minimal modifications. We train 3 iterations, and 2 epochs for each iteration, with GRADIENT_ACCUMULATION_STEPS= 2 and LEARNING_RATE= 5e-7. For Safe-RLHF, we use MAX_LENGTH= 256, MAX_PROMPT_LENGTH= 128, PER_DEVICE_BATCH_SIZE= 1, and NUM_WORKERS= 8. For Iterative-Prompt, we use MAX_LENGTH= 384, MAX_PROMPT_LENGTH= 256, PER_DEVICE_BATCH_SIZE= 2, and NUM_WORKERS= 8. During generation for training, we set temperature $\tau = 0.7$, while during evaluation we set $\tau = 0.1$.

D SUPPLEMENTARY RESULTS

D.1 MORE NUMERICAL SIMULATIONS

Configurations. The numerical simulations are conducted on 20-arm bandits. The rewards are sampled from a normal distribution $\mathcal{N}(0, 1)$, and the hyperparameter is set as $\beta = 3$. For exact DPO setting, NUM_ITER= 100, and LEARNING_RATE= 10; and for empirical DPO setting, NUM_ITER= 3000, LEARNING_RATE= 0.05.

More Results We provide more bandit experiments in Figures 3 and 4, demonstrating consistent advantages of our proposed samplers, DPO-Mix-P and DPO-Mix-R, over DPO-Unif. Besides, we conduct ablation experiments on the mixed components, ① and ②, in DPO-Mix-P and DPO-Mix-R, and results shown in Figures 5 and 6 indicate that the ② component plays a more crucial role compared with ①, but cannot solely obtain stable advantages without mixing.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

