EFFICIENT EXPLORATION FOR ITERATIVE NASH PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference alignment is central to improving large language models, but standard reward-based formulations can be restrictive when human preferences are cyclic, non-transitive. Nash Learning from Human Feedback (NLHF) addresses this limitation by modeling alignment as a preference game and targeting a Nash equilibrium. However, the learning-theoretic foundations of scalable NLHF remain limited. Existing regret guarantees requires training a general preference model and solving minimax oracles, while iterative NLHF methods are easier to implement but lack regret guarantees. In this paper, we study online iterative NLHF and identify exploration as the key obstacle. First, we show that standard iterative NLHF can suffer an exponential dependence on the KL-regularization parameter, revealing that implicit exploration through policy updates is insufficient for efficient exploration. Second, we propose Exploratory Nash Policy Optimization (ENPO) that combines SFT-based regularization with adversarial policy exploration. ENPO eliminates the *exponential dependency* without using any minimax oracles and explicit preference model estimations. Additionally, we present another theoretical algorithm: *Bonus-explorer ENPO* (BENPO) with an explicit exploration mechanism, which uses additionally oracles but can guarantee an $O(\log T)$ regret bound. A practical algorithm: *Direct ENPO* (DENPO) was implemented for LLM fine-tuning. The empirical demonstrations show that DENPO consistently outperforms existing similar RLHF and NLHF baselines.

1 INTRODUCTION

Large language models (LLMs) are increasingly trained to act according to human preferences, typically through *reinforcement learning from human feedback* (RLHF) (Christiano et al., 2017; Stiennon et al., 2020). Standard RLHF first fits a scalar reward model from pairwise comparisons and then optimizes a policy against that reward. This reduction has been empirically successful (Ziegler et al., 2019; Ouyang et al., 2022; Touvron et al., 2023; Achiam et al., 2023), but it can be restrictive when preferences are heterogeneous, cyclic, or otherwise not well described by a single Bradley-Terry (BT) reward model (Bradley & Terry, 1952). Game-theoretic alignment methods, including *Nash learning from human feedback* (NLHF) (Munos et al., 2024) and its variants (Chakraborty et al., 2024; Swamy et al., 2024; Ye et al., 2024; Rosset et al., 2024; Wu et al., 2025b; Zhang et al., 2025b; Wu et al., 2025a; 2026; Lee et al., 2026), instead treat pairwise preference as the payoff of a game and seek a Nash equilibrium (NE) policy.

Despite growing interest in game-theoretic alignment (Bai et al., 2022; Cui et al., 2024), its online learning foundations remain limited. Most existing analyses assume the preference game is fixed and known, and study how quickly equilibrium dynamics compute a Nash equilibrium (Munos et al., 2024; Wu et al., 2025b; Zhang et al., 2025a;b; Wang et al., 2025; Zhou et al., 2025; Tiapkin et al., 2025). This optimization view is useful, but it abstracts away a central difficulty in LLM alignment: the payoff is revealed only through costly, learner-chosen preference queries. Each update determines which responses are sampled, which opponents are compared, and therefore which parts of the game become statistically visible. Online NLHF is thus a coupled optimization and coverage problem. Existing learning results are comparatively scarce (Ye et al., 2024; Wu et al., 2025a; Lee et al., 2026). Among them, Ye et al. (2024) obtains an $O(\sqrt{T})$ regret bound without exponential dependence on the KL-regularization parameter, but requires solving minimax oracles and *explicit estimation* of a general preference model. In contrast, iterative NLHF methods (Zhang et al., 2025b) are simpler and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

more scalable because they update the policy directly through preference losses, but they currently lack regret guarantees. This leaves a central question: *can iterative NLHF retain its algorithmic simplicity while achieving provable regret guarantees?*

A key obstacle is exploration. Recent work on iterative direct preference optimization (DPO) under the BT model shows that purely on-policy sampling can suffer an *exponential dependence* on the KL-regularization modulus when high-quality responses are rarely sampled by the reference policy (Xie et al., 2025). This suggests that explicit exploration is necessary in online alignment. However, extending this idea from DPO to NLHF is highly nontrivial. The analysis of exploratory DPO relies on a scalar reward representation and the interpretation of DPO as implicit $Q^\star$-approximation (Rafailov et al., 2024). In NLHF, preferences may be non-transitive and need not admit any scalar reward representation. Moreover, the goal is not to maximize a reward, but to minimize a duality gap against *adaptive* opponent policies, which makes the equilibrium learning much harder. Thus, exploration in NLHF must cover not only informative responses, but also the opponent distributions that reveal poorly learned regions of the preference game.

To achieve this goal, we augment DPO in iterative NLHF with two exploration mechanisms. First, we introduce a *pessimistic* supervised fine-tuning-like (SFT) regularization term into the iterative preference loss. Unlike the optimism term used for exploratory DPO under the BT model (Xie et al., 2025), our term acts as a stabilizing regularizer: it strengthens the mirror-descent structure of the update and prevents exploration from destabilizing policy improvement. Second, we construct an adversarial policy maximization procedure that identifies opponent policies exposing poorly covered regions of the preference game, and we sample completions from these policies. This targeted exploration directly addresses the opponent distributions appearing in the duality gap, enabling regret control without requiring full preference-model training at every iteration.

Contributions. We propose an explicitly exploratory iterative NLHF method with provable regret guarantees. Our contributions can be summarized as follows:

1. First, we show that standard iterative NLHF can fail to cover adversarially informative responses and may incur an *exponential dependence* on the KL-regularization parameter. In contrast, we show that by combining SFT-based regularization with adversarial policy exploration, our method *Exploratory Nash Preference Optimization* (ENPO) can eliminate such an exponential dependence, while avoiding minimax oracles and preference model training. Additionally, we present another theoretical algorithm: *Bonus-explorer ENPO* (BENPO) with an explicit exploration mechanism, which requires additionally oracles but guarantees an $O(\log T)$ regret bound.
2. Second, we propose a practical algorithm that uses a surrogate to implement the adversarial policy exploration in ENPO, called *Direct ENPO* (DENPO). We implement DENPO to do post-training on `Llama-3-8B-Instruct` and evaluate it across various benchmarks. Experimental results show that DENPO consistently outperforms existing RLHF and NLHF baselines at matched computational cost, demonstrating its potential to improve the performance of NLHF alignment methods. This empirical evidence also highlights that our explicit policy exploration mechanism is useful not only theoretically but also in practical LLM alignment.

Related work. Our work is connected to the literature on game-theoretic preference alignment and online exploration. Due to space limitations, we defer our comments on other relevant topics to Appendix A. Recent works study game-theoretic alignment under general preference models. Early methods solve KL-regularized two-player minimax problems. In particular, NLHF learns a pairwise preference model and applies mirror descent against a geometric mixture of the current and reference policies (Munos et al., 2024), while exploratory NLHF provides offline pessimistic and online optimistic variants using conservative estimates or enhancer policies (Ye et al., 2024). MaxMin-RLHF addresses diverse users by learning subpopulation-specific rewards via EM and updating toward the currently worst-aligned group (Chakraborty et al., 2024). However, these methods involve equilibrium solving, which may be hard to stabilize in practice. More recent approaches avoid such oracles through iterative self-play: SPO uses one no-regret learner with on-policy win-rate rewards (Swamy et al., 2024), SPPO updates the policy according to empirical win rates (Wu et al., 2025b), and INPO optimizes a preference-pair square loss without explicit per-response win-rate estimation (Zhang et al., 2025b). Greedy sampling analyses show that simple empirical updates

can be statistically efficient under KL-regularized general preferences and generalized bilinear preferences (Wu et al., 2025a; Lee et al., 2026). MNPO extends previous works to multiplayer alignment with historical or heterogeneous opponents (Wu et al., 2026).

Online exploration was developed in reinforcement learning theory to exploit interactive feedback by deliberately sampling informative and diverse behaviors. Modern exploration methods provide algorithmic principles that adapt to problem structure and achieve near-optimal regret bounds (Jiang et al., 2017; Foster & Rakhlin, 2023). However, these approaches are often computationally intractable in general RL settings (Jiang et al., 2017; Jin et al., 2021a; Foster et al., 2021). In preference alignment, practical approaches either collect feedback without updating the policy model (Dwaracherla et al., 2024) or rely on passive on-policy sampling (Guo et al., 2024; Gao et al., 2024; Zhang et al., 2025b), which provides no explicit mechanism for novelty or diversity and can fail to discover rare high-quality responses. Exploratory approaches either impose restrictive assumptions (Novoseller et al., 2020; Xu et al., 2020; Saha et al., 2023; Wu & Sun, 2024; Zhan et al., 2024; Du et al., 2024; Das et al., 2025) or require computationally expensive procedures that are difficult to implement at LLM scale (Chen et al., 2022; Wang et al., 2023; Ye et al., 2024). Recently, Xie et al. (2025) have taken a first step toward simple and efficient exploration in RLHF under the BT reward model. Our work adapts the principle to online iterative NLHF under general preference models in LLM fine-tuning.

2 Preliminaries and technical background

Game-theoretic preference alignment. Modern LLMs are designed based on Transformer architectures (Vaswani et al., 2017): the input is a user prompt $\mathbf{x} \in \mathcal{V}^\star$ and a response $\mathbf{y} \in \mathcal{V}^\star$ is generated, where $\mathcal{V}^\star$ is the set of all finite-length sequences of tokens drawn from the vocabulary set $\mathcal{V}$. We cast an LLM as a policy $\pi_\theta(\mathbf{y}|\mathbf{x})$ which assigns a probability to each $\mathbf{y}$ given $\mathbf{x}$. Given a base LLM, the generated responses might not be helpful, safe, or reliable. Thus, techniques for aligning the LLM with human preferences are applied.

We assume access to a general preference oracle $P^\star : \mathcal{V}^\star \times \mathcal{V}^\star \times \mathcal{V}^\star \to [0, 1]$. Given a prompt $\mathbf{x}$ and two responses $\mathbf{y}, \mathbf{y}' \in \mathcal{V}^\star$, querying the oracle returns $I \sim \text{Bern}(P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}))$, where $I = 1$ means that $\mathbf{y}$ is preferred to $\mathbf{y}'$, and $I = 0$ means that $\mathbf{y}'$ is preferred to $\mathbf{y}$. We assume $P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}) + P^\star(\mathbf{y}' \succ \mathbf{y} \mid \mathbf{x}) = 1$ and $P^\star(\mathbf{y} \succ \mathbf{y} \mid \mathbf{x}) = \frac{1}{2}$. Given a pair of responses $(\mathbf{y}, \mathbf{y}')$, we define the preference distribution $\lambda_{P^\star}$ by

$$\lambda_{P^\star}(\mathbf{x}, \mathbf{y}, \mathbf{y}') = \begin{cases} (\mathbf{y}, \mathbf{y}'), & \text{with probability } P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}), \\ (\mathbf{y}', \mathbf{y}), & \text{with probability } 1 - P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}). \end{cases}$$

Accordingly, we write $(\mathbf{y}^+, \mathbf{y}^-) \sim \lambda_{P^\star}(\mathbf{x}, \mathbf{y}, \mathbf{y}')$ as a pair of preferred and dispreferred responses.

We consider *game-theoretic* preference alignment, which formulates alignment as a KL-regularized game (Munos et al., 2024; Ye et al., 2024). Formally, we consider pairs of policies π_1, π_2 that play an alignment game against each other:

$$\max_{\pi_1} \min_{\pi_2} \; J_\beta(\pi_1, \pi_2) := P^\star(\pi_1 \succ \pi_2) - \beta \text{KL}(\pi_1 \| \pi_{\text{ref}}) + \beta \text{KL}(\pi_2 \| \pi_{\text{ref}}),$$

where π_{ref} is the model after SFT, $\beta > 0$ is the KL-regularization parameter, and $P^\star(\pi_1 \succ \pi_2) = \mathbb{E}_{x \sim \rho, \mathbf{y} \sim \pi_1(\cdot|\mathbf{x}), \mathbf{y}' \sim \pi_2(\cdot|\mathbf{x})}[P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x})]$. The goal of both players is to maximize their win rate against the opponent while not deviating too far from π_{ref}.

Without loss of generality, we restrict our attention to the policy class containing the policies with the same support set as π_{ref}. A Nash equilibrium (NE) is then defined as $(\pi_1^\star, \pi_2^\star)$ satisfying

$$J_\beta(\pi_1, \pi_2^\star) \le J_\beta(\pi_1^\star, \pi_2^\star) \le J_\beta(\pi_1^\star, \pi_2), \quad \text{for any } (\pi_1, \pi_2).$$

Since the game is symmetric for the two players, as proven by Ye et al. (2024), the NE is unique and the two equilibrium policies coincide: $\pi_1^\star = \pi_2^\star = \pi^\star$. Since $J_\beta(\pi^\star, \pi^\star) = \frac{1}{2}$, we have $J_\beta(\pi^\star, \pi) \ge \frac{1}{2}$ for any policy π. For each policy π, we use the following duality gap to measure how well it approximates $\pi^\star$:

$$\text{gap}_\Pi(\pi) = \max_{\pi_1 \in \Pi} J_\beta(\pi_1, \pi) - \min_{\pi_2 \in \Pi} J_\beta(\pi, \pi_2). \tag{1}$$

In Appendix B.1, we show that this duality gap is directly connected to the regret bound used to measure policy quality.

Online exploration. Under the BT model, the preference oracle can be represented as $P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}) = \sigma(r^\star(\mathbf{x}, \mathbf{y}) - r^\star(\mathbf{x}, \mathbf{y}'))$, where $\sigma : \mathbb{R} \mapsto [0, 1]$ is the sigmoid function. Direct alignment methods (e.g., DPO (Rafailov et al., 2023)) optimize the policy π_θ over the dataset $D = \{(\mathbf{x}_i, \mathbf{y}_i^+, \mathbf{y}_i^-)\}_{i=1}^n$ by minimizing a loss of the following form:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-)\sim D}\left[\log\sigma\left(\beta\left(\log\frac{\pi_\theta(\mathbf{y}^+|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^+|\mathbf{x})} - \log\frac{\pi_\theta(\mathbf{y}^-|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^-|\mathbf{x})}\right)\right)\right].$$

In passive online DPO (Guo et al., 2024), the learner repeatedly samples responses from the current policy, queries the preference oracle, adds the resulting pair of preferred and dispreferred responses to the dataset, and updates the policy using the DPO loss. This procedure uses online feedback, but the learner only observes responses that the current policy already tends to generate. If the reference policy assigns very small probability to a high-reward response, then passive online DPO may fail to discover it for exponentially many rounds in the small-β regime. If an online algorithm does not have such an exponential dependency, we say the algorithm has *efficient exploration*. Thus, online feedback alone does not guarantee efficient exploration.

The key idea of XPO Xie et al. (2025) is to include an SFT-based optimism term when updating the policy using the DPO loss. At round t, sample $\mathbf{x}_t \sim \rho$, $\mathbf{y}_t \sim \pi_{\theta_t}(\cdot \mid \mathbf{x}_t)$, and $\mathbf{y}'_t \sim \pi_{\text{ref}}(\cdot \mid \mathbf{x}_t)$. After querying the preference oracle, we obtain a new data sample $(\mathbf{x}_t, \mathbf{y}_t^+, \mathbf{y}_t^-)$ and update the dataset by $D_t = D_{t-1} \cup \{(\mathbf{x}_t, \mathbf{y}_t^+, \mathbf{y}_t^-)\}$. Then, we compute the next policy by minimizing

$$\alpha\sum_{i=1}^t \log(\pi(\mathbf{y}'_i \mid \mathbf{x}_i)) - \mathbb{E}_{(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-)\sim D_t}\left[\log\sigma\left(\beta\left(\log\frac{\pi_\theta(\mathbf{y}^+|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^+|\mathbf{x})} - \log\frac{\pi_\theta(\mathbf{y}^-|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^-|\mathbf{x})}\right)\right)\right].$$

Xie et al. (2025) interpret this additional term through DPO's role as implicit $Q^\star$ approximation: the extra term implements optimism in the face of uncertainty. This yields efficient exploration and an improved regret bound.

Unfortunately, the existing analysis relies on assumptions that are absent in game-theoretic preference alignment. First, general preference oracles need not admit any scalar reward representation. Second, KL-regularized games do not admit a single objective to be maximized and lack an interpretation as implicit $Q^\star$ approximation. Finally, the goal of game-theoretic alignment is to minimize the duality gap against adversarial opponent policies. Thus, exploration must control comparisons against possible opponents, not only discover high-reward responses. This leaves open the question of how to achieve efficient exploration in iterative NLHF.

3 Exploratory Nash preference optimization (ENPO) and Sample efficiency

3.1 Deliberate exploration is necessary

In this subsection, we show that the naive exploration in iterative NLHF is insufficient to achieve efficient exploration, which requires the algorithm to eliminate exponentially dependency on the KL-regularization parameter.

Before presenting the theoretical analysis, we briefly review *mirror descent*, a canonical no-regret dynamics for computing Nash equilibria. In mirror descent dynamics, at each round, every player updates her strategy along the first-order improvement direction against the current opponent profile—which coincides with her own strategy in our symmetric setting—regularized by a KL divergence that anchors the new iterate to the previous one:

$$\pi^{\text{next}}_{\star,\eta}[\pi] = \operatorname*{argmax}_{\tilde{\pi}} \mathbb{E}_{\mathbf{x}\sim\rho}\Big[\eta \mathbb{E}_{\mathbf{y}\sim\tilde{\pi}(\cdot|\mathbf{x})}\Big[P^*(\mathbf{y} \succ \pi|\mathbf{x}) - \beta\log\Big(\frac{\pi(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}\Big)\Big] - \text{KL}(\tilde{\pi}(\cdot|\mathbf{x})\|\pi(\cdot|\mathbf{x}))\Big]. \quad (2)$$

Under an appropriate step-size schedule, the dynamics converges to the regularized NE and attains a sublinear regret bound. In essence, the INPO-type algorithm aims to track the mirror descent update at every round.

To study sample complexity guarantees for iterative NLHF methods, we make the following statistical assumptions. The second assumption guarantees the policy class Π is large enough to ensure every π's mirror descent update is covered by Π, unless π is not too close to NE[1].

[1]The second condition in Assumption 3.1 is analogous to the "policy realizability for general preferences" assumption in Huang et al. (2025), but is slightly relaxed to accommodate finite policy classes.

Assumption 3.1 *We assume that $P^\star$, Π satisfy:*

- ***Finite policy class**: Π is finite, and $\pi_{\mathrm{ref}}, \pi^\star \in \Pi$,*
- ***Mirror descent realizability**: For each η in a finite set of stepsizes, there exists a sufficiently small $\varepsilon_\eta \geq 0$ such that $\pi^{\mathrm{next}}_{\star,\eta}[\pi] \in \Pi$ for every $\pi \in \Pi$ satisfying $D_{\mathrm{KL}}(\pi^\star \| \pi) > \varepsilon_\eta$.*
- ***Bounded log-ratio**: For any $\pi, \pi' \in \Pi$, and any $\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \mathcal{V}^\star$, we have $\left|\log \frac{\pi(\mathbf{y}|\mathbf{x})}{\pi'(\mathbf{y}|\mathbf{x})}\right| \leq V/\beta$ where $V \geq 1$ is a constant.*
- ***Symmetric preferences**: For any $\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \mathcal{V}^\star$, $P^\star(\mathbf{y} \succ \mathbf{y}'|\mathbf{x}) + P^\star(\mathbf{y}' \succ \mathbf{y}|\mathbf{x}) = 1$.*

In contrast to offline algorithms such as DPO, Nash learning algorithms such as INPO intrinsically draw new samples from the policy that mimics the mirror descent trajectory, and thus naturally perform on-policy sampling with moving policies. It is therefore natural to ask whether this iterative sampling process helps with exploration, and whether it is sufficient to achieve efficient exploration without further modifications. The following proposition shows that, even with iterative sampling, INPO must still pay $\Omega(e^{1/\beta})$ samples to achieve any sublinear regret bound. We defer the formal algorithm statement and its proof to Appendix B.2.