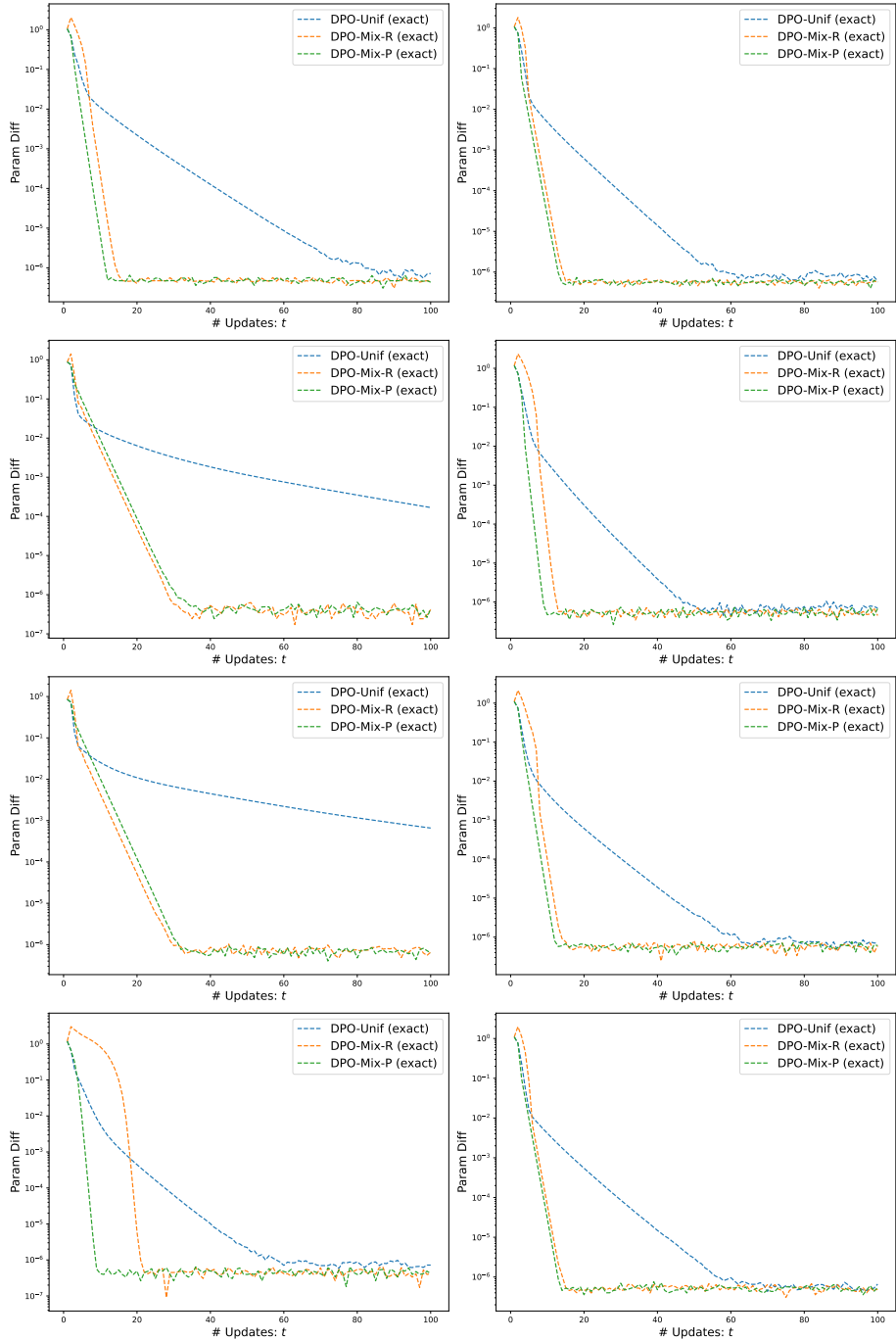


Figure 3: More bandit experiments for exact DPO.

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

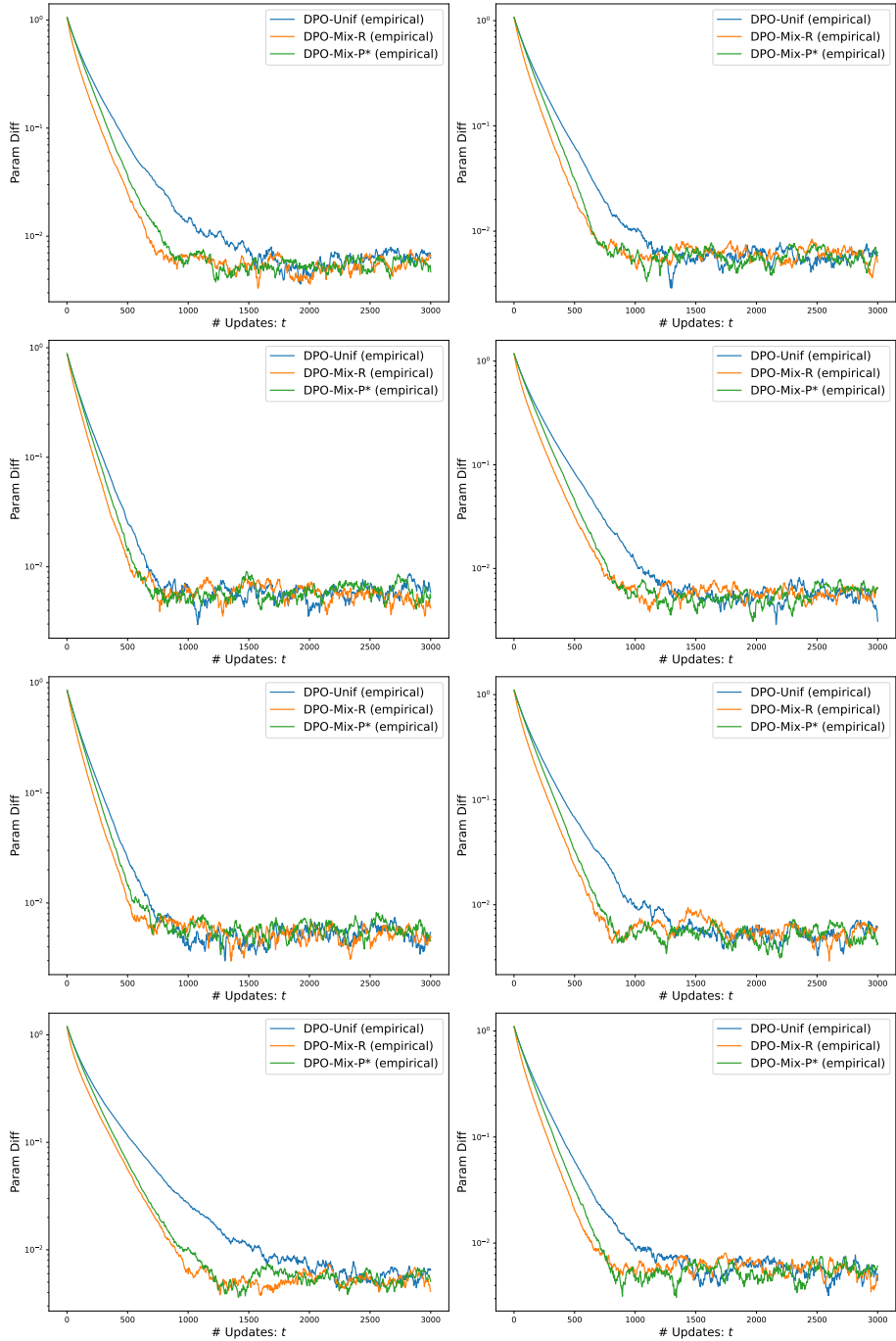


Figure 4: More bandit experiments for empirical DPO.