Proposition 3.2 *There exists an instance satisfying Assumption 3.1. Let π_t be a learned policy of INPO (described in Algorithm 2) on this instance after round t with constant batch size n. Then, for any $\delta \in (0,1)$, if $4n(T+1) \leq \delta e^{1/(4\beta)}$, with probability at least $1-\delta$, $\mathrm{gap}_\Pi(\pi^t)$ is at least a constant for all $t \leq T$ and $\mathrm{Reg}_T \geq \Omega(T)$.*

Recall that, we refer to sample complexity bounds that do *not* scale exponentially in $1/\beta$ as *sample efficient*. Since, without sample efficient, in the regime of small β iterative NLHF methods can suffer from an extremely large number of iterations to discover useful information, we target *sample efficient* as the main target in this paper.

3.2 Exploratory Nash preference optimization (ENPO)

In this subsection, we present our algorithmic solution to handle the non-efficient exploration issue in the iterative NLHF framework: Exploratary Nash Preference Optimization (ENPO).

We first introduce several additional notations and a main loss function: we use $\tau := (\mathbf{x}, \mathbf{y})$ to denote a pair of prompt and its response. Assume that, at round t, we have some empirical preference estimates for some τ against a policy π: $\widehat{P}_t(\tau \succ \pi)$, then we define $\widehat{\delta}_t(\tau;\pi) := \log \frac{\pi(\tau)}{\pi_{t-1}(\tau)} - \eta\beta \log \frac{\pi_{\mathrm{ref}}(\tau)}{\pi_{t-1}(\tau)} - \eta \widehat{P}_t(\tau \succ \pi_{t-1})$, and an empirical loss function:

$$L_{t-1}(\pi) := \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \left(\widehat{\delta}_t(\tau_{s,i};\pi) - \widehat{\delta}_t(\tau'_{s,i};\pi)\right)^2. \tag{3}$$

The minimizer of the population version of Eq. (3) corresponds the exact mirror descent target at the current round.

Each iteration t in ENPO has three steps. *(1) Data collection.* ENPO samples response pairs, each with one response from an earlier iterate and one from the explorer policy $\hat{\pi}_{t-1}$ chosen in the previous round, and adds them to the pairwise dataset $\mathcal{D}_2^t$. It also draws fresh responses from π_{t-1} to form $\mathcal{D}_1^t$. For each response τ in $\mathcal{D}_2^t$, ENPO estimates its win rate against the current policy: $\widehat{P}_t(\tau \succ \pi_{t-1})$. These estimates define the loss L_{t-1}. *(ii) Pessimistic exploitation.* The new policy π_t minimizes L_{t-1} plus an SFT-style term that rewards likelihood on the learner's own recent samples in $\mathcal{D}_1^t$. Unlike the optimism bonus in exploratory DPO, this term keeps π_t close to π_{t-1}, which strengthens the mirror-descent structure of the update. *(iii) Adversarial exploration.* ENPO then chooses the explorer $\hat{\pi}_t$ via solving a policy optimization problem. The explorer is a sampling policy under which π_t can still disagree most with some alternative policy, relative to how tightly the pairs collected so far constrain that disagreement. It thus targets opponent distributions that expose poorly covered regions of the preference game. After T rounds, ENPO outputs the uniform mixture of the iterates or the iterate with the smallest duality gap. We summarize ENPO in Algorithm 3.

Algorithm 1: Exploratory Nash Preference Optimization (ENPO) for Theoretical Analysis

Input: horizon T, schedules $(m_t, n_t, k_t)_{t=1}^T$, coefficients $\gamma, \lambda > 0$, stepsize $\eta > 0$, reference policy π_{ref}, pessimistic parameter α_t;

Initialize $\pi_{-1} \leftarrow \pi_{\text{ref}}$, $\pi_0 \leftarrow \pi_{\text{ref}}$, $\hat{\pi}_0 \leftarrow \pi_{\text{ref}}$;

Initialize $\mathcal{D}_1^0 \leftarrow \varnothing, \mathcal{D}_2^0 \leftarrow \varnothing$;

for $t = 1, \ldots, T$ **do**

Generate n_t response pairs $(\tau_{t-1,i}, \hat{\tau}_{t-1,i})$ via $\tau^{t-1,i} \sim \pi_{t-2}, \hat{\tau}^{t-1,i} \sim \hat{\pi}_{t-1}$;

Generate m_t responses $\tilde{\tau}_{t-1,i}$ via $\tilde{\tau}_{t-1,i} \sim \pi_{t-1}$;

Update preference datasets $D_1^t \leftarrow \{\tilde{\tau}_{t-1,i}\}_{i=1}^{m_t}$, $D_2^t \leftarrow D_2^{t-1} \cup \{(\tau_{t-1,i}, \hat{\tau}_{t-1,i})\}_{i=1}^{n_t}$;

Define $\mathcal{Y}_t^{\text{act}} := \{\tau, \tau' : (\tau, \tau') \in D_2^t\}$. Generate $c_{t,1}, \ldots, c_{t,k_t} \sim \pi_{t-1}$ independently. For every $a \in \mathcal{Y}_t^{\text{act}}$, query $Y_{t,a,j} \sim \text{Bern}(P^*(a \succ c_{t,j}))$ for $j = 1, \ldots, k_t$, and set $\widehat{P}_t(a \succ \pi_{t-1}) = \frac{1}{k_t} \sum_{j=1}^{k_t} Y_{t,a,j}$;

Compute

$$\pi_t \leftarrow \operatorname*{argmin}_{\pi \in \Pi} \left\{ -\alpha_t \sum\nolimits_{\tau \in D_1^t} \log \pi(\tau) + L_{t-1}(\pi) \right\}, \tag{4}$$

where $\mathcal{L}_{t-1}$ is defined in Eq. (3);

Compute

$$\hat{\pi}_t \leftarrow \operatorname*{argmax}_{\tilde{\pi} \in \Pi} \max_{\pi \in \Pi} \frac{\left(\frac{1}{m_t} \sum_{i=1}^{m_t} \log \frac{\pi_t(\tilde{\tau}_{t-1,i})}{\pi(\tilde{\tau}_{t-1,i})} - \mathbb{E}_{\tau' \sim \tilde{\pi}} \left[\log \frac{\pi_t(\tau')}{\pi(\tau')} \right] \right)^2}{\lambda + \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \left(\log \frac{\pi_t(\tau_{s,i})}{\pi(\tau_{s,i})} - \log \frac{\pi_t(\tau'_{s,i})}{\pi(\tau'_{s,i})} \right)^2}. \tag{5}$$

Output: Uniform$(\{\pi^t\}_{t=0}^{T-1})$ or $\pi_{\text{best}} \in \{\pi^t\}_{t=0}^{T-1}$ with the best duality gap.;

3.3 Sample efficiency guarantee of ENPO

The example in Section 3.1 suggests that relying solely on the exploration induced by the mirror descent trajectory is insufficient for achieving sample efficiency. More broadly, it has long been recognized in the RL literature that learning a Nash equilibrium is substantially harder than single-agent RL, due to strategic interaction and the need for coordinated exploration Zhang et al. (2020); Loftin et al. (2021); Huang et al. (2024). In particular, naïve optimistic exploration may waste samples on strategically irrelevant regions of the state space Loftin et al. (2021). The separation between general preferences and single-reward models has also been established in the offline RL setting (Cui & Du, 2022; Huang et al., 2025). More generally, computing an equilibrium is well known to be computationally hard Daskalakis et al. (2009).

The main difficulty of achieving sample efficiency under general preferences can be viewed from two perspectives. First, even in the symmetric-game setting, a player's effective reward depends on her own policy, so the learning objective is nonstationary rather than fixed. Second, game-theoretic performance measures are defined relative to either a Nash equilibrium or an adversarial policy, which may be completely out-of-sample. This gives rise to terms of the form

$$\mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y} \sim \pi_t(\cdot|\mathbf{x}), \mathbf{y}' \sim \pi(\cdot|\mathbf{x})} [\psi_t(\mathbf{x}, \mathbf{y}, \mathbf{y}') - \psi'_t(\mathbf{x}, \mathbf{y}, \mathbf{y}')]$$

in the regret decomposition, where π may be *completely unpredictable* at time t. As a result, even abundant in-sample data may provide little control over this term, leading to arbitrarily poor worst-case performance. In particular, this suggests that simple modifications such as proximal exploration in Xie & Jiang (2020) are unlikely to achieve sample efficiency for equilibrium learning.

Next, we present that, the nontrivial "pessimistic exploitation + optimistic exploration" structure in ENPO turns out to align naturally with mirror descent dynamics. By strengthening the regularization in the mirror descent update and then performing deliberate exploration, ENPO is sufficient to achieve efficient exploration.

Theorem 3.3 *Assume that $P^\star, \Pi$ satisfy Assumption 3.1 with $\varepsilon_\eta \le c/T$ for a stepsize $\eta = 2\beta/(1 + 4\beta^2) < 1$. Let $M_T := T(T+1)(T+2)/3$ and*

$$L_T := \max\{1, L_\Pi, L_P, \log(eT)\}, \text{ where } L_\Pi := \log\left(16T|\Pi|^2\delta^{-1}\right), \; L_P := \log\left(16M_T\delta^{-1}\right).$$

Set $m_t = n_t = k_t = t$, $\lambda = 16(1 + (V^2 + 1)/\beta^2)L_T^2$, $\gamma = \sqrt{(1 \vee \mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda))/(\lambda T)}$, $\alpha_t = 2/(\gamma t)$, *and stepsize* $\eta = 2\beta/(1 + 4\beta^2) < 1$ *in Algorithm 3. Then, with probability at least* $1 - \delta$,

$$\mathrm{Reg}_T^{\mathrm{best}} \le \max\left\{ \frac{1 + 2e^{\eta(2V+1)}}{\eta} \left(B + 20\sqrt{(1 \vee \mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda))\lambda T} \right), (1 + 4V)\sqrt{\frac{cT}{2}} \right\}.$$

The proof of this theorem utilizes the regret decomposition that is analogous to that in Xie & Jiang (2020), and the connection to online data coverage via the deliberate exploration step. We provide a complete proof of Theorem 3.3 in Appendix B.3.

Discussion. This is the first sample-efficiency result of this type available in the literature. Although iterative Nash preference optimization has dominated empirical practice and achieved significant progress, the only known sample-complexity result directly based on a mirror-descent oracle is that of Huang et al. (2025), which is offline and therefore does not guarantee no-regret performance when the reference policy is suboptimal. In the more natural online setting, sample-complexity guarantees are scarce, let alone sample-efficient guarantees. Our result also resolves an open problem posed by Lee et al. (2026), which asks for efficient algorithms based on mirror-descent oracles.

In our written algorithm, we sample an $O(t^3)$-sized batch at each iteration, which could potentially be improved by using improved sampling strategies. In this paper, our focus on closing the major gap on the exponential dependency in iterative NLHF methods, and we do not optimize the batch-size samples. Notice that, even under the BT model, the general XPO algorithm of Xie et al. (2025) also requires t samples per iteration. Their algorithm can avoid this only when samples can be fully reused, i.e., when a stationary policy is used as the sampler; this is incompatible with efficient learning in the game setting.

4 Bonus-exploer ENPO and Logarithmic regret bound

Our ENPO algorithm achieves sample efficiency without relying on either a minimax oracle or an explicit preference model, making it a natural basis for practical algorithms. At the same time, more complex oracle-based methods remain a central paradigm in RL, where access to best-response or minimax planning oracles often enables sharper regret guarantees. Recent results further suggest that, in KL-regularized zero-sum Markov games, oracle-style optimistic best-response sampling can improve the standard $O(\sqrt{T})$ rate to logarithmic regret. Motivated by this gap, we next study efficient exploration under access to minimax oracles and ask whether one can design an exploration strategy that achieves $O(\log T)$ regret in our preference-based setting.

We further design a minimax-oracle variant of ENPO, called Bonus-Explorer ENPO (BENPO), which is equipped with minimax oracles and best-response computation. This variant achieves an improved $O(\log T)$ regret bound without any $e^{1/\beta}$ dependence. Its analysis builds on recent techniques from KL-regularized RL (Zhao et al., 2025; Wu et al., 2025a; Lee et al., 2026); in particular, we exploit the structure of KL regularization to convert first-order errors into *second-order* errors, which ultimately enables the logarithmic rate. We conclude with the main theorem here, and defer the formal algorithm and analysis to Appendix B.4.

Theorem 4.1 (Logarithmic regret bound of Algorithm 4) *Assume that* $\mathcal{P}$ *is a finite class of anti-symmetric preference functions with* $P^\star \in \mathcal{P}$, *and let* $d_{\mathcal{P}}$ *be its eluder coefficient (see Appendix B.4.2). Then, with probability at least* $1 - \delta$, *Algorithm 4 with* $\lambda \in (0, 1]$ *and* $\gamma_t(\delta) = 4\log(2T|\mathcal{P}|/\delta) + \lambda$ *output iterates* $\{\pi_t\}_{t=1}^T$ *that satisfy*

$$\sum_{t=1}^{T} \mathrm{gap}_{\Pi}(\pi_t) \le O\left(\frac{d_{\mathcal{P}} \log(T|\mathcal{P}|/\delta)}{\beta} \right),$$

where the duality gap (1) is taken over all policies with the same support set as π_{ref}.

Discussion. To the best of our knowledge, BENPO is the first algorithm in the NLHF literature to *simultaneously* achieve $O(\log T)$ regret and avoid the $\exp(1/\beta)$ regularization scaling under

standard reverse-KL regularization and general function approximation. In comparison, Wu et al. (2025a); Lee et al. (2026) achieve $O(\log T)$ regret bounds but still incur an $\exp(1/\beta)$ dependence under KL regularization, while Nayak et al. (2025) avoid this dependence but require linear function approximation. We note, however, that all existing algorithms achieving $O(\log T)$ regret rely on minimax oracles, leaving open whether logarithmic regret can be attained using only optimization oracles.

One may observe that BENPO is closely related to the Optimistic ELHF algorithm of Ye et al. (2024). Indeed, both algorithms learn a minimax policy together with an explorer that maximizes a UCB-type bonus. The key difference is that Optimistic ELHF trains a policy-level explorer, whereas `BENPO` searches over the *response space* for each context $\mathbf{x}$. This finer-grained exploration structure is what enables our sharper analysis.

5 Direct ENPO and Experiments

5.1 Direct ENPO (DENPO)

In practice, we use lighter algorithm in the applications such as LLM fine-tuning. We therefore present *Direct ENPO* (DENPO), a practical instantiation that keeps ENPO's two exploration mechanisms but replaces these oracles with simple surrogates. The policy update becomes a few gradient steps on the iterative preference loss plus the SFT-style term on self-generated responses, so each round costs about the same as standard iterative NLHF. For exploration, DENPO does not solve Eq. (15). Instead, it use a surrogate which approximates the adversarial opponent without a separate optimization. We defer a full introduction to DENPO to Appendix C.1.

As a result, DENPO uses the exact same "pessimistic exploitation + optimistic exploration" structure as that in ENPO, with an even lighter implementation. This design makes it perfectly compatible with LLM fine-tuning, and inherit the useful explicit exploration mechanism. In the subsequent numerical experiments, we show that DENPO consistently outperforms RLHF and NLHF baselines on `Llama-3-8B-Instruct` at matched computational cost.

5.2 Numerical experiments

Experiment setup. We compare our DENPO algorithm with XPO (Xie et al., 2025) and INPO (Zhang et al., 2025b), using `Llama-3-8B-Instruct` as the base model. For INPO and DENPO, we sweep $1/\eta$ over a series of stepsize parameter choices: $\{7.5, 3.75, 1.875, 0.9375\}\times 10^{-3}$. The corresponding KL regularization coefficients β are $\{2.5, 2.5, 1.25, 0.625\} \times 10^{-3}$, respectively, for both methods. For ease of presentation, we use the subscripts $1/2$, $1/4$, and $1/8$ to denote scaling $1/\eta$ by these factors relative to the default value $1/\eta = 7.5 \times 10^{-3}$.

All methods use the same 20,000 UltraFeedback prompts per iteration for $T = 3$ iterations, with `Skywork-Reward-V2-Llama-3.1-8B-40M` providing training-time preference feedback, learning rate 10^{-6}, and one epoch per optimization stage. Training-time generation uses temperature 0.7, top-p 0.9, and at most 512 new tokens. Let K be the number of generated completions in a batch. INPO samples $K = 8$ responses per prompt from the current policy; XPO samples $K = 8$ responses from each of the current and reference policies. Both retain the top 70% of candidate pairs by reward margin. ENPO uses the single-pair configuration with $K = 2$ and $n_z = 4$ comparator responses from the current policy, trains on all prompts without a replay buffer, and updates both the policy and the adversary in each iteration. The comparison matches prompt budgets and iteration counts, with method-dependent computational costs. We present the matched computational cost empirical demonstrations in Appendix C.2. Experiments are conducted on a node with $8 \times$ H100 GPUs.

Evaluation. For each iteration-3 policy, we report four chat-quality metrics. We include the performances across three iterations in Appendix C.3.

We use two AlpacaEval-2 metrics (length-controlled) win rates under `gpt-4o-mini` and `gpt-4-turbo`. We also report the MT-Bench mean score under `gpt-4o-mini` and the Arena-Hard-v0.1 win-rate against `gpt-4-0314`, judged by `gpt-4o-mini` and estimated using Bradley-Terry scoring with bootstrap resampling over 500 prompts. AlpacaEval-2 uses its 805-prompt

Table 1: Iteration-3 performance, reported as mean $\pm$ sample standard deviation across three training runs. AE2-mini and AE2-turbo are AlpacaEval-2 length-controlled win-rates (%) under the respective judges; MT-Bench is the mean score; Arena-Hard is the win-rate (%). Higher is better for all metrics. Bold marks the highest mean in each column, including ties at the displayed precision.

Policy	$\eta^{-1}(\times 10^{-3})$	AE2-mini	AE2-turbo	MT-Bench	Arena-Hard
Base model	—	34.83	28.61	6.64	36.40
XPO	—	47.06 ± 1.46	40.75 ± 0.59	6.91 ± 0.05	45.33 ± 2.57
INPO	7.5	53.69 ± 0.38	45.81 ± 0.52	6.99 ± 0.05	50.83 ± 1.85
$\text{INPO}_{1/2}$	3.75	52.09 ± 2.59	44.15 ± 1.82	6.92 ± 0.07	48.57 ± 2.03
$\text{INPO}_{1/4}$	1.875	46.25 ± 2.89	39.61 ± 1.82	6.47 ± 0.11	38.70 ± 2.36
$\text{INPO}_{1/8}$	0.9375	30.24 ± 2.40	26.77 ± 1.62	5.38 ± 0.35	17.67 ± 5.39
DENPO	7.5	52.03 ± 1.34	44.29 ± 0.84	6.93 ± 0.07	50.70 ± 2.08
$\text{DENPO}_{1/2}$	3.75	52.56 ± 3.14	45.54 ± 0.87	6.97 ± 0.14	53.07 ± 2.22
$\text{DENPO}_{1/4}$	1.875	57.57 ± 2.86	47.48 ± 2.11	$\mathbf{7.04} \pm 0.08$	51.43 ± 3.04
$\text{DENPO}_{1/8}$	0.9375	$\mathbf{58.33} \pm 1.00$	$\mathbf{47.93} \pm 1.49$	$\mathbf{7.04} \pm 0.09$	$\mathbf{55.10} \pm 2.17$

benchmark, and MT-Bench contains 80 two-turn questions. For every trained configuration, Table 1 reports the mean and sample standard deviation across three training runs. These standard deviations describe variation across runs.

Results and discussion. By Table 1, $\text{DENPO}_{1/8}$ achieves the highest mean scores on both AlpacaEval-2 judges and Arena-Hard, and ties $\text{DENPO}_{1/4}$ on MT-Bench. Compared with default INPO, it improves the mean AE2-mini, AE2-turbo, and Arena-Hard win-rates by 4.64, 2.12, and 4.27 percentage points, respectively. At the same $1/\eta = 0.9375 \times 10^{-3}$, INPO falls below the initial-model baseline on all four metrics in every run, whereas DENPO remains above the baseline throughout these evaluations. This robust contrast supports DENPO's greater empirical stability and superior performance across the tested configurations.

One can see that DENPO performs better when the stepsize parameter η becomes larger (when η^{-1} becomes smaller). This identify that the normal INPO algorithm cannot work with a large stepsize, while the stabilization and the active exploration can help the algorithm work along the right policy optimization direction even when we move toward at a mirror descent target with a large stepsize.

It is also worth highlighting the robustness of DENPO's performance. Across three random seeds, the standard deviations of DENPO's scores are substantially smaller than its improvements over the baselines. This indicates that DENPO's gains are consistent across runs and unlikely to be explained by random variation alone.

6 Conclusion

We studied online iterative NLHF under general preference models and identified exploration as the key obstacle to provable regret guarantees. We showed that standard iterative NLHF, which relies on implicit exploration through its own policy updates, can incur an exponential dependence on the KL-regularization parameter. To address this, we proposed ENPO, which augments the iterative preference loss with two mechanisms: a pessimistic SFT-style regularizer that stabilizes the mirror-descent update, and an adversarial policy search that samples from opponents exposing poorly covered regions of the preference game. ENPO solves a major problem in online iterative NLHF by achieving *sample efficiency*: a regret bound without an exponential dependence on the KL-regularization parameter, while keeping the direct policy-optimization structure of iterative NLHF and requires neither minimax oracles nor preference-model training. Its variant BENPO improves the regret to $O(\log T)$ given additional oracles. Finally, our practical algorithm DENPO replaces these oracles with surrogates and consistently outperforms RLHF and NLHF baselines on `Llama-3-8B-Instruct` at matched computational cost. These results show that explicit exploration is useful not only in theory but also in practical LLM alignment. An interesting future direction is to reduce the per-iteration batch sample sizes required by our analysis.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and to help check and revise the proofs of our theoretical results. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving them, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results. Generating synthetic datasets, translation, cleaning or reformatting datasets, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, draft parts of the paper, and edit the paper to improve readability.

We have reviewed all AI-assisted work. All AI-suggested changes to the proofs were independently verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or new data collection, and uses only publicly available models and datasets in accordance with their licenses. We are not aware of any ethical concerns beyond those common to LLM alignment research.

REPRODUCIBILITY STATEMENT

For our theoretical results, all assumptions are stated in the main text, and complete proofs are provided in the appendix. For our experiments, the main text and appendix describe the models, datasets, baselines, hyperparameters, and evaluation protocols. Anonymized source code for DENPO is included in the supplementary material.

REFERENCES

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *ArXiv Preprint: 2303.08774*, 2023.

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv Preprint: 2204.05862*, 2022.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, et al. Human alignment of large language models through online preference optimisation. In *International Conference on Machine Learning*, pp. 5409–5435. PMLR, 2024.

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. T. Wang, S. Marks, C-R. Ségerie, M. Carroll, A. Peng, P. J. K. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Biyik, A. D. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

S. Cen, J. Mei, K. Goshvadi, H. Dai, T. Yang, S. Yang, D. Schuurmans, Y. Chi, and B. Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. In *ICLR*, 2025. URL https://openreview.net/forum?id=SQnitDuow6.

S. Chakraborty, J. Qiu, H. Yuan, A. Koppel, D. Manocha, F. Huang, A. S. Bedi, and M. Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *ICML*, pp. 6116–6135. PMLR, 2024.

J. Chen and N. Jiang. Offline reinforcement learning under value and density-ratio realizability: The power of gaps. In *UAI*, pp. 378–388. PMLR, 2022.

P. L. Chen, X. Chen, W. Yin, and T. Lin. ComPO: Preference alignment via comparison oracles. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=0lNwIIHWhZ.

P. L. Chen, X. Li, X. Chen, and T. Lin. Reward-free alignment for conflicting objectives. *ArXiv Preprint: 2602.02495*, 2026a.

X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *ICML*, pp. 3773–3793. PMLR, 2022.

Y. Chen, Y. He, M. I. Jordan, and F. Yao. How sampling shapes LLM alignment: From one-shot optima to iterative dynamics. *ArXiv Preprint: 2602.12180*, 2026b.

V. Chernozhukov, M. Demirer, G. Lewis, and V. Syrgkanis. Semi-parametric efficient policy learning with continuous actions. In *NeurIPS*, pp. 15039–15049, 2019.

P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, pp. 4302–4310, 2017.

V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mosse, E. Pacuit, S. Russell, H. Schoelkopf, et al. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *ICML*, pp. 9346–9360. PMLR, 2024.

T. Coste, U. Anwar, R. Kirk, and D. Krueger. Reward model ensembles help mitigate overoptimization. In *ICLR*, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx.

G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin, et al. UltraFeedback: Boosting language models with scaled AI feedback. In *ICML*, pp. 9722–9744. PMLR, 2024.

Qiwen Cui and Simon S Du. When are offline two-player zero-sum markov games solvable? *Advances in Neural Information Processing Systems*, 35:25779–25791, 2022.

J. Dai and E. Fleisig. Mapping social choice theory to RLHF. In *ICLR Workshop on Reliable and Responsible Foundation Models*, 2024. URL https://openreview.net/forum?id=iYthn3WATc.

C. Dann, N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. On oracle-efficient PAC RL with rich observations. In *NeurIPS*, pp. 1429–1439, 2018.

N. Das, S. Chakraborty, A. Pacchiano, and S. R. Chowdhury. Active preference optimization for sample efficient RLHF. In *ECML-KDD*, pp. 96–112. Springer, 2025.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

S. S. Du, S. Kakade, J. D. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in RL. In *ICML*, pp. 2826–2836. PMLR, 2021.

Y. Du, A. Winnicki, G. Dalal, S. Mannor, and R. Srikant. Exploration-driven policy optimization in RLHF: Theoretical insights on efficient data utilization. In *ICML*, pp. 11830–11887. PMLR, 2024.

Y. Duan, Z. Jia, and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *ICML*, pp. 2701–2709. PMLR, 2020.

J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

V. Dwaracherla, S. M. Asghari, B. Hao, and B. Van Roy. Efficient exploration for LLMs. In *ICML*, pp. 12215–12227. PMLR, 2024.

J. Eisenstein, C. Nagpal, A. Agarwal, A. Beirami, A. N. D'Amour, K. D. Dvijotham, A. Fisch, K. A. Heller, S. R. Pfohl, D. Ramachandran, P. Shaw, and J. Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *COLM*, 2024. URL https://openreview.net/forum?id=5u1GpUkKtG.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Model alignment as prospect theoretic optimization. In *ICML*, pp. 12634–12651, 2024.

A. M. Farahmand, R. Munos, and C. Szepesvári. Error propagation for approximate policy and value iteration. In *NeurIPS*, pp. 568–576, 2010.

A. Fisch, J. Eisenstein, V. Zayats, A. Agarwal, A. Beirami, C. Nagpal, P. Shaw, and J. Berant. Robust preference optimization through reward model distillation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=E2zKNuwNDc.

D. J. Foster and A. Rakhlin. Foundations of reinforcement learning and interactive decision making. *ArXiv Preprint: 2312.16730*, 2023.

D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *ArXiv Preprint: 2112.13487*, 2021.

D. J. Foster, N. Golowich, and Y. Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *COLT*, pp. 3969–4043. PMLR, 2023.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

G. Gabbianelli, G. Neu, and M. Papini. Importance-weighted offline learning done right. In *ALT*, pp. 614–634. PMLR, 2024.

Z. Gao, J. D. Chang, W. Zhan, O. Oertell, G. Swamy, K. Brantley, T. Joachims, J. A. Bagnell, J. D. Lee, and W. Sun. REBEL: Reinforcement learning via regressing relative rewards. In *NeurIPS*, pp. 52354–52400, 2024.

N. Golowich, A. Moitra, and D. Rohatgi. Exploration is harder than prediction: Cryptographically separating reinforcement learning from supervised learning. In *FOCS*, pp. 1953–1967. IEEE, 2024.

S. Guo, B. Zhang, T. Liu, T. Liu, M. Khalman, F. Llinares, A. Rame, T. Mesnard, Y. Zhao, B. Piot, et al. Direct language model alignment from online AI feedback. *ArXiv Preprint: 2402.04792*, 2024.

A. Huang, W. Zhan, T. Xie, J. D. Lee, W. Sun, A. Krishnamurthy, and D. J. Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via Chi-squared preference optimization. In *ICLR*, 2025. URL https://openreview.net/forum?id=hXm0Wu2U9K.

Jiawei Huang, Niao He, and Andreas Krause. Model-based rl for mean-field games is not statistically harder than single-agent rl. *arXiv preprint arXiv:2402.05724*, 2024.

N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *ICML*, pp. 1704–1713. PMLR, 2017.

C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *NeurIPS*, pp. 13406–13418, 2021a.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *ICML*, pp. 5084–5096. PMLR, 2021b.

N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

D. Kane, S. Liu, S. Lovett, and G. Mahajan. Computational-statistical gap in reinforcement learning. In *COLT*, pp. 1282–1302. PMLR, 2022.

T. Kozuno, W. Yang, N. Vieillard, T. Kitamura, Y. Tang, J. Mei, P. Ménard, M. G. Azar, M. Valko, R. Munos, et al. KL-entropy-regularized RL with a generative model is minimax optimal. *ArXiv Preprint: 2205.14211*, 2022.

J. Lee, M. Hong, K-S. Jun, C. Yun, and S-Y. Yun. Regularized online RLHF with generalized bilinear preferences. *ArXiv Preprint: 2602.23116*, 2026.

Z. Li, Z. Yang, and M. Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. In *ICML Workshop Interactive Learning with Implicit Human Feedback*, 2023. URL https://openreview.net/forum?id=gxM2AUFMsK.

K. Liu, Q. Long, Z. Shi, W. J. Su, and J. Xiao. Statistical impossibility and possibility of aligning LLMs with human preferences: From Condorcet paradox to Nash equilibrium. *ArXiv Preprint: 2503.10990*, 2025.

T. Liu, Y. Zhao, R. Joshi, M. Khalman, M. Saleh, P. J. Liu, and J. Liu. Statistical rejection sampling improves preference optimization. In *ICLR*, 2024a. URL https://openreview.net/forum?id=xbjSwwrQOe.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch reinforcement learning without great exploration. In *NeurIPS*, pp. 1264–1274, 2020.

Z. Liu, M. Lu, S. Zhang, B. Liu, H. Guo, Y. Yang, J. Blanchet, and Z. Wang. Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. In *NeurIPS*, pp. 138663–138697, 2024b.

Robert Loftin, Aadirupa Saha, Sam Devlin, and Katja Hofmann. Strategically efficient exploration in competitive multi-agent reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 1587–1596. PMLR, 2021.

R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation, 2005.

R-R. Maura-Rivero, M. Lanctot, F. Visin, and K. Larson. Jackpot! alignment as a maximal lottery. *ArXiv Preprint: 2501.19266*, 2025.

Y. Meng, M. Xia, and D. Chen. SimPO: Simple preference optimization with a reference-free reward. In *NeurIPS*, pp. 124198–124235, 2024.

A. Mishra. AI alignment and social choice: Fundamental limitations and policy implications. *ArXiv Preprint: 2310.16048*, 2023.

R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, D. Guo, Y. Tang, M. Geist, T. Mesnard, C. Fiegel, et al. Nash learning from human feedback. In *ICML*, pp. 36743–36768. PMLR, 2024.

Anupam Nayak, Tong Yang, Osman Yagan, Gauri Joshi, and Yuejie Chi. Achieving logarithmic regret in kl-regularized zero-sum markov games. *arXiv preprint arXiv:2510.13060*, 2025.

G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. *ArXiv Preprint: 1705.07798*, 2017.

E. Novoseller, Y. Wei, Y. Sui, Y. Yue, and J. Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *UAI*, pp. 1029–1038. PMLR, 2020.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, pp. 27730–27744, 2022.

A. Pal, D. Karkhanis, S. Dooley, M. Roberts, S. Naidu, and C. White. Smaug: Fixing failure modes of preference optimisation with DPO-positive. *ArXiv Preprint: 2402.13228*, 2024.

R. Park, R. Rafailov, S. Ermon, and C. Finn. Disentangling length from quality in direct preference optimization. In *ACL*, pp. 4998–5017, 2024.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, pp. 53728–53741, 2023.

R. Rafailov, J. Hejna, R. Park, and C. Finn. From r to q^*: Your language model is secretly a q-function. In *COLM*, 2024. URL https://openreview.net/forum?id=kEVcNxtqXk.

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *NeurIPS*, pp. 11702–11716, 2021.

M. Rita, F. Strub, R. Chaabouni, P. Michel, E. Dupoux, and O. Pietquin. Countering reward over-optimization in LLM with demonstration-guided reinforcement learning. In *Findings of ACL*, pp. 12447–12472, 2024.

C. Rosset, C-A. Cheng, A. Mitra, M. Santacroce, A. Awadallah, and T. Xie. Direct Nash optimization: Teaching language models to self-improve with general preferences. *ArXiv Preprint: 2404.03715*, 2024.

D. Russo and B. V. Roy. Eluder dimension and the sample complexity of optimistic exploration. In *NeurIPS*, pp. 2256–2264, 2013.

A. Saha, A. Pacchiano, and J. Lee. Dueling RL: Reinforcement learning with trajectory preferences. In *AISTATS*, pp. 6263–6289. PMLR, 2023.

R. Shi, R. Zhou, and S. S. Du. The crucial role of samplers in online direct preference optimization. In *ICLR*, 2025a. URL https://openreview.net/forum?id=F6z3utfcYw.

Z. Shi, K. Liu, Q. Long, W. J. Su, and J. Xiao. Fundamental limits of game-theoretic LLM alignment: Smith consistency and preference matching. *ArXiv Preprint: 2505.20627*, 2025b.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations*, 2023.

Y. Song, G. Swamy, A. Singh, J. Bagnell, and W. Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *NeurIPS*, pp. 12243–12270, 2024.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In *NeurIPS*, pp. 3008–3021, 2020.

W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *COLT*, pp. 2898–2933. PMLR, 2019.

G. Swamy, C. Dann, R. Kidambi, S. Wu, and A. Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *ICML*, pp. 47345–47377. PMLR, 2024.

Y. Tang, Z. Guo, Z. Zheng, D. Calandriello, R. Munos, M. Rowland, P. H. Richemond, M. Valko, B. Pires, and B. Piot. Generalized preference optimization: A unified approach to offline alignment. In *ICML*, pp. 47725–47742, 2024.

D. Tiapkin, D. Belomestny, D. Calandriello, E. Moulines, R. Munos, A. Naumov, P. Perrault, Y. Tang, M. Valko, and P. Menard. Fast rates for maximum entropy exploration. In *ICML*, pp. 34161–34221. PMLR, 2023.

D. Tiapkin, D. Belomestny, D. Calandriello, E. Moulines, A. Naumov, P. Perrault, M. Valko, and P. Menard. Demonstration-regularized RL. In *ICLR*, 2024. URL https://openreview.net/forum?id=lF2aip4Scn.

D. Tiapkin, D. Calandriello, D. Belomestny, E. Moulines, A. Naumov, K. Rasul, M. Valko, and P. Menard. Accelerating Nash learning from human feedback via mirror prox. *ArXiv Preprint: 2505.19731*, 2025.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv Preprint: 2302.13971*, 2023.

M. Uehara and W. Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *ICLR*, 2022. URL https://openreview.net/forum?id=tyrJsbKAe6.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pp. 6000–6010, 2017.

C. Wang, Y. Jiang, C. Yang, H. Liu, and Y. Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *ICLR*, 2024a. URL https://openreview.net/forum?id=2cRzmWXK9N.

L. Wang, A. Krishnamurthy, and A. Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *AISTATS*, pp. 766–774. PMLR, 2024b.

M. Wang, C. Ma, Q. Chen, L. Meng, Y. Han, J. Xiao, Z. Zhang, J. Huo, W. J. Su, and Y. Yang. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *ICLR*, 2025. URL https://openreview.net/forum?id=PDnEDS244P.

R. Wang, R. Salakhutdinov, and L. F. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *NeurIPS*, pp. 6123–6135, 2020.

Y. Wang, Q. Liu, and C. Jin. Is RLHF more difficult than standard RL? a theoretical perspective. In *NeurIPS*, pp. 76006–76032, 2023.

D. Wu, C. Shi, J. Yang, and C. Shen. Greedy sampling is provably efficient for RLHF. In *NeurIPS*, 2025a. URL https://openreview.net/forum?id=jQH0gdwsuT.

F. Wu, X. Huang, W. Xuan, Z. Zhang, Y. Xiao, G. Wan, X. Li, B. Hu, P. Xia, J. Leskovec, and Y. Choi. Multiplayer Nash preference optimization. In *ICLR*, 2026. URL https://openreview.net/forum?id=x7aLhLMVn1.

R. Wu and W. Sun. Making RL with preference-based feedback efficient via randomization. In *ICLR*, 2024. URL https://openreview.net/forum?id=Pe2lo3QOvo.

Y. Wu, Z. Sun, H. Yuan, K. Ji, Y. Yang, and Q. Gu. Self-play preference optimization for language model alignment. In *ICLR*, 2025b. URL https://openreview.net/forum?id=a3PmRgAB5T.

J. Xiao, Z. Li, X. Xie, E. Getzen, C. Fang, Q. Long, and W. J. Su. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552):2154–2164, 2025.

T. Xie and N. Jiang. $q^\star$ approximation schemes for batch reinforcement learning: A theoretical comparison. In *UAI*, pp. 550–559. PMLR, 2020.

T. Xie, C-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offine reinforcement learning. In *NeurIPS*, pp. 6683–6694, 2021.

T. Xie, D. J. Foster, Y. Bai, N. Jiang, and S. M. Kakade. The role of coverage in online reinforcement learning. In *ICLR*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

T. Xie, D. J. Foster, A. Krishnamurthy, C. Rosset, A. H. Awadallah, and A. Rakhlin. Exploratory preference optimization: Harnessing implicit q^*-approximation for sample-efficient RLHF. In *ICLR*, 2025. URL https://openreview.net/forum?id=QYigQ6gXNw.

W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *ICML*, pp. 54715–54754, 2024.

H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. Van Durme, K. Murray, and Y. J. Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *ICML*, pp. 55204–55224, 2024.

Y. Xu, R. Wang, L. F. Yang, A. Singh, and A. Dubrawski. Preference-based reinforcement learning with finite-time guarantees. In *NeurIPS*, pp. 18784–18794, 2020.

C. Ye, W. Xiong, Y. Zhang, H. Dong, N. Jiang, and T. Zhang. Online iterative reinforcement learning from human feedback with general preference model. In *NeurIPS*, pp. 81773–81807, 2024.

A. Zanette, M. J. Wainwright, and E. Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *NeurIPS*, pp. 13626–13640, 2021.