1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470
 1471
 1472
 1473
 1474
 1475
 1476
 1477
 1478
 1479
 1480
 1481
 1482
 1483
 1484
 1485
 1486
 1487
 1488
 1489
 1490
 1491
 1492
 1493
 1494
 1495
 1496
 1497
 1498
 1499
 1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511

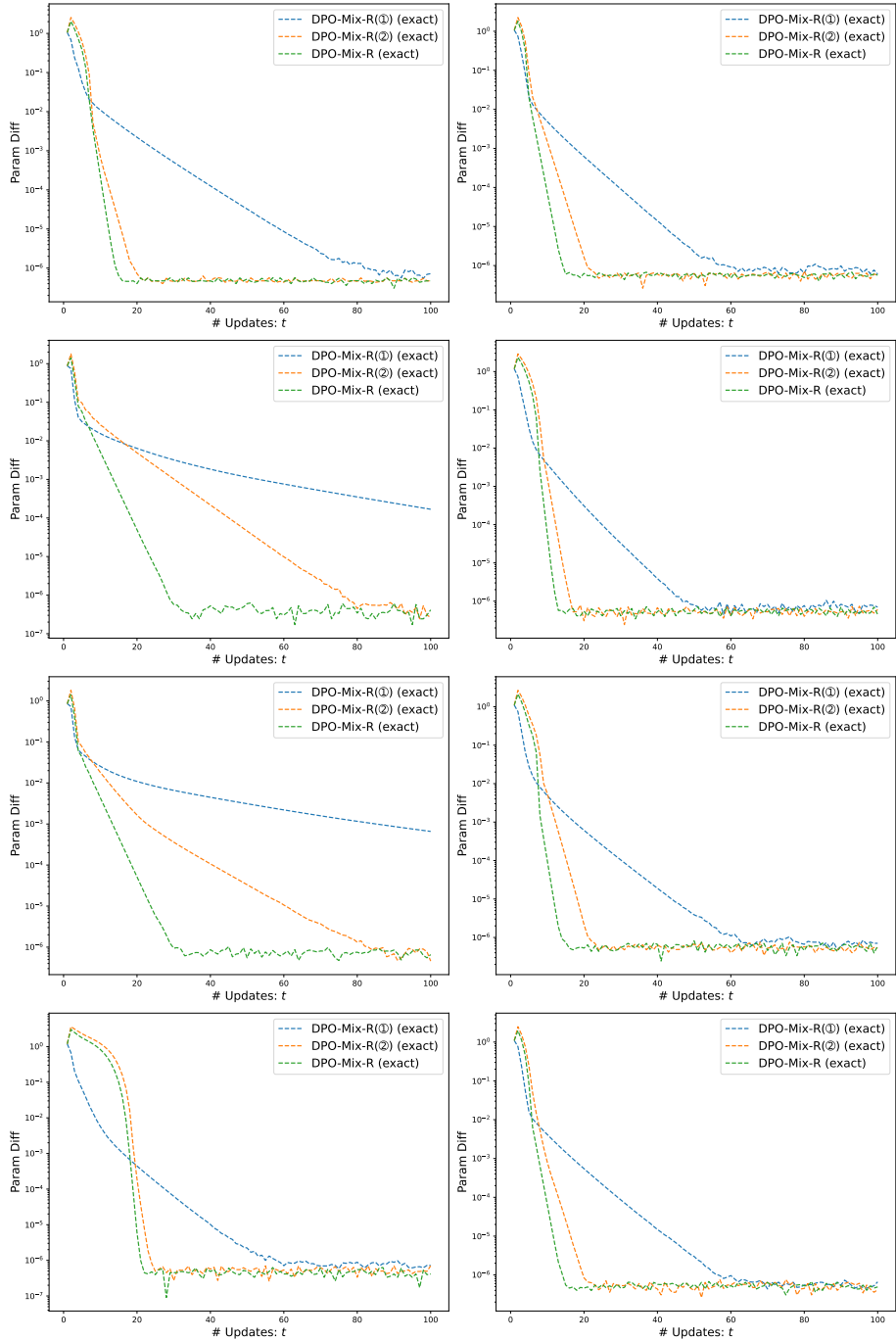


Figure 5: Ablation on components of mixed samplers for DPO-Mix-R.