W. Zhan, B. Huang, A. Huang, N. Jiang, and J. D. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *COLT*, pp. 2730–2775. PMLR, 2022.

W. Zhan, M. Uehara, W. Sun, and J. D. Lee. Provable reward-agnostic preference-based reinforcement learning. In *ICLR*, 2024. URL https://openreview.net/forum?id=yTBXeXdbMf.

Kaiqing Zhang, Sham M. Kakade, Tamer Başar, and Lin F. Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

X. Zhang, J-F. Ton, W. Shen, H. Wang, and Y. Liu. Overcoming reward overoptimization via adversarial policy optimization with lightweight uncertainty estimation. *ArXiv Preprint: 2403.05171*, 2024.

Y. Zhang, D. Yu, T. Ge, L. Song, Z. Zeng, H. Mi, N. Jiang, and D. Yu. Improving LLM general preference alignment via optimistic online mirror descent. In *NeurIPS*, 2025a. URL https://openreview.net/forum?id=kZstGANG8D.

Y. Zhang, D. Yu, B. Peng, L. Song, Y. Tian, M. Huo, N. Jiang, H. Mi, and D. Yu. Iterative Nash policy optimization: Aligning LLMs with general preferences via no-regret learning. In *ICLR*, 2025b. URL https://openreview.net/forum?id=Pujt3ADZgI.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025.

Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *ArXiv Preprint: 2305.10425*, 2023.

R. Zhou, M. Fazel, and S. S. Du. Extragradient preference optimization (EGPO): Beyond last-iterate convergence for Nash learning from human feedback. In *COLM*, 2025. URL https://openreview.net/forum?id=EP7mAqx2BO.

B. Zhu, M. I. Jordan, and J. Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *ICML*, pp. 43037–43067, 2023.

B. D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, pp. 1433–1438, 2008.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *ArXiv Preprint: 1909.08593*, 2019.

Appendices

CONTENTS

1 Introduction **1**

2 Preliminaries and technical background **3**

3 Exploratory Nash preference optimization (ENPO) and Sample efficiency **4**

3.1 Deliberate exploration is necessary 4

3.2 Exploratory Nash preference optimization (ENPO) 5

3.3 Sample efficiency guarantee of ENPO 6

4 Bonus-exploer ENPO and Logarithmic regret bound **7**

5 Direct ENPO and Experiments **8**

5.1 Direct ENPO (DENPO) 8

5.2 Numerical experiments 8

6 Conclusion **9**

A Further Related Works **18**

B Theoretical analysis of efficient sample complexity guarantees **19**

B.1 Problem setup 19

B.2 Proofs of the necessity of efficient explorations 22

B.2.1 Iterative Nash preference optimization is *not* sufficient for efficient exploration 22

B.3 Proofs of efficient exploration guarantee for ENPO 25

B.3.1 Algorithm 25

B.3.2 Additional notations 25

B.3.3 Proof of Theorem 3.3 26

B.3.4 Auxiliary Propositions and Lemmas for Theorem B.10 30

B.4 Proofs of improved sample efficiency of BENPO 41

B.4.1 Algorithm 41

B.4.2 Assumptions and additional notations 42

B.4.3 Main proof of sample efficiency 42

B.4.4 Technical lemmas 45

C Additional details in DENPO and numerical experiments **46**

C.1 Details on Direct ENPO (DENPO) 46

C.2 Computational cost and budget-matched comparisons 48

C.3 Performance across training iterations 49

A Further Related Works

We comment on other topics, including more discussions on game-theoretic alignment methods, theoretical and empirical preference alignment methods, and theoretical reinforcement learning (RL). For an overview of preference alignment methods, we refer to the recent survey (Casper et al., 2023).

More discussions on game-theoretic alignment methods. Beyond algorithmic progress, game-theoretic alignment is motivated by a more structural question: what alignment properties are induced by treating pairwise human preferences as a game rather than compressing them into a scalar reward? This perspective is especially relevant when preferences are cyclic, non-transitive, or incompatible with the BT reward model, where standard RLHF might fail to satisfy social choice and diversity desiderata (Bradley & Terry, 1952; Luce, 2005; Mishra, 2023; Conitzer et al., 2024; Dai & Fleisig, 2024; Xiao et al., 2025). Recent results have shown that the appeal of NLHF is not merely computational: with raw pairwise preference as the payoff, NLHF is Condorcet consistent, Smith consistent under a no-tie condition, and can preserve preference diversity through mixed strategies when no single response dominates all alternatives (Maura-Rivero et al., 2025; Liu et al., 2025). However, Shi et al. (2025b) emphasize that raw preference is only one possible payoff choice and study the more general game with payoff $\Psi(P(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}))$, where $\Psi(t) = t$ recovers standard NLHF and $\Psi(t) = \log(\frac{t}{1-t})$ connects to reward-based objectives under the BT reward model. Their analysis shows both robustness and limits: Condorcet consistency only requires preserving the majority threshold, while Smith consistency requires an anti-symmetry condition $\Psi(t) + \Psi(1-t) = 2\Psi(\frac{1}{2})$, implying that practical preference models must preserve pairwise anti-symmetry to guarantee support on the Smith set. At the same time, mixed strategies provide only a partial notion of diversity. In the stronger sense of preference matching, which asks the learned policy to match a target distribution reflecting human preference diversity, Shi et al. (2025b) prove an impossibility result: no smooth and learnable mapping of pairwise preferences can generally guarantee a unique Nash equilibrium equal to the target policy, even under the standard BT model (Xiao et al., 2025; Shi et al., 2025b). These results suggest that game-theoretic alignment is well motivated by social choice consistency and robustness to non-transitive preferences, but its objective has axiomatic limitations that are different from the exploration mechanism studied in this paper.

Theoretical preference alignment methods. Recent progress in theoretical preference alignment has developed along several complementary directions. Under the BT reward model, early work established finite-sample guarantees for RLHF and preference-based reinforcement learning in structured settings such as tabular MDPs, linear MDPs, linearly parameterized policies, and KL-constrained policy classes (Novoseller et al., 2020; Xu et al., 2020; Saha et al., 2023; Zhu et al., 2023; Li et al., 2023; Zhan et al., 2024; Xiong et al., 2024). Another line connects alignment to offline reinforcement learning theory, where the central challenge is distribution shift between the reference policy and candidate aligned policies, leading to analyses based on coverage, uncertainty quantification, and regularization (Liu et al., 2020; Jin et al., 2021b; Rashidinejad et al., 2021; Xie et al., 2021; Uehara & Sun, 2022; Zhan et al., 2022; Chen & Jiang, 2022). Recent work studies direct preference objectives themselves: DPO provides a policy-only formulation of KL-regularized preference optimization (Rafailov et al., 2023), while subsequent theory considers broader divergence choices, generalized preference losses, and additional regularization mechanisms (Wang et al., 2024a; Tang et al., 2024; Liu et al., 2024b; Cen et al., 2025; Fisch et al., 2025). XPO and χPO are representative of this recent shift toward simple direct objectives with end-to-end guarantees under general function approximation or offline alignment assumptions (Xie et al., 2025; Huang et al., 2025). Other works conduct a rigorous analysis through the lens of dataset coverage to differentiate offline DPO and online RLHF (Song et al., 2024) and report faster convergence rate than the information-theoretic lower bounds for online reward maximization in RL by exploiting the structure induced by KL regularization (Shi et al., 2025a). Beyond scalar-reward preference models, another emerging theoretical direction studies general, possibly non-transitive preferences through game-theoretic solution concepts, replacing reward maximization with duality-gap guarantees (Munos et al., 2024; Swamy et al., 2024; Rosset et al., 2024; Ye et al., 2024).

Empirical preference alignment methods. Empirical preference alignment methods have mainly focused on improving the data construction, objective design, and robustness of direct fine-tuning pipelines. Since offline preference optimization is sensitive to the source and quality of comparison

pairs, several works modify the preference data itself: SLiC-HF calibrates sequence likelihoods using preference rankings, while statistical rejection sampling constructs more informative preference pairs from refined model samples rather than relying only on static SFT-generated data (Zhao et al., 2023; Liu et al., 2024a). A second line redesigns the alignment loss to address practical failure modes. KTO replaces pairwise comparisons with prospect-theoretic desirable/undesirable feedback (Ethayarajh et al., 2024); length-aware variants study verbosity and length-quality confounding (Park et al., 2024); contrastive preference objectives and SimPO adjust reward parameterization, normalization, or margins to improve performance (Xu et al., 2024; Meng et al., 2024). More recent variants broaden empirical preference alignment beyond a fixed pairwise-loss formulation, including comparison-oracle optimization (Chen et al., 2025) and reward-free multi-objective alignment under conflicting preferences (Chen et al., 2026a). Another empirical direction studies reward hacking and overoptimization, proposing practical stabilizers such as reward-model ensembles, uncertainty-aware or adversarial policy optimization, demonstration-guided calibration, and positive-response correction (Coste et al., 2024; Eisenstein et al., 2024; Rita et al., 2024; Pal et al., 2024; Zhang et al., 2024). Finally, recent work on sampling dynamics shows that sampling can change ranking behavior and concentration, while iterative deployment choices can lead to entropy collapse (Chen et al., 2026b).

Theoretical reinforcement learning (RL). The RL theory relevant to modern alignment has developed around structural complexity, coverage, regularization, and computational tractability. One line studies sample-efficient RL with rich function approximation by identifying complexity measures under which prediction error can be converted into policy improvement, including eluder-dimension, contextual-decision-process, general value-function-approximation, bilinear-class, and decision-estimation-coefficient frameworks (Russo & Roy, 2013; Sun et al., 2019; Wang et al., 2020; Du et al., 2021; Xie et al., 2023; Foster et al., 2023). These results clarify when statistically efficient learning is possible, but they also reveal a persistent computational-statistical tension: many general optimistic or version-space methods require oracle access, constrained optimization, or other procedures that are difficult to implement at scale (Dann et al., 2018; Kane et al., 2022; Golowich et al., 2024). A second line analyzes offline and off-policy learning through coverage and concentrability, where guarantees depend on how errors under a behavior distribution transfer to comparator policies or value functions (Farahmand et al., 2010; Xie & Jiang, 2020; Zanette et al., 2021). This viewpoint is closely related to policy learning and off-policy evaluation in statistics and econometrics, where doubly robust and semiparametric methods are used to control distribution shift under observational data (Chernozhukov et al., 2019; Kallus & Uehara, 2020; Athey & Wager, 2021). A third line studies regularized RL objectives: maximum-entropy and KL-regularized MDPs provide soft dynamic-programming identities and policy representations (Ziebart et al., 2008; Ziebart, 2010; Neu et al., 2017), while recent finite-sample analyses establish sharper guarantees for entropy- or KL-regularized learning in tabular settings (Kozuno et al., 2022; Tiapkin et al., 2023; 2024). Finally, χ^2-based weighting and regularization have emerged as tools for uncertainty-sensitive offline policy evaluation and optimization (Duchi & Namkoong, 2019; Duan et al., 2020; Wang et al., 2024b; Gabbianelli et al., 2024).

B Theoretical analysis of efficient sample complexity guarantees

B.1 Problem setup

In Section 2, we describe policies as conditional LLM policies $\pi(\mathbf{y} \mid \mathbf{x})$, where prompts $\mathbf{x}$ are drawn from a fixed distribution ρ and responses $\mathbf{y}$ belong to $\mathcal{V}^\star$. We use the same convention in this section. To keep formulas short, we often suppress the prompt and *write a response atom as* τ; formally, all expectations of the form

$$\mathbb{E}_{\tau\sim\pi,\tau'\sim\pi'}[f(\tau,\tau')] \equiv \mathbb{E}_{\mathbf{x}\sim\rho,\ \mathbf{y}\sim\pi(\cdot|\mathbf{x}),\ \mathbf{y}'\sim\pi'(\cdot|\mathbf{x})}[f(\mathbf{y},\mathbf{y}';\mathbf{x})],$$

where f can be a general function of two response atoms $\tau := (\mathbf{x},\mathbf{y})$ and $\tau' := (\mathbf{x},\mathbf{y}')$. This notion implies that the two responses are always compared under the same prompt.

Similarly,

$$P^\star(\pi \succ \pi') := \mathbb{E}_{\mathbf{x}\sim\rho,\ \mathbf{y}\sim\pi(\cdot|\mathbf{x}),\ \mathbf{y}'\sim\pi'(\cdot|\mathbf{x})}[P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x})].$$

The KL divergence is the prompt-averaged conditional KL,

$$D_{\mathrm{KL}}(\pi\|\pi') := \mathbb{E}_{\mathbf{x}\sim\rho}[D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{x})\|\pi'(\cdot \mid \mathbf{x}))]\,.$$

When the prompt is suppressed, $\mathcal{Y}$ denotes the finite response support at a fixed prompt; equivalently, all statements below hold pointwise in $\mathbf{x}$ and are then integrated over $\mathbf{x} \sim \rho$.

Let Π be a finite class of conditional policies. Recall that J_β is defined in Section 2 as

$$\begin{aligned} J_\beta(\pi_1, \pi_2) =& P^\star(\pi_1 \succ \pi_2) - \beta \mathrm{KL}(\pi_1 \| \pi_{\mathrm{ref}}) + \beta \mathrm{KL}(\pi_2 \| \pi_{\mathrm{ref}}) \\ =& \mathbb{E}_{\mathbf{x}\sim\rho}\left[\mathbb{E}_{\mathbf{y}_1\sim\pi_1(\cdot|\mathbf{x}),\mathbf{y}_2\sim\pi_2(\cdot|\mathbf{x})}[P^\star(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})]\right] \\ & - \mathbb{E}_{\mathbf{x}\sim\rho}\left[\beta D_{\mathrm{KL}}(\pi_1(\cdot \mid \mathbf{x}) \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})) + \beta D_{\mathrm{KL}}(\pi_2(\cdot \mid \mathbf{x}) \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x}))\right]. \end{aligned}$$

The regularized regret is defined as

$$\begin{aligned} \mathrm{Reg}_T :=& \max_{\pi\in\Pi} \sum_{t=1}^{T} \left(J_\beta(\pi, \pi_{t-1}) - \frac{1}{2} \right) \\ =& \max_{\pi\in\Pi} \sum_{t=1}^{T} \left(P^\star(\pi \succ \pi_{t-1}) - \frac{1}{2} - \beta(D_{\mathrm{KL}}(\pi \| \pi_{\mathrm{ref}}) - D_{\mathrm{KL}}(\pi_{t-1} \| \pi_{\mathrm{ref}})) \right). \end{aligned} \tag{6}$$

Notice that the regularized regret above is connected to the duality gap defined in Section 2.

Proposition B.1 *For any $\hat{\pi} \in \Pi$, the duality gap satisfies*

$$\frac{1}{2}\mathrm{gap}(\hat{\pi}) = \max_{\pi\in\Pi} J_\beta(\pi, \hat{\pi}) - \frac{1}{2} = \frac{1}{2} - \min_{\pi\in\Pi} J_\beta(\hat{\pi}, \pi). \tag{7}$$

Let $\pi_0, \ldots, \pi_{t-1}$ be a sequence of policies in Π. Then, $\mathrm{Reg}_T \le \frac{1}{2}\sum_{t=1}^{T} \mathrm{gap}(\pi_{t-1})$, and $\frac{1}{2}\mathrm{gap}(\frac{1}{T}\sum_{t=1}^{T}\pi_{t-1}) \le \frac{1}{T}\mathrm{Reg}_T$.

Proof of Proposition B.1. For any $\hat{\pi} \in \Pi$, we have

$$\begin{aligned} \mathrm{gap}(\hat{\pi}) =& \max_{\pi\in\Pi} J_\beta(\pi, \hat{\pi}) - \min_{\pi\in\Pi} J_\beta(\hat{\pi}, \pi) \\ =& \left(\max_{\pi\in\Pi} J_\beta(\pi, \hat{\pi}) - \frac{1}{2} \right) - \left(\min_{\pi\in\Pi} J_\beta(\hat{\pi}, \pi) - \frac{1}{2} \right) \\ =& \left(\max_{\pi\in\Pi} J_\beta(\pi, \hat{\pi}) - \frac{1}{2} \right) + \left(\max_{\pi\in\Pi} -J_\beta(\hat{\pi}, \pi) + \frac{1}{2} \right) \\ =& \left(\max_{\pi\in\Pi} J_\beta(\pi, \hat{\pi}) - \frac{1}{2} \right) + \left(\max_{\pi\in\Pi} J_\beta(\pi, \hat{\pi}) - 1 + \frac{1}{2} \right) \\ =& 2\left(\max_{\pi\in\Pi} J_\beta(\pi, \hat{\pi}) - \frac{1}{2} \right). \end{aligned}$$

This proves Eq. (7). Then, we have

$$\mathrm{Reg}_T = \max_{\pi\in\Pi} \sum_{t=1}^{T} \left(J_\beta(\pi, \pi_{t-1}) - \frac{1}{2} \right) \le \sum_{t=1}^{T} \left(\max_{\pi\in\Pi} J_\beta(\pi, \pi_{t-1}) - \frac{1}{2} \right) = \frac{1}{2}\sum_{t=1}^{T} \mathrm{gap}(\pi_{t-1}),$$

and

$$\begin{aligned} \frac{1}{T}\mathrm{Reg}_T = \max_{\pi\in\Pi} \frac{1}{T}\sum_{t=1}^{T} \left(J_\beta(\pi, \pi_{t-1}) - \frac{1}{2} \right) \ge& \max_{\pi\in\Pi} J_\beta\left(\pi, \frac{1}{T}\sum_{t=1}^{T}\pi_{t-1} \right) - \frac{1}{2} \\ =& \frac{1}{2}\mathrm{gap}\left(\frac{1}{T}\sum_{t=1}^{T}\pi_{t-1} \right). \end{aligned}$$

This completes the proof. □

We restate the antisymmetric property of the ground true preferences as the following assumption.

Assumption B.2 (Antisymmetric preference) *For every prompt $\mathbf{x} \in \mathcal{V}^\star$ and every pair of responses $\mathbf{y}, \mathbf{y}' \in \mathcal{V}^\star$,*

$$P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}) + P^\star(\mathbf{y}' \succ \mathbf{y} \mid \mathbf{x}) = 1.$$

Consequently, $P^\star(\mathbf{y} \succ \mathbf{y} \mid \mathbf{x}) = \frac{1}{2}$ and $P^\star(\pi \succ \pi) = \frac{1}{2}$ for every policy π.

Assumption B.3 (Finite policy class) *The prompt-response support considered in the proof is finite, the policy class Π is finite.*

Recall that the symmetric Nash equilibrium $\pi^\star$ is defined as

$$(\pi^\star, \pi^\star) \in \arg\max_{\pi_1} \arg\min_{\pi_2} \; J_\beta(\pi_1, \pi_2), \tag{8}$$

and the unconstrained mirror descent update based on a policy $\pi \in \Pi$ as

$$\pi^+_{\star,\eta}[\pi] = \operatorname*{argmax}_{\tilde{\pi}} \mathbb{E}_{\mathbf{x}\sim\rho}\Big[\eta \mathbb{E}_{\mathbf{y}\sim\tilde{\pi}(\cdot|\mathbf{x})}\Big[P^*(\mathbf{y} \succ \pi|\mathbf{x}) - \beta \log\Big(\frac{\pi(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}\Big)\Big] - \text{KL}(\tilde{\pi}(\cdot|\mathbf{x})\|\pi(\cdot|\mathbf{x}))\Big], \tag{9}$$

which is the solution of a KL-proximal problem obtained from the exact game gradient. Note that in the subsequent appendix we use $\pi^+_{\star,\eta}[\pi]$ instead of $\pi^{\text{next}}_{\star,\eta}[\pi]$ for simplicity. This exact population mirror descent update is equivalent to the conditional policy

$$\pi^+_{\star,\eta}[\pi](\mathbf{y} \mid \mathbf{x}) := \frac{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})^{\eta\beta}\pi(\mathbf{y} \mid \mathbf{x})^{1-\eta\beta}\exp(\eta P^\star(\mathbf{y} \succ \pi \mid \mathbf{x}))}{\sum_{\mathbf{u}\in\mathcal{Y}} \pi_{\text{ref}}(\mathbf{u} \mid \mathbf{x})^{\eta\beta}\pi(\mathbf{u} \mid \mathbf{x})^{1-\eta\beta}\exp(\eta P^\star(\mathbf{u} \succ \pi \mid \mathbf{x}))}, \tag{10}$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{V}^\star$. In the analysis of algorithms, we often use the following notation: at round t,

$$\pi^\star_t := \pi^+_{\star,\eta}[\pi_{t-1}].$$

Assumption B.4 (Policy realizability) *The unique Nash equilibrium $\pi^\star \in \Pi$. Moreover, $\pi^+_{\star,\eta}[\pi] \in \Pi$ for all $\pi \in \Pi$, where the definitions of $\pi^\star$ and $\pi^+_{\star,\eta}[\pi]$ can be found in Eq.* (8) *and Eq.* (9)*.*

Assumption B.5 (Bounded logarithmic ratio) *For every prompt $\mathbf{x} \in \mathcal{V}^\star$ and every response $\mathbf{y} \in \mathcal{V}^\star$, and all $\pi, \pi' \in \Pi$,*

$$\left|\log \frac{\pi(\mathbf{y} \mid \mathbf{x})}{\pi'(\mathbf{y} \mid \mathbf{x})}\right| \le \frac{V}{\beta},$$

where V is a constant.

We will write

$$B := \frac{V}{\beta}.$$

For any policy $\pi \in \Pi$, define

$$P^\star(\mathbf{y} \succ \pi \mid \mathbf{x}) := \mathbb{E}_{\mathbf{u}\sim\pi(\cdot|\mathbf{x})}[P^\star(\mathbf{y} \succ \mathbf{u} \mid \mathbf{x})].$$

Denote

$$h_t(\pi, \mathbf{x}, \mathbf{y}, \mathbf{y}') := \log \frac{\pi(\mathbf{y} \mid \mathbf{x})}{\pi(\mathbf{y}' \mid \mathbf{x})} - \eta\beta \log \frac{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}' \mid \mathbf{x})} - (1 - \eta\beta) \log \frac{\pi_{t-1}(\mathbf{y} \mid \mathbf{x})}{\pi_{t-1}(\mathbf{y}' \mid \mathbf{x})}.$$

Proposition B.6 *For any $\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \mathcal{V}^\star$, and a distribution μ over $\mathcal{V}^\star$ given $\mathbf{x}$, let*

$$L_1 = \mathbb{E}_{\hat{\mathbf{y}}\sim\mu, I_t(\mathbf{y})\sim\text{Bern}(P^\star(\mathbf{y}\succ\hat{\mathbf{y}}|\mathbf{x})), I_t(\mathbf{y}')\sim\text{Bern}(P^\star(\mathbf{y}'\succ\hat{\mathbf{y}}|\mathbf{x}))}\Big(h_t(\pi, \mathbf{x}, \mathbf{y}, \mathbf{y}') - \eta\big(I_t(\mathbf{y}) - I_t(\mathbf{y}')\big)\Big)^2,$$

and

$$L_2 = \Big(h_t(\pi, \mathbf{x}, \mathbf{y}, \mathbf{y}') - \eta\big(P^\star(\mathbf{y} \succ \mu \mid \mathbf{x}) - P^\star(\mathbf{y}' \succ \mu \mid \mathbf{x})\big)\Big)^2.$$

Then, L_1 is equivalent to L_2 up to a constant.

Algorithm 2: Iterative Nash Preference Optimization (INPO (Zhang et al., 2025b))

Input: Number of rounds T, per-round sample sizes $\{n_t\}_{t=1}^{T}$, reference policy π_{ref}, stepsize $\eta > 0$;

Initialize $\pi^0 \leftarrow \pi_{\text{ref}},\ D^t_{\text{pref}} \leftarrow \varnothing$;

for $t = 0, 1, \ldots, T$ **do**

Generate n_t response pairs $\{(\mathbf{x}, \mathbf{y}, \mathbf{y}')\}$ via $\mathbf{x} \sim \rho, \mathbf{y}, \mathbf{y}' \sim \pi_t$;

Query preference label $(\mathbf{x}, \mathbf{y}_+, \mathbf{y}_-) \sim \lambda_{P^\star}(\mathbf{x}, \mathbf{y}, \mathbf{y}')$;

Update datasets $D^t_{\text{pref}} \leftarrow D^t_{\text{pref}} \cup \{(\mathbf{x}, \mathbf{y}_+, \mathbf{y}_-)\}$;

Compute π_{t+1} via

$$\pi_{t+1} \leftarrow \operatorname*{argmin}_{\pi \in \Pi} \left\{ \sum\nolimits_{(\mathbf{x}, \mathbf{y}_+, \mathbf{y}_-) \in D^t_{\text{pref}}} \left(h_t(\pi, \mathbf{x}, \mathbf{y}_+, \mathbf{y}_-) - \frac{\eta}{2} \right)^2 \right\},$$

where h_t is defined in Eq. (11).

Output: Uniform$(\{\pi^t\}_{t=1}^{T})$;

Proof. Fix $\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \mathcal{V}^\star$ and write

$$h \coloneqq h_t(\pi, \mathbf{x}, \mathbf{y}, \mathbf{y}'), \qquad Z \coloneqq I_t(\mathbf{y}) - I_t(\mathbf{y}').$$

Conditional on $\hat{\mathbf{y}} \sim \mu$, the indicators $I_t(\mathbf{y})$ and $I_t(\mathbf{y}')$ are independent Bernoullis with means $P^*(\mathbf{y} \succ \hat{\mathbf{y}} \mid \mathbf{x})$ and $P^*(\mathbf{y}' \succ \hat{\mathbf{y}} \mid \mathbf{x})$, thus the tower property gives

$$\begin{aligned}
\mathbb{E}[Z] =& \mathbb{E}_{\hat{\mathbf{y}} \sim \mu, I_t(\mathbf{y}) \sim \text{Bern}(P^\star(\mathbf{y} \succ \hat{\mathbf{y}} | \mathbf{x})), I_t(\mathbf{y}') \sim \text{Bern}(P^\star(\mathbf{y}' \succ \hat{\mathbf{y}} | \mathbf{x}))} \big(I_t(\mathbf{y}) - I_t(\mathbf{y}') \big) \\
=& \mathbb{E}_{\hat{\mathbf{y}} \sim \mu} \left[\mathbb{E}_{I_t(\mathbf{y}) \sim \text{Bern}(P^\star(\mathbf{y} \succ \hat{\mathbf{y}} | \mathbf{x})), I_t(\mathbf{y}') \sim \text{Bern}(P^\star(\mathbf{y}' \succ \hat{\mathbf{y}} | \mathbf{x}))} \left[I_t(\mathbf{y}) - I_t(\mathbf{y}') \mid \hat{\mathbf{y}} \right] \right] \\
=& \mathbb{E}_{\hat{\mathbf{y}} \sim \mu} \big[P^\star(\mathbf{y} \succ \hat{\mathbf{y}} \mid \mathbf{x}) - P^\star(\mathbf{y}' \succ \hat{\mathbf{y}} \mid \mathbf{x}) \big] = P^\star(\mathbf{y} \succ \mu \mid \mathbf{x}) - P^\star(\mathbf{y}' \succ \mu \mid \mathbf{x}).
\end{aligned}$$

Applying the bias-variance identity to the square loss in ηZ,

$$L_1 = \mathbb{E}\big[(h - \eta Z)^2\big] = \big(h - \eta \mathbb{E}[Z]\big)^2 + \eta^2 \mathrm{Var}(Z) = L_2 + \eta^2 \mathrm{Var}(Z).$$

By independence of $I_t(\mathbf{y}), I_t(\mathbf{y}')$ given $\hat{\mathbf{y}}$, the law of total variance yields

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathbb{E}_{\hat{\mathbf{y}} \sim \mu}[\mathrm{Var}(Z \mid \hat{\mathbf{y}})] + \mathrm{Var}_{\hat{\mathbf{y}} \sim \mu}(\mathbb{E}[Z \mid \hat{\mathbf{y}}]) \\
&= \mathbb{E}_{\hat{\mathbf{y}} \sim \mu} \left[P^\star(\mathbf{y} \succ \hat{\mathbf{y}} | \mathbf{x}) \big(1 - P^\star(\mathbf{y} \succ \hat{\mathbf{y}} | \mathbf{x})\big) + P^\star(\mathbf{y}' \succ \hat{\mathbf{y}} | \mathbf{x}) \big(1 - P^\star(\mathbf{y}' \succ \hat{\mathbf{y}} | \mathbf{x})\big) \right] \\
&\qquad + \mathrm{Var}_{\hat{\mathbf{y}} \sim \mu} \big(P^\star(\mathbf{y} \succ \hat{\mathbf{y}} \mid \mathbf{x}) - P^\star(\mathbf{y}' \succ \hat{\mathbf{y}} \mid \mathbf{x}) \big),
\end{aligned}$$

which depends only on $(\mathbf{x}, \mathbf{y}, \mathbf{y}', \mu, P^\star)$ and not on π. Hence, $L_1 = L_2 + C$ for a constant C independent of π. □

B.2 Proofs of the necessity of efficient explorations

B.2.1 Iterative Nash preference optimization is *not* sufficient for efficient exploration

Define

$$h_t(\pi, \mathbf{x}, \mathbf{y}_+, \mathbf{y}_-) := \log \left(\frac{\pi(\mathbf{y}_+|\mathbf{x}) (\pi_{\text{ref}}(\mathbf{y}_-|\mathbf{x}))^{\eta\beta} (\pi^t(\mathbf{y}_-|\mathbf{x}))^{1-\eta\beta}}{\pi(\mathbf{y}_-|\mathbf{x}) (\pi_{\text{ref}}(\mathbf{y}_+|\mathbf{x}))^{\eta\beta} (\pi^t(\mathbf{y}_+|\mathbf{x}))^{1-\eta\beta}} \right). \tag{11}$$

The following counterexample is constructed similarly to that of Xie et al. (2025). However, our counterexample must satisfy much stricter assumptions, especially mirror descent realizability, and is therefore harder to find. Note that we define $\pi^+_{\star,\eta}[\pi]$ as in Eq. (10).

Assumption B.7 (Relaxed policy realizability) *The unique symmetric Nash policy satisfies $\pi^\star \in \Pi$. Moreover, for the stepsize η used in the algorithm, there exists $\varepsilon_\eta \geq 0$ such that $\pi^+_{\star,\eta}[\pi] \in \Pi$ for every $\pi \in \Pi$ satisfying $D_{\mathrm{KL}}(\pi^\star \| \pi) > \varepsilon_\eta$.*

Assumption B.7 weakens Assumption B.4 by requiring mirror-target realizability only away from the Nash policy. If the current policy is already inside the ε_η-neighborhood of $\pi^\star$, then the policy is already a small-gap candidate, and exact realizability of its next mirror target is no longer needed for a best-iterate guarantee.

Lemma B.8 *Suppose Assumptions B.2, B.3 and B.7 hold. If $q \in \Pi$ satisfies $D_{\mathrm{KL}}(\pi^\star \| q) \le \varepsilon$, then*

$$\mathrm{gap}(q) \le \sqrt{2}(1+4V)\sqrt{\varepsilon}.$$

Proof. By Proposition B.1,

$$\frac{1}{2}\mathrm{gap}(q) = \max_{\pi \in \Pi} J_\beta(\pi, q) - \frac{1}{2}.$$

Fix any $\pi \in \Pi$. Since $\pi^\star$ is the symmetric Nash policy, $J_\beta(\pi, \pi^\star) \le \frac{1}{2}$. Hence

$$J_\beta(\pi, q) - \frac{1}{2} \le J_\beta(\pi, q) - J_\beta(\pi, \pi^\star).$$

Let $p_x := \pi^\star(\cdot \mid \mathbf{x})$ and $q_x := q(\cdot \mid \mathbf{x})$. The preference part satisfies

$$\begin{aligned} |P^\star(\pi \succ q) - P^\star(\pi \succ \pi^\star)| &\le \mathbb{E}_{\mathbf{x}\sim\rho}[\mathrm{TV}(q_x, p_x)] \le \sqrt{\mathbb{E}_{\mathbf{x}\sim\rho}\mathrm{TV}(q_x, p_x)^2} \\ &\le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\pi^\star \| q)} \le \sqrt{\frac{\varepsilon}{2}}, \end{aligned}$$

where we used Cauchy–Schwarz and Pinsker's inequality.

It remains to control the KL-regularization difference. For each prompt $\mathbf{x}$,

$$D_{\mathrm{KL}}(q_x \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})) - D_{\mathrm{KL}}(p_x \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})) = D_{\mathrm{KL}}(q_x \| p_x) + \left\langle q_x - p_x, \log \frac{p_x}{\pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})} \right\rangle.$$

By Assumption B.5,

$$\left\| \log \frac{p_x}{\pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})} \right\|_\infty \le \frac{V}{\beta}, \qquad \left\| \log \frac{q_x}{p_x} \right\|_\infty \le \frac{V}{\beta}.$$

Moreover,

$$D_{\mathrm{KL}}(q_x \| p_x) \le D_{\mathrm{KL}}(q_x \| p_x) + D_{\mathrm{KL}}(p_x \| q_x) = \left\langle q_x - p_x, \log \frac{q_x}{p_x} \right\rangle \le \frac{V}{\beta} \|q_x - p_x\|_1.$$

Therefore

$$D_{\mathrm{KL}}(q_x \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})) - D_{\mathrm{KL}}(p_x \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})) \le \frac{2V}{\beta} \|q_x - p_x\|_1.$$

Averaging over $\mathbf{x} \sim \rho$ and applying Pinsker again,

$$\beta(D_{\mathrm{KL}}(q \| \pi_{\mathrm{ref}}) - D_{\mathrm{KL}}(\pi^\star \| \pi_{\mathrm{ref}})) \le 2V\, \mathbb{E}_x \|q_x - p_x\|_1 \le 2\sqrt{2}V\sqrt{\varepsilon}.$$

Combining the preference and KL parts gives, uniformly over $\pi \in \Pi$,

$$J_\beta(\pi, q) - \frac{1}{2} \le \frac{1}{\sqrt{2}}\sqrt{\varepsilon} + 2\sqrt{2}V\sqrt{\varepsilon} = \frac{1+4V}{\sqrt{2}}\sqrt{\varepsilon}.$$

Taking the maximum over $\pi \in \Pi$ and using Proposition B.1 proves

$$\mathrm{gap}(q) \le \sqrt{2}(1+4V)\sqrt{\varepsilon}.$$

□

Proof of Proposition 3.2. Consider the following bandit instance with general preferences, i.e., a single fixed prompt $\mathbf{x}$ and three responses $\mathcal{Y} = \{a, b, c\}$ with

$$P^\star(a \succ b) = 1, P^\star(b \succ c) = 1, P^\star(a \succ c) = \frac{1}{2}.$$

Let $\beta \le \frac{1}{16\log(2)}$. Let the reference policy be $\pi_{\text{ref}}(a) = \epsilon, \pi_{\text{ref}}(b) = 1 - 2\epsilon$ and $\pi_{\text{ref}}(c) = \epsilon$, where $\epsilon := e^{-\frac{1}{4\beta}} \le \frac{1}{16}$.

Consider a policy class

$$\Pi = \left\{\pi_{\text{ref}}, (\pi_{\text{ref}})^1_{\star,\eta}, \dots, (\pi_{\text{ref}})^{\lceil (1/\beta^2)\log(8/\epsilon)\rceil}_{\star,\eta}, \pi^\star\right\},$$

where $(\pi_{\text{ref}})^t_{\star,\eta}$ is defined inductively by $(\pi_{\text{ref}})^t_{\star,\eta} = \pi^+_{\star,\eta}[(\pi_{\text{ref}})^{t-1}_{\star,\eta}]$ for all $t \ge 1$, and $\pi^\star$ is Nash equilibrium of the game $\max_{\pi_1\in\Delta(\mathcal{Y})}\min_{\pi_2\in\Delta(\mathcal{Y})} J_\beta(\pi_1,\pi_2)$.

We first show that Assumptions B.5 and B.7 holds with Π. Assumption B.7 is satisfied by the construction of Π and the contraction of KL divergence between the mirror descent iterates and $\pi^\star$. In particular, we use the contraction guarantees to show that $(\pi_{\text{ref}})^{\lceil (1/\beta^2)\log(8/\epsilon)\rceil}_{\star,\eta}$ has to satisfy $D_{\text{KL}}(\pi^\star \| (\pi_{\text{ref}})^{\lceil (1/\beta^2)\log(8/\epsilon)\rceil}_{\star,\eta}) \le \varepsilon$.

To verify Assumption B.5, we prove the following fact: if $\pi \in \Delta(\mathcal{Y})$ satisfies $\min_{\tau\in\mathcal{Y}}\pi(\tau) \ge \epsilon^5$, then $\pi^+_\star[\pi](\tau) \ge \epsilon^5$ for all $\tau \in \mathcal{Y}$. For any $\tau \in \mathcal{Y}$, $\pi^+_\star[\pi]$ has a closed-form solution as shown in Eq. (10). Denote

$$Z := \sum_{\tau\in\mathcal{Y}} (\pi_{\text{ref}}(\tau))^{\eta\beta}(\pi(\tau))^{1-\eta\beta} e^{\eta P^\star(\tau\succ\pi)}.$$

Then, we can upper bound Z by

$$Z \le e^\eta \sum_{\tau\in\mathcal{Y}} (\pi_{\text{ref}}(\tau))^{\eta\beta}(\pi(\tau))^{1-\eta\beta} \le e^\eta \Big(\sum_{\tau\in\mathcal{Y}} \pi_{\text{ref}}(\tau)\Big)^{\eta\beta}\Big(\sum_{\tau\in\mathcal{Y}}\pi(\tau)\Big)^{1-\eta\beta} = e^\eta,$$

where the first inequality is because $P^\star(\tau \succ \pi) \le 1$ for all $\tau \in \mathcal{Y}$ and $\pi \in \Delta(\mathcal{Y})$, and the second inequality is by Hölder's inequality with $p = \frac{1}{\eta\beta}$ and $q = \frac{1}{1-\eta\beta}$. Thus,

$$\pi^+_\star(\tau) = \frac{1}{Z}(\pi_{\text{ref}}(\tau))^{\eta\beta}(\pi(\tau))^{1-\eta\beta} e^{\eta P^\star(\tau\succ\pi)} \ge e^{-\eta}\epsilon^{\eta\beta}\left(\epsilon^5\right)^{1-\eta\beta} = (\epsilon^{-4\beta})^{-\eta}\epsilon^{\eta\beta}\epsilon^{5(1-\eta\beta)} = \epsilon^5,$$

where the first inequality is because $Z \le e^\eta$, $\pi_{\text{ref}}(\tau) \ge \epsilon, \pi(\tau) \ge \epsilon^5$ for all $\tau \in \mathcal{Y}$, and $e^{\eta P^\star(\cdot\succ\cdot)} \ge 1$. Since $\min_{\tau\in\mathcal{Y}}\pi_{\text{ref}}(\tau) \ge \epsilon \ge \epsilon^5$, we have $\min_{\tau\in\mathcal{Y}}\pi(\tau) \ge \epsilon^5$ for all $\pi \in \Pi \setminus \{\pi^\star\}$. It remains to show $\min_{\tau\in\mathcal{Y}^*}\pi^\star(\tau) \ge \epsilon^5$. Note that the Nash equilibrium can be characterized by the following equality (Calandriello et al., 2024, Eq. (12)):

$$\pi^\star(\tau) = \frac{\pi_{\text{ref}}(\tau)\exp\left(P^\star(\tau\succ\pi^\star)/\beta\right)}{\sum_{\tau'\in\mathcal{Y}}\pi_{\text{ref}}(\tau')\exp\left(P^\star(\tau'\succ\pi^\star)/\beta\right)}. \tag{12}$$

Since $\exp\left(P^\star(\tau\succ\pi^\star)/\beta\right) \ge 1$, and $\sum_{\tau'\in\mathcal{Y}}\pi_{\text{ref}}(\tau')\exp\left(P^\star(\tau'\succ\pi^\star)/\beta\right) \le e^{1/\beta}$, we have

$$\pi^\star(\tau) \ge \pi_{\text{ref}}(\tau)e^{-1/\beta} \ge \epsilon e^{-1/\beta} = e^{-\frac{1}{4\beta}}e^{-\frac{1}{\beta}} = e^{-\frac{5}{4\beta}} = \epsilon^5.$$

Thus, we have $\min_{\tau\in\mathcal{Y}}\pi(\tau) \ge \epsilon^5$ for all $\pi \in \Pi$. Clearly, we have $|\log(\frac{\pi(\mathbf{y})}{\pi'(\mathbf{y})})| \le \log(1/\epsilon^5) = \frac{5}{4\beta}$ for all $\pi, \pi' \in \Pi$. Hence, Assumption B.5 holds for Π with $V = 5/4$.