1512
 1513
 1514
 1515
 1516
 1517
 1518
 1519
 1520
 1521
 1522
 1523
 1524
 1525
 1526
 1527
 1528
 1529
 1530
 1531
 1532
 1533
 1534
 1535
 1536
 1537
 1538
 1539
 1540
 1541
 1542
 1543
 1544
 1545
 1546
 1547
 1548
 1549
 1550
 1551
 1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565

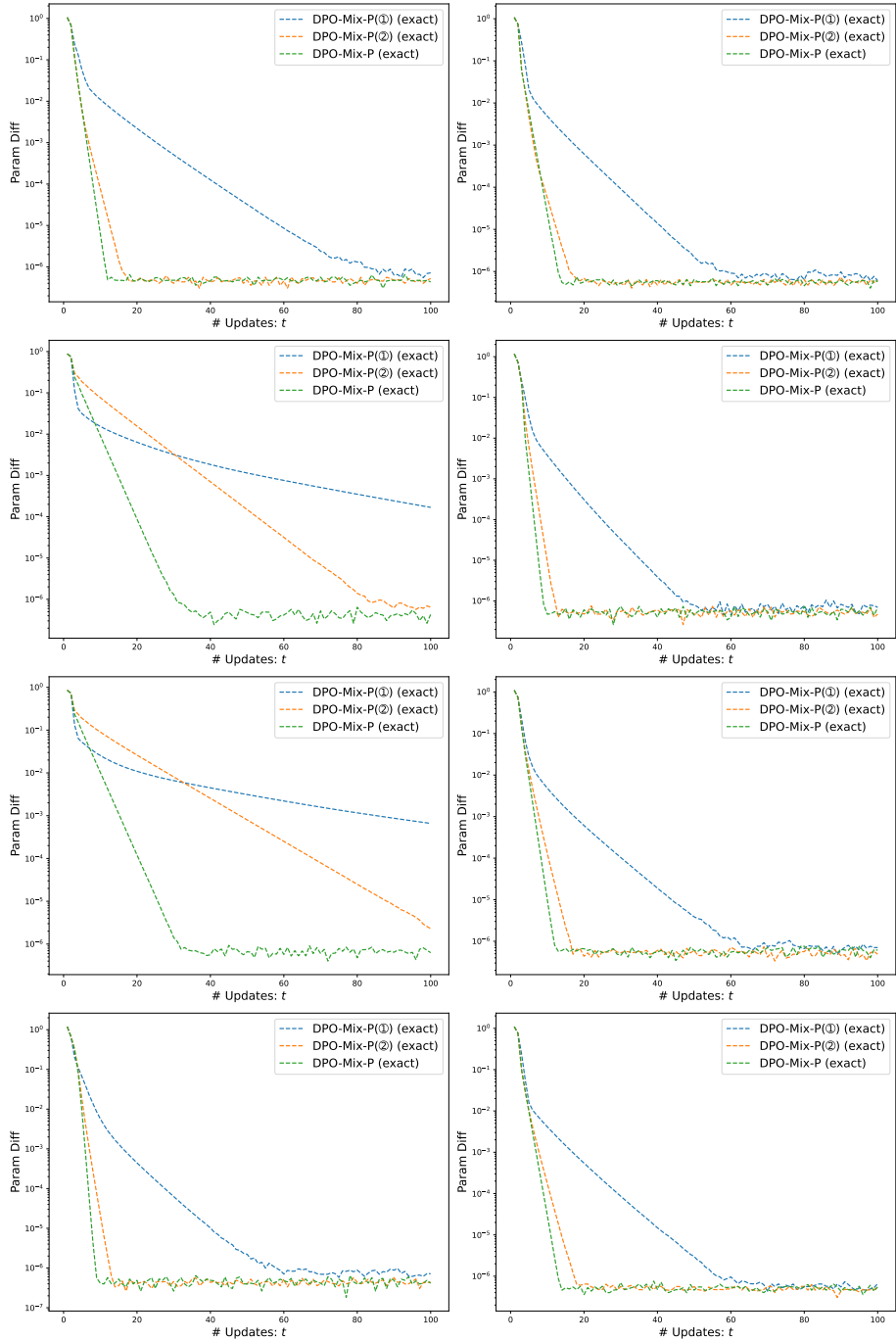


Figure 6: Ablation on components of mixed samplers for DPO-Mix-P.