We now consider applying Algorithm 2 to learn a Nash policy. For each round t, define the event $\mathcal{E}_t := \{\text{actions } a, c \text{ is not sampled at the round } t\}$. We establish the following lemma and present its proof after this proof.

Lemma B.9 *For each round t, define the event $\mathcal{E}_t := \{$actions a, c is not sampled at the round $t\}$. Then, whenever $\mathcal{E}_0, \mathcal{E}_1, \dots, \mathcal{E}_{t-1}$ happen, we can choose the next policy $\pi^t = \pi_{\text{ref}}$, and therefore*

- $\mathbb{P}(\mathcal{E}_t \,|\, \mathcal{E}_0, \mathcal{E}_1, \dots, \mathcal{E}_{t-1}) \ge 1 - 4n_t\epsilon$, *and*
- *the duality gap of π^t is lower bounded by a constant:* $\text{gap}_\Pi(\pi^t) \ge c$, *where c is the KL divergence between π_{ref} and $\pi^\star$.*

For any $\delta \in (0,1)$, suppose that $4n(T+1)\epsilon \le \delta$. By Lemma B.9, conditional on $E_0, \dots, E_{t-1}$, $\mathbb{P}(E_t \mid E_0, \dots, E_{t-1}) \ge 1 - 4n\epsilon$. Therefore, by iterated conditioning and Bernoulli's inequality,

$$\begin{aligned}\mathbb{P}(E_0 \cap E_1 \cap \cdots \cap E_T) &= \mathbb{P}(E_0)\prod_{t=0}^{T-1}\mathbb{P}(E_{t+1} \mid E_0, \dots, E_t) \\ &\ge (1-4n\epsilon)^{T+1} \ge 1 - 4n(T+1)\epsilon \ge 1 - \delta.\end{aligned} \tag{13}$$

Since $\epsilon = e^{-1/(4\beta)}$, the assumed condition is equivalently $4n(T+1) \le \delta e^{1/(4\beta)}$. Combining Eq. (13) with the second point of Lemma B.9 completes the proof. □

Proof of Lemma B.9. If $\mathcal{E}_0, \mathcal{E}_1, \ldots, \mathcal{E}_{t-1}$ happen, the online dataset D_{pref}^{t-1} only contains b. In this case, the loss function in Algorithm 2 is a constant that is independent of π, and thus π_{ref} is a valid minimizer of the loss function. Then,

$$\mathbb{P}(\mathcal{E}_t \,|\, \mathcal{E}_0, \mathcal{E}_1, \ldots, \mathcal{E}_{t-1}) = (1 - \pi^t(a) - \pi^t(c))^{2n_t} = (1-2\epsilon)^{2n_t} \ge 1 - 4n_t\epsilon,$$

where the last inequality is because $(1+x)^a \ge 1 + ax$ for any $x \ge -1$ and $a \ge 1$.

For the second claim, since $\pi^\star \in \Pi$, Proposition B.1 gives

$$\frac{1}{2}\text{gap}_\Pi(\pi_{\text{ref}}) = \max_{\pi_1 \in \Pi} J_\beta(\pi_1, \pi_{\text{ref}}) - \frac{1}{2} \ge J_\beta(\pi^\star, \pi_{\text{ref}}) - \frac{1}{2}.$$

Set $f(\pi_2) := J_\beta(\pi^\star, \pi_2)$. By the Nash property $\pi^\star \in \arg\min_{\pi_2} f(\pi_2)$ and $f(\pi^\star) = \frac{1}{2}$. Writing the preference term as the linear form $P^\star(\pi^\star \succ \pi_2) = \langle g, \pi_2 \rangle$ with $g(y') := P^\star(\pi^\star \succ y' \mid x)$, and using the exact three-point identity

$$\text{KL}(\pi_2 \| \pi_{\text{ref}}) - \text{KL}(\pi^\star \| \pi_{\text{ref}}) = \text{KL}(\pi_2 \| \pi^\star) + \big\langle \log(\pi^\star/\pi_{\text{ref}}), \pi_2 - \pi^\star \big\rangle,$$

we get

$$f(\pi_2) - f(\pi^\star) = \big\langle g + \beta \log(\pi^\star/\pi_{\text{ref}}), \pi_2 - \pi^\star \big\rangle + \beta \text{KL}(\pi_2 \| \pi^\star).$$

First-order optimality of $\pi^\star$ over $\Delta(\mathcal{Y})$ makes the inner product nonnegative (the constant component of $\nabla f(\pi^\star)$ annihilates $\pi_2 - \pi^\star$, which has zero sum), so $f(\pi_2) - f(\pi^\star) \ge \beta \text{KL}(\pi_2 \| \pi^\star)$. Taking $\pi_2 = \pi_{\text{ref}}$,

$$\frac{1}{2}\text{gap}_\Pi(\pi_{\text{ref}}) \ge f(\pi_{\text{ref}}) - f(\pi^\star) \ge \beta \text{KL}(\pi_{\text{ref}} \| \pi^\star).$$

Since $\pi^\star$ has full support and $\pi_{\text{ref}} \ne \pi^\star$, the constant $c := 2\beta \text{KL}(\pi_{\text{ref}} \| \pi^\star) > 0$ is independent of t and T, giving $\text{gap}_\Pi(\pi^t) \ge c$. □

B.3 Proofs of efficient exploration guarantee for ENPO

B.3.1 Algorithm

We restate the formal statement of Exploratory Nash Preference Optimization (ENPO) is in Algorithm 3.

B.3.2 Additional notations

For a prompt-response atom $\tau = (\mathbf{x}, \mathbf{y})$, write $\pi(\tau)$ as shorthand for $\pi(\mathbf{y} \mid \mathbf{x})$ inside likelihood ratios. For any $\pi, \pi' \in \Pi$, define the population residual

$$\delta(\tau; \pi, \pi') := \log \frac{\pi(\tau)}{\pi'(\tau)} - \eta\beta \log \frac{\pi_{\text{ref}}(\tau)}{\pi'(\tau)} - \eta P^\star(\tau \succ \pi'), \tag{16}$$

where $P^\star(\tau \succ \pi')$ means $P^\star(\mathbf{y} \succ \pi' \mid \mathbf{x})$ when $\tau = (\mathbf{x}, \mathbf{y})$. For a general $\pi \in \Pi$, define

$$\Delta_t(\tau, \tau'; \pi) := \delta(\tau; \pi, \pi_{t-1}) - \delta(\tau'; \pi, \pi_{t-1}).$$

At round t, write

$$\delta_t(\tau) := \delta(\tau; \pi_t, \pi_{t-1}), \qquad \Delta_t(\tau, \tau') := \Delta_t(\tau, \tau'; \pi_t) = \delta_t(\tau) - \delta_t(\tau'),$$

where $\tau = (\mathbf{x}, \mathbf{y})$ and $\tau' = (\mathbf{x}, \mathbf{y}')$ share the same prompt. We define the empirical residuals at round t by substituting $P^\star$ with the empirical preferences $\widehat{P}_t$ at round t:

$$\widehat{\delta}_t(\tau; \pi) := \log \frac{\pi(\tau)}{\pi_{t-1}(\tau)} - \eta\beta \log \frac{\pi_{\text{ref}}(\tau)}{\pi_{t-1}(\tau)} - \eta \widehat{P}_t(\tau \succ \pi_{t-1}), \tag{17}$$

and

$$\widehat{\Delta}_t(\tau, \tau'; \pi) := \widehat{\delta}_t(\tau; \pi) - \widehat{\delta}_t(\tau'; \pi).$$

Algorithm 3: Exploratory Nash Preference Optimization (ENPO)

Input: horizon T, schedules $(m_t, n_t, k_t)_{t=1}^T$, coefficients $\gamma, \lambda > 0$, stepsize $\eta > 0$, reference policy π_{ref}, pessimistic parameter α_t;

Initialize $\pi_{-1} \leftarrow \pi_{\text{ref}}$, $\pi_0 \leftarrow \pi_{\text{ref}}$, $\hat{\pi}_0 \leftarrow \pi_{\text{ref}}$;

Initialize $\mathcal{D}_1^0 \leftarrow \varnothing, \mathcal{D}_2^0 \leftarrow \varnothing$;

for $t = 1, \ldots, T$ **do**

Generate n_t response pairs $(\tau_{t-1,i}, \hat{\tau}_{t-1,i})$ via $\tau^{t-1,i} \sim \pi_{t-2}, \hat{\tau}^{t-1,i} \sim \hat{\pi}_{t-1}$;

Generate m_t responses $\tilde{\tau}_{t-1,i}$ via $\tilde{\tau}_{t-1,i} \sim \pi_{t-1}$;

Update preference datasets $D_1^t \leftarrow \{\tilde{\tau}_{t-1,i}\}_{i=1}^{m_t}$, $D_2^t \leftarrow D_2^{t-1} \cup \{(\tau_{t-1,i}, \hat{\tau}_{t-1,i})\}_{i=1}^{n_t}$;

Define $\mathcal{Y}_t^{\text{act}} := \{\tau, \tau' : (\tau, \tau') \in D_2^t\}$. Generate $c_{t,1}, \ldots, c_{t,k_t} \sim \pi_{t-1}$ independently. For every $a \in \mathcal{Y}_t^{\text{act}}$, query $Y_{t,a,j} \sim \text{Bern}(P^*(a \succ c_{t,j}))$ for $j = 1, \ldots, k_t$, and set $\widehat{P}_t(a \succ \pi_{t-1}) = \frac{1}{k_t} \sum_{j=1}^{k_t} Y_{t,a,j}$;

Compute

$$\pi_t \leftarrow \operatorname*{argmin}_{\pi \in \Pi} \left\{ -\alpha_t \sum\nolimits_{\tau \in D_1^t} \log \pi(\tau) + L_{t-1}(\pi) \right\}, \tag{14}$$

where $\mathcal{L}_{t-1}$ is defined in Eq. (18);

Compute

$$\hat{\pi}_t \leftarrow \operatorname*{argmax}_{\tilde{\pi} \in \Pi} \max_{\pi \in \Pi} \frac{\left(\frac{1}{m_t} \sum_{i=1}^{m_t} \log \frac{\pi_t(\tilde{\tau}_{t-1,i})}{\pi(\tilde{\tau}_{t-1,i})} - \mathbb{E}_{\tau' \sim \tilde{\pi}} \left[\log \frac{\pi_t(\tau')}{\pi(\tau')} \right] \right)^2}{\lambda + \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \left(\log \frac{\pi_t(\tau_{s,i})}{\pi(\tau_{s,i})} - \log \frac{\pi_t(\tau'_{s,i})}{\pi(\tau'_{s,i})} \right)^2}. \tag{15}$$

Output: Uniform$(\{\pi^t\}_{t=0}^{T-1})$ or $\pi_{\text{best}} \in \{\pi^t\}_{t=0}^{T-1}$ with the best duality gap.;

The empirical learner loss in Algorithm 3 is

$$L_{t-1}(\pi) := \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \widehat{\Delta}_t(\tau_{s,i}, \tau'_{s,i}; \pi)^2. \tag{18}$$

At round t, Algorithm 3 only estimates preferences for the active prompt-response atoms appearing in the historical comparison dataset,

$$\mathcal{Y}_t^{\text{act}} := \{(\mathbf{x}_{s,i}, \mathbf{y}_{s,i}), (\mathbf{x}_{s,i}, \mathbf{y}'_{s,i}) : 0 \le s < t,\ 1 \le i \le n_{s+1}\}.$$

For an active atom $a = (\mathbf{x}, \mathbf{y}) \in \mathcal{Y}_t^{\text{act}}$, the proof version of Algorithm 3 uses prompt-conditional opponents: draw $c_{t,a,j} \sim \pi_{t-1}(\cdot \mid \mathbf{x})$ and query $Y_{t,a,j} \sim \text{Bern}(P^\star(\mathbf{y} \succ c_{t,a,j} \mid \mathbf{x}))$, for $j = 1, \ldots, k_t$. Reusing a common set of opponents is a practical variant, but the theorem below is proved for the prompt-conditional version stated in the algorithm. All preference-estimation concentration below is imposed only on $\mathcal{Y}_t^{\text{act}}$, not on the entire response space.

The SEC class is

$$\Psi := \left\{ \psi_{\rho',\pi}(\tau, \tau') = \log \frac{\rho'(\tau)}{\pi(\tau)} - \log \frac{\rho'(\tau')}{\pi(\tau')} : \rho', \pi \in \Pi \right\}.$$

For a fixed floor $\lambda > 0$, define

$$\text{SEC}_{\text{NLHF}}(\Psi, \lambda) := \sup_{\psi_1, \ldots, \psi_T \in \Psi} \sup_{\substack{\pi_0, \ldots, \pi_{T-1} \in \Pi \\ \widehat{\pi}_1, \ldots, \widehat{\pi}_T \in \Pi}} \sum_{t=1}^{T} \frac{\left(\mathbb{E}_{\tau \sim \pi_{t-1}, \tau' \sim \widehat{\pi}_t} \psi_t(\tau, \tau') \right)^2}{\lambda \vee \sum_{s=1}^{t-1} \mathbb{E}_{\tau \sim \pi_{s-1}, \tau' \sim \widehat{\pi}_s} \psi_t(\tau, \tau')^2}. \tag{19}$$

B.3.3 Proof of Theorem 3.3

Proof strategy. The proof of the regret bound consists of a deterministic regret bound Theorem B.11 and a high-probability concentration step Lemma B.24. The deterministic part shows that accurate empirical estimates imply small regret; the concentration part verifies that these accuracy events hold with high probability for the stated sample sizes.

To facilitate the proof, we first restate the main theorem with complete notations.

Theorem B.10 *Suppose Assumptions B.2 to B.5 hold. Let* $M_T := T(T+1)(T+2)/3$ *and*

$$L_T := \max\{1, L_\Pi, L_P, \log(eT)\},\ \textit{where}\ L_\Pi := \log\left(16T|\Pi|^2\delta^{-1}\right),\ L_P := \log\left(16M_T\delta^{-1}\right).$$

Set the exploration coefficient $\alpha_t = 2/(\gamma m_t)$*, the sample counts* $m_t = n_t = k_t = t$*, the SEC parameters* $\overline{H} := 1 + B^2 + \beta^{-2}$*,* $\lambda := 16\overline{H}L_T^2$*,* $S_T \geq 1 \vee \mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda)$*,* $\gamma := \sqrt{S_T/(\lambda T)}$*, and the constant stepsize* $\eta := 2\beta/(1+4\beta^2)$*. Then, with probability at least* $1-\delta$*,*

$$\mathrm{Reg}_T \leq \frac{1+2e^{\eta(2V+1)}}{\eta}\left(B + 20\sqrt{S_T\lambda T}\right) \leq 3e^{3V}\eta^{-1}\left(B + 20\sqrt{S_T\lambda T}\right). \tag{20}$$

Thus, Algorithm 3 achieves $\widetilde{O}(\sqrt{T})$ *regret bound.*

We first prove a deterministic version of the regret bound. It assumes the empirical quantities used by the learner and explorer are accurate, and postpones concentration to the end of Appendix B.3.4.

Theorem B.11 *Suppose Assumptions B.2 to B.5 hold, and choose a constant stepsize satisfying* $0 < \eta \leq 4\beta/(1+4\beta^2)$*. Let* $B := V/\beta$*,* $c_\eta := 1 - \eta\beta - \eta/(4\beta)$ *and* $R_\eta := \eta(2V+1)$*. Assume* $c_\eta > 0$*. On any event on which Proposition B.20 holds for both* $\pi = \widetilde{\pi}_T$ *and* $\pi = \pi^\star$*, where*

$$\widetilde{\pi}_T \in \arg\max_{\pi\in\Pi} \sum_{t=1}^T \left(P^\star(\pi \succ \pi_{t-1}) - \frac{1}{2} - \beta(D_{\mathrm{KL}}(\pi\|\pi_{\mathrm{ref}}) - D_{\mathrm{KL}}(\pi_{t-1}\|\pi_{\mathrm{ref}}))\right),$$

with ℓ_t *defined in Eq.* (29)*, we have*

$$\mathrm{Reg}_T \leq \frac{1+e^{R_\eta}/c_\eta}{\eta}\left[B + \frac{16}{\gamma}\mathrm{SEC}_{\mathrm{NLHF}}(\Psi,\lambda) + \frac{3\gamma\lambda T}{8} + \sum_{t=1}^T \epsilon_{m,t} + \frac{18}{\gamma\lambda}\sum_{t=1}^T \epsilon_{m,t}^2 + 4\gamma\eta^2\sum_{t=1}^T t\epsilon_{p,t}^2\right]. \tag{21}$$

Proof of Theorem B.11. Apply Lemma B.15 with comparator $\widetilde{\pi}_T$ and sum over t:

$$\begin{aligned}\eta\mathrm{Reg}_T &\leq \sum_{t=1}^T (D_{\mathrm{KL}}(\widetilde{\pi}_T\|\pi_{t-1}) - D_{\mathrm{KL}}(\widetilde{\pi}_T\|\pi_t^\star)) + \sum_{t=1}^T D_{\mathrm{KL}}(\pi_{t-1}\|\pi_t^\star)\\ &= D_{\mathrm{KL}}(\widetilde{\pi}_T\|\pi_0) - D_{\mathrm{KL}}(\widetilde{\pi}_T\|\pi_T) + E_T(\widetilde{\pi}_T) + \sum_{t=1}^T D_{\mathrm{KL}}(\pi_{t-1}\|\pi_t^\star).\end{aligned}$$

Since $\pi_0 = \pi_{\mathrm{ref}}$ and $\widetilde{\pi}_T \in \Pi$, Assumption B.5 gives $D_{\mathrm{KL}}(\widetilde{\pi}_T\|\pi_0) \leq B$. Also $D_{\mathrm{KL}}(\widetilde{\pi}_T\|\pi_T) \geq 0$. Therefore

$$\eta\mathrm{Reg}_T \leq B + E_T(\widetilde{\pi}_T) + \sum_{t=1}^T D_{\mathrm{KL}}(\pi_{t-1}\|\pi_t^\star).$$

By Lemma B.21,

$$\eta\mathrm{Reg}_T \leq B + E_T(\widetilde{\pi}_T) + \frac{e^{R_\eta}}{c_\eta}\left(D_{\mathrm{KL}}(\pi^\star\|\pi_0) + E_T(\pi^\star)\right).$$

Again $D_{\mathrm{KL}}(\pi^\star\|\pi_0) \leq B$. Apply Proposition B.20 to $\pi = \widetilde{\pi}_T$ and $\pi = \pi^\star$ to obtain Eq. (21). □

It remains to instantiate the deterministic bound with the concentration events proved in Appendix B.3.4 with specific choices of parameters.

Proof of Theorem B.10. Lemma B.24 has shown the following three concentration events hold with high probability, i.e., for all $\rho', \pi, \nu \in \Pi$,

$$\begin{aligned}&|N_t^{\rho',\pi}(\nu) - \widehat{N}_t^{\rho',\pi}(\nu)| \leq \epsilon_{m,t}, \quad \epsilon_{m,t} := B\sqrt{\frac{2L_\Pi}{m_t}};\\ &\frac{1}{2}(\lambda + H_t(\pi)) \leq \lambda + \widehat{H}_t(\pi) \leq 2(\lambda + H_t(\pi));\\ &|\widehat{P}_t(a \succ \pi_{t-1}) - P^\star(a \succ \pi_{t-1})| \leq \epsilon_{p,t}, \quad \epsilon_{p,t}^2 := \frac{L_P}{2k_t},\end{aligned}$$

where $N_t^{\rho',\pi}(\nu), \widehat{N}_t^{\rho',\pi}(\nu), \widehat{H}_t(\pi), H_t(\pi)$ are defined in Eqs. (39) and (40).

The first event controls the numerator of the explorer objective: the empirical extrapolation shift is close to its population value. The second one controls the denominator: the empirical coverage energy is within a constant factor of the population coverage energy after adding λ. The third one controls the preference estimates used in the learner's residual loss on the historical comparison data set. Together, these events allow replacing the population quantities in the deterministic regret bound by the empirical quantities computed by Algorithm 3.

On the above three events, Proposition B.20 is valid for every $\pi \in \Pi$, and it allows us to invoke the deterministic regret bound Theorem B.11,

$$\mathrm{Reg}_T \le \frac{1+e^{R_\eta}/c_\eta}{\eta}\left[B + \frac{16}{\gamma}S_T + \frac{3\gamma\lambda T}{8} + \sum_{t=1}^{T}\epsilon_{m,t} + \frac{18}{\gamma\lambda}\sum_{t=1}^{T}\epsilon_{m,t}^2 + 4\gamma\eta^2\sum_{t=1}^{T}t\epsilon_{p,t}^2\right],$$

where we use $S_T = 1 \vee \mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda)$. Since Eq. (63) gives $c_\eta = 1/2$ and $R_\eta = \eta(2V+1)$, the prefactor $(1+e^{R_\eta}/c_\eta)/\eta$ becomes $(1+2e^{\eta(2V+1)})/\eta$.

It remains to upper bound the bracket in Eq. (21). Let $A_T := \sqrt{S_T\lambda T}$ and we will try to bound each term in the bracket using a constant multiple of A_T. Since $\gamma = \sqrt{S_T/(\lambda T)}$, we have

$$\frac{16}{\gamma}S_T + \frac{3\gamma\lambda T}{8} = \left(16 + \frac{3}{8}\right)A_T. \tag{22}$$

Since we take $m_t = t$,

$$\epsilon_{m,t} = B\sqrt{\frac{2L_\Pi}{t}} \le \sqrt{\frac{2\overline{H}L_T}{t}},$$

where we used $B^2 \le \overline{H}$ and $L_\Pi \le L_T$. Hence, using $\sum_{t=1}^{T} t^{-1/2} \le 2\sqrt{T}$ and $\sum_{t=1}^{T} t^{-1} \le \log(eT) \le L_T$, we have

$$\sum_{t=1}^{T}\epsilon_{m,t} \le 2\sqrt{2\overline{H}L_T T}, \qquad \sum_{t=1}^{T}\epsilon_{m,t}^2 \le 2\overline{H}L_T^2. \tag{23}$$

Since $S_T \ge 1$, $L_T \ge 1$, and $\lambda = 16\overline{H}L_T^2$, the linear ϵ_m term satisfies

$$\sum_{t=1}^{T}\epsilon_{m,t} \le A_T. \tag{24}$$

For the quadratic ϵ_m term,

$$\frac{18}{\gamma\lambda}\sum_{t=1}^{T}\epsilon_{m,t}^2 \le \frac{36\overline{H}L_T^2}{\gamma\lambda} = \frac{36\overline{H}L_T^2}{S_T\lambda}A_T \le \frac{9}{4}A_T, \tag{25}$$

where the last inequality uses $S_T \ge 1$ and $\lambda = 16\overline{H}L_T^2$. Next, $k_t = t$ and $L_P \le L_T$, so

$$\sum_{t=1}^{T}t\epsilon_{p,t}^2 = \sum_{t=1}^{T}\frac{tL_P}{2k_t} \le \frac{TL_T}{2}. \tag{26}$$

Since $\eta \le 1$, $\overline{H} \ge 1$, and $L_T \ge 1$,

$$4\gamma\eta^2\sum_{t=1}^{T}t\epsilon_{p,t}^2 \le 2\gamma\eta^2 L_T T = \frac{2\eta^2 L_T}{\lambda}A_T \le \frac{1}{8}A_T. \tag{27}$$

Combining Eq. (22), Eq. (24), Eq. (25), and Eq. (27), the non-B part of the bracket in Eq. (21) is at most

$$\left(16 + \frac{3}{8} + 1 + \frac{9}{4} + \frac{1}{8}\right)A_T = 19.75A_T \le 20A_T.$$

The desired inequality Eq. (20) then follows by noticing that $V \ge 1$ and $\eta \le 1$. □

Theorem B.10 is stated under exact mirror-target realizability. We next consider a relaxed version: if realizability only holds away from a small neighborhood of $\pi^\star$, the same proof yields a best-candidate guarantee with an additional approximation term.

Define best response

$$\pi^{\mathrm{best}} \in \operatorname*{arg\,min}_{\pi \in \{\pi_0, \pi_1, \ldots, \pi_{T-1}, \bar{\pi}_T\}} \mathrm{gap}(\pi),$$

and the best-candidate regret

$$\mathrm{Reg}_T^{\mathrm{best}} := T\left(\max_{\nu \in \Pi} J_\beta(\nu, \pi^{\mathrm{best}}) - \frac{1}{2}\right) = \frac{T}{2}\mathrm{gap}(\pi^{\mathrm{best}}),$$

where the equality follows from Proposition B.1.

Corollary B.12 *Suppose Assumptions B.2, B.3, B.5 and B.7. Run Algorithm 3 for T rounds with the same choices of parameters as in Theorem B.10. Then, with probability at least $1-\delta$,*

$$\mathrm{Reg}_T^{\mathrm{best}} \le \max\left\{\frac{1+2e^{\eta(2V+1)}}{\eta}\left(B + 20\sqrt{S_T \lambda T}\right), \frac{T}{\sqrt{2}}(1+4V)\sqrt{\varepsilon_\eta}\right\}.$$

If $\varepsilon_\eta = O(1/T)$, then $\mathrm{Reg}_T^{\mathrm{best}} = \widetilde{O}(\sqrt{T})$.

Proof. Let $\mathcal{C}_T := \{\pi_0, \pi_1, \ldots, \pi_{T-1}, \bar{\pi}_T\}$. Define the first entrance time

$$\tau := \inf\left\{s \in \{0, \ldots, T-1\} : D_{\mathrm{KL}}(\pi^\star \| \pi_s) \le \varepsilon_\eta\right\}.$$

First suppose $\tau = \infty$. Then for every update round $t = 1, \ldots, T$, $D_{\mathrm{KL}}(\pi^\star \| \pi_{t-1}) > \varepsilon_\eta$. By relaxed realizability, $\pi^+_{\star,\eta}[\pi_{t-1}] \in \Pi$ for every realized round. Inspecting the proof of Theorem B.10, Assumption B.4 is used only to ensure that these realized exact mirror targets belong to Π. Therefore the proof of Theorem B.10 applies on the full horizon and gives the same trajectory regret Reg_T.

Since $\bar{\pi}_T \in \mathcal{C}_T$, by Proposition B.1, we have

$$\mathrm{Reg}_T^{\mathrm{best}} = \frac{T}{2}\mathrm{gap}(\pi^{\mathrm{best}}) \le \frac{T}{2}\mathrm{gap}(\bar{\pi}_T) \le \mathrm{Reg}_T.$$

Now suppose $\tau < \infty$. Then $D_{\mathrm{KL}}(\pi^\star \| \pi_\tau) \le \varepsilon_\eta$. By Lemma B.8,

$$\mathrm{gap}(\pi_\tau) \le \sqrt{2}(1+4V)\sqrt{\varepsilon_\eta}.$$

Since $\pi_\tau \in \mathcal{C}_T$, the best candidate satisfies

$$\mathrm{Reg}_T^{\mathrm{best}} = \frac{T}{2}\mathrm{gap}(\pi^{\mathrm{best}}) \le \frac{T}{2}\mathrm{gap}(\pi_\tau) \le \frac{T}{\sqrt{2}}(1+4V)\sqrt{\varepsilon_\eta}.$$

Combining the two cases gives

$$\mathrm{Reg}_T^{\mathrm{best}} \le \max\left\{\mathrm{Reg}_T, \frac{T}{\sqrt{2}}(1+4V)\sqrt{\varepsilon_\eta}\right\}.$$

Substituting the bound on Reg_T from Theorem B.10 proves the displayed high-probability bound. □

Finally, we can invoke a result in Xie et al. (2025) to upper bound $\mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda)$ by coverage and some logarithmic factor in T.

Corollary B.13 *For $\pi, \widehat{\pi} \in \Pi$, define the same-prompt pair distribution*

$$\nu_{\pi,\widehat{\pi}}(\mathbf{x}, \mathbf{y}, \mathbf{y}') := \rho(\mathbf{x})\pi(\mathbf{y}|\mathbf{x})\widehat{\pi}(\mathbf{y}'|\mathbf{x}).$$

Define

$$C_{\mathrm{cov}}^{\mathrm{pair}}(\Pi) := \inf_{\mu} \sup_{\pi, \widehat{\pi} \in \Pi} \left\| \frac{\nu_{\pi,\widehat{\pi}}}{\mu} \right\|_\infty.$$

Under the same assumptions as in Theorem B.10

$$\mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda) \le c\, C_{\mathrm{cov}}^{\mathrm{pair}}(\Pi) \log(eT),$$

where $c > 0$ is a universal constant.

Proof. This is a direct adaptation of Xie et al. (2025). □

B.3.4 Auxiliary Propositions and Lemmas for Theorem B.10

We now collect the auxiliary facts used in Appendix B.3.3. The first part concerns the population mirror update, the second part controls extrapolation and local empirical error, and the last part verifies the required concentration events.

The following result shows that exact mirror target has a simple residual structure: its population residual is constant across responses for the same prompt. This identity allows us to decompose KL error into a local term and an extrapolation term.

Proposition B.14 *For every $t \geq 1$ and every prompt $\mathbf{x}$, the function $\mathbf{y} \mapsto \delta((\mathbf{x}, \mathbf{y}); \pi_t^\star, \pi_{t-1})$ is constant over responses $\mathbf{y}$. More precisely,*

$$\delta((\mathbf{x}, \mathbf{y}); \pi_t^\star, \pi_{t-1}) = -\log Z_t(\mathbf{x}),$$

where $Z_t(\mathbf{x})$ is the denominator in Eq. (10) with $\pi = \pi_{t-1}$. Consequently, for any two prompt-response atoms $\tau = (\mathbf{x}, \mathbf{y})$ and $\tau' = (\mathbf{x}, \mathbf{y}')$ sharing the same prompt,

$$\Delta_t(\tau, \tau') = \log \frac{\pi_t(\tau)}{\pi_t^\star(\tau)} - \log \frac{\pi_t(\tau')}{\pi_t^\star(\tau')}.$$

Moreover, for any $\pi \in \Pi$,

$$D_{\mathrm{KL}}(\pi \| \pi_t) - D_{\mathrm{KL}}(\pi \| \pi_t^\star) = \mathbb{E}_{\tau \sim \pi_{t-1}} \left[\log \pi_t^\star(\tau) - \log \pi_t(\tau)\right] + \mathbb{E}_{\tau \sim \pi_{t-1}, \tau' \sim \pi} \left[\Delta_t(\tau, \tau')\right]. \quad (28)$$

Proof. Fix a prompt $\mathbf{x}$. By Eq. (10),

$$\log \pi_t^\star(\mathbf{y} \mid \mathbf{x}) = \eta\beta \log \pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x}) + (1 - \eta\beta) \log \pi_{t-1}(\mathbf{y} \mid \mathbf{x}) + \eta P^\star(\mathbf{y} \succ \pi_{t-1} \mid \mathbf{x}) - \log Z_t(\mathbf{x}).$$

Substituting this display into Eq. (16) gives

$$\delta((\mathbf{x}, \mathbf{y}); \pi_t^\star, \pi_{t-1}) = -\log Z_t(\mathbf{x}),$$

which is independent of $\mathbf{y}$ for the fixed prompt $\mathbf{x}$. Therefore

$$\delta_t(\mathbf{x}, \mathbf{y}) - \delta((\mathbf{x}, \mathbf{y}); \pi_t^\star, \pi_{t-1}) = \log \frac{\pi_t(\mathbf{y} \mid \mathbf{x})}{\pi_t^\star(\mathbf{y} \mid \mathbf{x})}.$$

Taking differences between $\mathbf{y}$ and $\mathbf{y}'$ proves the first identity. For Eq. (28), note that all policies have the same prompt marginal ρ, so the conditional constant $-\log Z_t(\mathbf{x})$ has the same expectation under π and under π_{t-1}. Thus

$$D_{\mathrm{KL}}(\pi \| \pi_t) - D_{\mathrm{KL}}(\pi \| \pi_t^\star) = \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y}' \sim \pi(\cdot|\mathbf{x})} \left[\log \frac{\pi_t^\star(\mathbf{y}' \mid \mathbf{x})}{\pi_t(\mathbf{y}' \mid \mathbf{x})}\right]$$

can be rewritten by adding and subtracting the expectation under $\mathbf{y} \sim \pi_{t-1}(\cdot \mid \mathbf{x})$, yielding Eq. (28). □

The next lemma is the one-step mirror-descent inequality. It converts regularized gain against any comparator into a KL telescoping term plus the movement of the exact mirror target.

Lemma B.15 *For every $\pi \in \Pi$ and every $t \geq 1$, define*

$$\ell_t(\pi) := P^\star(\pi \succ \pi_{t-1}) - \frac{1}{2} - \beta(D_{\mathrm{KL}}(\pi \| \pi_{\mathrm{ref}}) - D_{\mathrm{KL}}(\pi_{t-1} \| \pi_{\mathrm{ref}})).$$

Then

$$\eta \ell_t(\pi) \leq D_{\mathrm{KL}}(\pi \| \pi_{t-1}) - D_{\mathrm{KL}}(\pi \| \pi_t^\star) + D_{\mathrm{KL}}(\pi_{t-1} \| \pi_t^\star). \quad (29)$$

Proof. Let

$$g_t(\tau) := P^\star(\tau \succ \pi_{t-1}) - \beta \log \frac{\pi_{t-1}(\tau)}{\pi_{\mathrm{ref}}(\tau)}.$$

The exact update satisfies

$$\pi_t^\star = \arg\min_{u \in \Delta(\mathcal{Y})} \{\langle -\eta g_t, u \rangle + D_{\mathrm{KL}}(u \| \pi_{t-1})\}.$$

The Bregman three-point inequality for the KL prox step gives, for all π,

$$\eta\langle g_t, \pi - \pi_t^\star\rangle \le D_{\mathrm{KL}}(\pi\|\pi_{t-1}) - D_{\mathrm{KL}}(\pi\|\pi_t^\star) - D_{\mathrm{KL}}(\pi_t^\star\|\pi_{t-1}). \tag{30}$$

Add $\eta\langle g_t, \pi_t^\star - \pi_{t-1}\rangle$ to both sides. The closed form gives

$$\eta g_t(\tau) = \log\frac{\pi_t^\star(\tau)}{\pi_{t-1}(\tau)} + \log Z_t(x),$$

so the constant cancels when integrated against $\pi_t^\star - \pi_{t-1}$. Hence

$$\eta\langle g_t, \pi_t^\star - \pi_{t-1}\rangle = \sum_\tau(\pi_t^\star(\tau) - \pi_{t-1}(\tau))\log\frac{\pi_t^\star(\tau)}{\pi_{t-1}(\tau)} = D_{\mathrm{KL}}(\pi_t^\star\|\pi_{t-1}) + D_{\mathrm{KL}}(\pi_{t-1}\|\pi_t^\star).$$

Combining with Eq. (30),

$$\eta\langle g_t, \pi - \pi_{t-1}\rangle \le D_{\mathrm{KL}}(\pi\|\pi_{t-1}) - D_{\mathrm{KL}}(\pi\|\pi_t^\star) + D_{\mathrm{KL}}(\pi_{t-1}\|\pi_t^\star).$$

Finally,

$$\begin{aligned}\langle g_t, \pi - \pi_{t-1}\rangle &= P^\star(\pi \succ \pi_{t-1}) - \frac{1}{2} - \beta(D_{\mathrm{KL}}(\pi\|\pi_{\mathrm{ref}}) - D_{\mathrm{KL}}(\pi_{t-1}\|\pi_{\mathrm{ref}})) + \beta D_{\mathrm{KL}}(\pi\|\pi_{t-1})\\ &\ge \ell_t(\pi),\end{aligned}$$

which proves the claim. □

We next use the KL regularization to show that the exact population mirror map contracts toward the Nash policy. This contraction is the key reason a constant stepsize is possible.

Lemma B.16 (Contraction of mirror descent with constant stepsize) *Let $\pi^\star$ be the unique symmetric Nash equilibrium of the KL-regularized game. Let $\pi^+ := \pi^+_{*,\eta}(\pi)$ be defined by Eq.* (10)*. If*

$$0 < \eta \le \frac{4\beta}{1+4\beta^2} \le 1, \tag{31}$$

then, with $c_\eta := 1 - \eta\beta - \eta/(4\beta) \ge 0$,

$$D_{\mathrm{KL}}(\pi^\star\|\pi^+) \le (1-\eta\beta)D_{\mathrm{KL}}(\pi^\star\|\pi) - c_\eta D_{\mathrm{KL}}(\pi^+\|\pi). \tag{32}$$

In particular, $D_{\mathrm{KL}}(\pi^\star\|\pi^+) \le (1-\eta\beta)D_{\mathrm{KL}}(\pi^\star\|\pi)$.

Proof. The proof is pointwise in the prompt $\mathbf{x}$; we suppress $\mathbf{x}$ and write distributions over the finite response set $\mathcal{Y}$. Integrating the resulting inequality over $\mathbf{x}\sim\rho$ gives the stated prompt-averaged KL inequality. Let $h(p) = \sum_\tau p(\tau)\log p(\tau)$ be negative entropy. Define the operator

$$F(p) = A(p) + B(p), \qquad A(p)(\tau) := -P^\star(\tau \succ p), \qquad B(p) := \beta(\nabla h(p) - \nabla h(\pi_{\mathrm{ref}})).$$

The Nash equilibrium $\pi^\star$ satisfies the variational inequality

$$\langle F(\pi^\star), u - \pi^\star\rangle \ge 0 \qquad \forall u \in \Delta(\mathcal{Y}). \tag{33}$$

The exact update is equivalently

$$\pi^+ = \operatorname*{arg\,min}_{u\in\Delta(\mathcal{Y})}\left\{\eta\langle F(\pi), u\rangle + D_{\mathrm{KL}}(u\|\pi)\right\}. \tag{34}$$

Indeed, expanding F gives exactly Eq. (10) up to constants independent of u.