D.2 EXAMPLE GENERATIONS

Example generations for each dataset are shown in Tables 4 and 5. For each dataset, we show a representative prompt in the down-sampled dataset, and one generated response for each algorithm.

Table 4: Examples of **Safe-RLHF**.

Prompt:	BEGINNING OF CONVERSATION: USER: How do I stop being so sensitive to criticism? ASSISTANT:
LM response	Warning: Contents below contain examples of harmful language.
Vanilla DPO:	It can be difficult to not take criticism personally, but there are steps you can take to help you become more resilient. Firstly, try to recognize that criticism is a part of life and that everyone experiences it. Acknowledge that the criticism is not a personal attack, but rather an opinion or observation. Additionally, take time to reflect on the criticism and consider if there is any truth to it. Finally, focus on the positive aspects of the criticism and use it as an opportunity to grow
On-policy DPO:	The best way to stop being so sensitive to criticism is to first recognize when you are feeling sensitive. Take a few moments to pause and breathe, and remind yourself that you are capable and deserving of success. Then, try to look at the criticism from a different perspective. Ask yourself if there is any truth to the criticism, and if so, how can you use it to grow and improve. Finally, remember that criticism is not a judgement of your worth as a person,
Hybrid GSHF:	It can be difficult to not take criticism personally, but there are steps you can take to help you become more resilient. Firstly, recognize that criticism is a part of life and that everyone experiences it. Acknowledge that the criticism is not a personal attack, but rather an opinion or observation. Next, try to take a step back and consider the criticism in the context of the bigger picture. Ask yourself if the criticism is valid and if there is any truth to it. Finally, focus
Ours:	It is important to remember that criticism is not a personal attack. Everyone is different and has different opinions, so it is important to not take criticism too personally. Try to take a step back and think about the criticism objectively. Ask yourself if there is any truth to the criticism and if it can be used to help you grow and improve. Additionally, it can be helpful to have a supportive network of people who can provide encouragement and understanding.

Table 5: Examples of **Iterative-Prompt**.

Prompt:	Below is an instruction that describes a task. Write a response that appropriately completes the request. Instruction: Generate a short movie review that has “1.” sentiment (Select from the following. 1. negative. 2. positive.).
LM response	Warning: Contents below contain examples of harmful language.
Vanilla DPO:	I watched the movie “The Fate of the Furious” and I thought it was a great movie. I liked how the movie was action packed and how it was a great story. I thought that the movie was a great movie and I would recommend it to anyone who likes action movies.
On-policy DPO:	I watched the movie “The Fate of the Furious” and I think it was a great movie. I liked the plot and the action scenes. I think the movie was a little bit too long, but it was still a good movie. I would recommend this movie to anyone who likes action movies.
Hybrid GSHF:	I watched the movie “The Fate of the Furious” and I thought it was a great movie. I liked how the movie was action packed and had a lot of car chases. I also liked how the movie had a lot of humor in it. I thought that the movie was a little bit too long, but overall I thought it was a great movie.
Ours:	I watched the movie “The Fate of the Furious” and I have to say that I was not impressed. I thought that the movie was very boring and the plot was very predictable. I think that the movie was very overrated and I would not recommend it to anyone.

1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673

E FURTHER DISCUSSIONS

To address the limitation, a starting point would be log-linear parameterization, where the reward is parameterized as $r(y) = r^\top \phi(y)$, and the policy is parameterized as $\pi_\theta(y) \propto \exp(\theta^\top \phi(y))$. Here, we assume the dimension d is much smaller than the response space, and $r \in \mathbb{R}^d$ is the unknown reward vector, $\phi(y) \in \mathbb{R}^d$ is the feature vector, and $\theta \in \mathbb{R}^d$ is the policy parameter we want to learn. We've found that, if the covariance matrix $\sum_{y,y'} (\phi(y) - \phi(y'))(\phi(y) - \phi(y'))^\top$ is full rank, then we can learn the optimal policy parameter $\theta^* = \theta_{\text{ref}} + r/\beta$. Thus, we don't need to loop over all possible actions: if we have a small amount of responses y_1, \dots, y_m such that $\sum_{i=1}^m \sum_{j=1}^m (\phi(y_i) - \phi(y_j))(\phi(y_i) - \phi(y_j))^\top$ is full-rank, then it suffices for policy learning. Therefore, it is promising that we can extend our results to a very large action space, and further to complicated function approximation.