We first record three structural facts. Let P be the matrix with entries $P_{\tau,u} = P^\star(\tau \succ u)$. Assumption B.2 gives $P + P^\top = \mathbf{1}\mathbf{1}^\top$. Since both $u - v$ and $\pi - v$ have zero sum for distributions, for all $u, v \in \Delta(\mathcal{Y})$,

$$\langle A(u) - A(v), u - v\rangle = -(u-v)^\top P(u-v) = -\frac{1}{2}(u-v)^\top(P+P^\top)(u-v) = 0.$$

Thus A is monotone with equality. Also,

$$\|A(u) - A(v)\|_\infty = \max_\tau\left|\sum_a\left(P_{\tau,a} - \frac{1}{2}\right)(v(a) - u(a))\right| \le \frac{1}{2}\|u - v\|_1.$$

Finally, the entropy identity gives

$$\langle B(u) - B(v), u - v\rangle = \beta\big(D_{\mathrm{KL}}(u\|v) + D_{\mathrm{KL}}(v\|u)\big). \tag{35}$$

Using

$$\eta\langle B(\pi), u\rangle = \eta\beta D_{\mathrm{KL}}(u\|\pi_{\mathrm{ref}}) - \eta\beta D_{\mathrm{KL}}(u\|\pi) + \text{constant in } u,$$

we can write the prox objective as

$$\Psi(u) = \eta\langle A(\pi), u\rangle + \eta\beta D_{\mathrm{KL}}(u\|\pi_{\mathrm{ref}}) + (1-\eta\beta)D_{\mathrm{KL}}(u\|\pi),$$

where $\eta\beta \le 1$ follows from Eq. (31). The two KL weights sum to one, hence Ψ is 1-strongly convex with respect to the entropy Bregman divergence. Since π^+ minimizes Ψ, the Bregman three-point inequality for a 1-strongly convex function with respect to negative entropy gives

$$\Psi(\pi^\star) - \Psi(\pi^+) \ge D_{\mathrm{KL}}(\pi^\star\|\pi^+). \tag{36}$$

On the other hand, Eq. (33) and the strong convexity of $u \mapsto \langle A(\pi^\star), u\rangle + \beta D_{\mathrm{KL}}(u\|\pi_{\mathrm{ref}})$ imply

$$\eta\beta\big(D_{\mathrm{KL}}(\pi^+\|\pi_{\mathrm{ref}}) - D_{\mathrm{KL}}(\pi^\star\|\pi_{\mathrm{ref}})\big) \ge \eta\langle A(\pi^\star), \pi^\star - \pi^+\rangle + \eta\beta D_{\mathrm{KL}}(\pi^+\|\pi^\star). \tag{37}$$

Adding Eq. (36) and Eq. (37), the $D_{\mathrm{KL}}(\cdot\|\pi_{\mathrm{ref}})$ terms cancel and we obtain

$$\begin{aligned} D_{\mathrm{KL}}(\pi^\star\|\pi^+) \le{}& \eta\langle A(\pi) - A(\pi^\star), \pi^\star - \pi^+\rangle + (1-\eta\beta)D_{\mathrm{KL}}(\pi^\star\|\pi) \\ &- (1-\eta\beta)D_{\mathrm{KL}}(\pi^+\|\pi) - \eta\beta D_{\mathrm{KL}}(\pi^+\|\pi^\star). \end{aligned} \tag{38}$$

Insert and subtract $A(\pi^+)$ in the inner product. By monotonicity with equality,

$$\langle A(\pi) - A(\pi^\star), \pi^\star - \pi^+\rangle = \langle A(\pi^+) - A(\pi), \pi^+ - \pi^\star\rangle.$$

Using the Lipschitz property of A, Hölder's inequality, and Pinsker's inequality,

$$\begin{aligned} \eta\left|\langle A(\pi^+) - A(\pi), \pi^+ - \pi^\star\rangle\right| &\le \frac{\eta}{2}\|\pi^+ - \pi\|_1\|\pi^+ - \pi^\star\|_1 \\ &\le \eta\sqrt{D_{\mathrm{KL}}(\pi^+\|\pi)D_{\mathrm{KL}}(\pi^+\|\pi^\star)} \\ &\le \frac{\eta}{4\beta}D_{\mathrm{KL}}(\pi^+\|\pi) + \eta\beta D_{\mathrm{KL}}(\pi^+\|\pi^\star), \end{aligned}$$

where the last step applies the AM–GM inequality. Substituting this into Eq. (38) cancels the $\eta\beta D_{\mathrm{KL}}(\pi^+\|\pi^\star)$ term and yields

$$D_{\mathrm{KL}}(\pi^\star\|\pi^+) \le (1-\eta\beta)D_{\mathrm{KL}}(\pi^\star\|\pi) - \left(1 - \eta\beta - \frac{\eta}{4\beta}\right)D_{\mathrm{KL}}(\pi^+\|\pi).$$

This proves Eq. (32). □

The previous contraction controls one KL direction. The next lemma compares the reverse direction using the fact that one exact mirror step changes log-probabilities by a bounded amount.

Lemma B.17 *For every $t \ge 1$,*

$$D_{\mathrm{KL}}(\pi_{t-1}\|\pi_t^\star) \le e^{R_\eta}D_{\mathrm{KL}}(\pi_t^\star\|\pi_{t-1}), \qquad R_\eta := \eta(2V+1).$$

Proof. By the closed form Eq. (10),

$$\log\frac{\pi_t^\star(\mathbf{y}\mid\mathbf{x})}{\pi_{t-1}(\mathbf{y}\mid\mathbf{x})} = a_t(\mathbf{x},\mathbf{y}) - \log Z_t(\mathbf{x}), \qquad Z_t(\mathbf{x}) = \sum_u \pi_{t-1}(u\mid\mathbf{x})e^{a_t(\mathbf{x},u)}.$$

where

$$a_t(\mathbf{x},\mathbf{y}) := \eta\beta\log\frac{\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}{\pi_{t-1}(\mathbf{y}\mid\mathbf{x})} + \eta P^\star(\mathbf{y}\succ\pi_{t-1}\mid\mathbf{x}),$$

The range of $\eta\beta\log(\pi_{\mathrm{ref}}/\pi_{t-1})$ over τ is at most $2\eta V$ by Assumption B.5, and the range of $\eta P^\star(\tau\succ\pi_{t-1})$ is at most η. Since $Z_t(\mathbf{x})$ is a $\pi_{t-1}(\cdot\mid\mathbf{x})$-weighted average of $e^{a_t(\mathbf{x},u)}$, $\log Z_t(\mathbf{x})$ lies between $\min_u a_t(\mathbf{x},u)$ and $\max_u a_t(\mathbf{x},u)$. Therefore

$$\left|\log\frac{\pi_t^\star(\mathbf{y}\mid\mathbf{x})}{\pi_{t-1}(\mathbf{y}\mid\mathbf{x})}\right| \le \max_u a_t(\mathbf{x},u) - \min_u a_t(\mathbf{x},u) \le \eta(2V+1) =: R_\eta.$$

Let $p = \pi_t^\star$, $\bar{\pi} = \pi_{t-1}$, and $r = p/\bar{\pi}$. Then $r \in [e^{-R_\eta}, e^{R_\eta}]$ pointwise. The f-divergence representation gives

$$D_{\mathrm{KL}}(\bar{\pi}\|p) = \mathbb{E}_{\bar{\pi}}[r - 1 - \log r], \qquad D_{\mathrm{KL}}(p\|\bar{\pi}) = \mathbb{E}_{\bar{\pi}}[r\log r - r + 1].$$

For $r \in [e^{-R_\eta}, e^{R_\eta}]$, the functions $f(r) = r - 1 - \log r$ and $g(r) = r\log r - r + 1$ satisfy $f(1) = g(1) = f'(1) = g'(1) = 0$ and

$$f''(r) = \frac{1}{r^2} \le e^{R_\eta}\frac{1}{r} = e^{R_\eta} g''(r).$$

Integrating twice around $r = 1$ gives $f(r) \le e^{R_\eta} g(r)$ for all such r. Taking expectation under $\bar{\pi}$ proves the claim. □

For $\pi \in \Pi$, define pairwise log-ratio direction

$$\psi_{t,\pi}(\tau, \tau') := \log\frac{\pi_t(\tau)}{\pi(\tau)} - \log\frac{\pi_t(\tau')}{\pi(\tau')}.$$

For concentration statements we also need a version with an arbitrary numerator policy $\rho' \in \Pi$:

$$\psi_{\rho',\pi}(\tau, \tau') := \log\frac{\rho'(\tau)}{\pi(\tau)} - \log\frac{\rho'(\tau')}{\pi(\tau')}.$$

We now introduce the quantities optimized by the adversarial explorer. The numerator measures extrapolation to a candidate opponent, while the denominator measures how well past exploratory data cover the corresponding log-ratio direction.

For $\rho', \pi, \nu \in \Pi$, define

$$\begin{aligned}
N_t^{\rho',\pi}(\nu) &:= \mathbb{E}_{\tau\sim\pi_{t-1},\tau'\sim\nu}\psi_{\rho',\pi}(\tau,\tau'),\\
\widehat{N}_t^{\rho',\pi}(\nu) &:= \frac{1}{m_t}\sum_{i=1}^{m_t}\log\frac{\rho'(\widetilde{\tau}_{t-1,i})}{\pi(\widetilde{\tau}_{t-1,i})} - \mathbb{E}_{\tau'\sim\nu}\log\frac{\rho'(\tau')}{\pi(\tau')},\\
H_t^{\rho',\pi} &:= \sum_{s=0}^{t-1}\mathbb{E}_{\tau\sim\pi_{s-1},\tau'\sim\widehat{\pi}_s}\psi_{\rho',\pi}(\tau,\tau')^2,\\
\widehat{H}_t^{\rho',\pi} &:= \sum_{s=0}^{t-1}\frac{1}{n_{s+1}}\sum_{i=1}^{n_{s+1}}\psi_{\rho',\pi}(\tau_{s,i},\tau'_{s,i})^2.
\end{aligned} \tag{39}$$

Here $N_t^{\rho',\pi}(\nu)$ denotes the population extrapolation shift, measuring how much the log-ratio direction changes when moving from π_{t-1} to candidate opponent ν, and $\widehat{N}_t^{\rho',\pi}(\nu)$ is its empirical version; $H_t^{\rho',\pi}$ denotes the population coverage energy, accumulating squared size of this direction under past exploration pairs.

We can also define their round-t version

$$N_t(\nu,\pi) := N_t^{\pi_t,\pi}(\nu), \widehat{N}_t(\nu,\pi) := \widehat{N}_t^{\pi_t,\pi}(\nu), H_t(\pi) := H_t^{\pi_t,\pi}, \text{ and } \widehat{H}_t(\pi) := \widehat{H}_t^{\pi_t,\pi}. \tag{40}$$

The following lemma bounds the extrapolation error by the SEC term, up to empirical estimation errors and a quadratic remainder.

Lemma B.18 *Assume that for every $t \le T$, every $\pi \in \Pi$, and every $\nu \in \Pi$,*

$$|N_t(\nu,\pi) - \widehat{N}_t(\nu,\pi)| \le \epsilon_{m,t}, \qquad \frac{1}{2}(\lambda + H_t(\pi)) \le \lambda + \widehat{H}_t(\pi) \le 2(\lambda + H_t(\pi)), \tag{41}$$

where $N_t(\nu,\pi)$, $\widehat{N}_t(\nu,\pi)$, $H_t(\pi)$, $\widehat{H}_t(\pi)$ are defined in Eqs. (39) *and* (40)*. Then for every $\pi \in \Pi$,*

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\pi_{t-1},\tau'\sim\pi}\Delta_t(\tau,\tau') \le \frac{16}{\gamma}\mathrm{SEC}_{\mathrm{NLHF}}(\Psi,\lambda) + \frac{\gamma\lambda T}{4} + \frac{18}{\gamma\lambda}\sum_{t=1}^{T}\epsilon_{m,t}^2\\
+ \frac{\gamma}{4}\sum_{t=1}^{T}\sum_{s=0}^{t-1}\mathbb{E}_{\tau\sim\pi_{s-1},\tau'\sim\widehat{\pi}_s}\Delta_t(\tau,\tau')^2.
\end{aligned} \tag{42}$$

Proof. By Proposition B.14, $\Delta_t(\tau, \tau') = \psi_{t,\pi_t^\star}(\tau, \tau')$, and this function belongs to Ψ by Assumption B.4. Define $D_t := \lambda + H_t(\pi_t^\star)$. For any real x and $a > 0$, $x \le x^2/(\gamma a) + \gamma a/4$. Therefore,

$$\mathbb{E}_{\pi_{t-1},\pi}\Delta_t \le \frac{N_t(\pi, \pi_t^\star)^2}{\gamma D_t} + \frac{\gamma\lambda}{4} + \frac{\gamma}{4}H_t(\pi_t^\star). \tag{43}$$

Using Eq. (41),

$$\frac{N_t(\pi, \pi_t^\star)^2}{\gamma D_t} \le \frac{2\widehat{N}_t(\pi, \pi_t^\star)^2}{\gamma D_t} + \frac{2\epsilon_{m,t}^2}{\gamma\lambda} \le \frac{4\widehat{N}_t(\pi, \pi_t^\star)^2}{\gamma(\lambda + \widehat{H}_t(\pi_t^\star))} + \frac{2\epsilon_{m,t}^2}{\gamma\lambda}.$$

By the definition of $\widehat{\pi}_t$ in Algorithm 3,

$$\frac{\widehat{N}_t(\pi, \pi_t^\star)^2}{\lambda + \widehat{H}_t(\pi_t^\star)} \le \max_{r\in\Pi} \frac{\widehat{N}_t(\widehat{\pi}_t, r)^2}{\lambda + \widehat{H}_t(r)}.$$

Again using Eq. (41), for every $r \in \Pi$,

$$\frac{\widehat{N}_t(\widehat{\pi}_t, r)^2}{\lambda + \widehat{H}_t(r)} \le 4\frac{N_t(\widehat{\pi}_t, r)^2}{\lambda + H_t(r)} + \frac{4\epsilon_{m,t}^2}{\lambda}.$$

Combining the last three displays yields

$$\frac{N_t(\pi, \pi_t^\star)^2}{\gamma D_t} \le \frac{16}{\gamma}\max_{r\in\Pi} \frac{N_t(\widehat{\pi}_t, r)^2}{\lambda + H_t(r)} + \frac{18\epsilon_{m,t}^2}{\gamma\lambda}.$$

Substitute this into Eq. (43) and sum over t. For each $r \in \Pi$, $H_t(r)$ contains the SEC denominator sum $\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{s-1},\widehat{\pi}_s}[\psi_{t,r}^2]$ and possibly the additional nonnegative $s = 0$ term. Hence

$$\lambda + H_t(r) \ge \lambda \vee \sum_{s=1}^{t-1} \mathbb{E}_{\tau\sim\pi_{s-1},\tau'\sim\widehat{\pi}_s} \psi_{t,r}(\tau, \tau')^2.$$

By the SEC definition Eq. (19), the sum of the max-ratio terms is bounded by $\mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda)$. This proves Eq. (42). □

The preceding lemma controls the extrapolation part but leaves a quadratic remainder. The next lemma controls the local, on-sample part of the update and produces a negative copy of the same quadratic term. This is where the learner's empirical risk minimization step is used: because π_t minimizes the empirical residual loss, it cannot be much worse than the exact mirror target $\pi_t^\star$ on the sampled data. The remaining error comes only from finite-sample estimation of preferences.

Lemma B.19 *Fix $t \ge 1$. Suppose the following events hold:*

$$\mathbb{E}_{\tau\sim\pi_{t-1}}[\log \pi_t^\star(\tau) - \log \pi_t(\tau)] \le \frac{1}{m_t}\sum_{i=1}^{m_t}[\log \pi_t^\star(\widetilde{\tau}_{t-1,i}) - \log \pi_t(\widetilde{\tau}_{t-1,i})] + \epsilon_{m,t}, \tag{44}$$

$$\sum_{s=0}^{t-1} \mathbb{E}_{\pi_{s-1},\widehat{\pi}_s} \Delta_t(\tau, \tau')^2 \le \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \Delta_t(\tau_{s,i}, \tau'_{s,i})^2 + \frac{\lambda}{2}, \tag{45}$$

and for every empirical historical pair, with

$$e_t(a) := \widehat{P}_t(a \succ \pi_{t-1}) - P^\star(a \succ \pi_{t-1}),$$

we have

$$|e_t(a)| \le \epsilon_{p,t} \qquad \textit{for all } a \in \{\tau_{s,i}, \tau'_{s,i} : 0 \le s < t,\ 1 \le i \le n_{s+1}\}. \tag{46}$$

Then, under Eq. (14)*,*

$$\mathbb{E}_{\tau\sim\pi_{t-1}}[\log \pi_t^\star(\tau) - \log \pi_t(\tau)] + \frac{\gamma}{4}\sum_{s=0}^{t-1} \mathbb{E}_{\pi_{s-1},\widehat{\pi}_s} \Delta_t(\tau, \tau')^2 \le \epsilon_{m,t} + \frac{\gamma\lambda}{8} + 4\gamma\eta^2 t\epsilon_{p,t}^2. \tag{47}$$

Proof. For brevity, define the empirical averages

$$\widehat{S}_t(\pi) := \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \widehat{\Delta}_t(\tau_{s,i}, \tau'_{s,i}; \pi)^2,$$

$$S_t(\pi) := \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \Delta_t(\tau_{s,i}, \tau'_{s,i}; \pi)^2.$$

From Eq. (44) and Eq. (45),

$$\begin{aligned}
&\mathbb{E}_{\pi_{t-1}}[\log \pi_t^\star - \log \pi_t] + \frac{\gamma}{4} \sum_{s=0}^{t-1} \mathbb{E}_{\pi_{s-1}, \widehat{\pi}_s} \Delta_t^2 \\
&\qquad \le \frac{1}{m_t} \sum_{i=1}^{m_t} [\log \pi_t^\star(\widetilde{\tau}_{t-1,i}) - \log \pi_t(\widetilde{\tau}_{t-1,i})] + \frac{\gamma}{4} S_t(\pi_t) + \epsilon_{m,t} + \frac{\gamma\lambda}{8}. \qquad (48)
\end{aligned}$$

The ERM optimality of π_t against the feasible comparator $\pi_t^\star$ gives

$$\frac{1}{m_t} \sum_{i=1}^{m_t} [\log \pi_t^\star(\widetilde{\tau}_{t-1,i}) - \log \pi_t(\widetilde{\tau}_{t-1,i})] + \frac{\gamma}{2} \widehat{S}_t(\pi_t) \le \frac{\gamma}{2} \widehat{S}_t(\pi_t^\star). \qquad (49)$$

This is exactly where $\alpha_t = 2/(\gamma m_t)$ is used. For an empirical pair (a, b), write

$$u := \Delta_t(a, b; \pi_t), \qquad z := \eta(e_t(a) - e_t(b)).$$

Then $\widehat{\Delta}_t(a, b; \pi_t) = u - z$. Also, by Proposition B.14, $\Delta_t(a, b; \pi_t^\star) = 0$, and hence $\widehat{\Delta}_t(a, b; \pi_t^\star) = -z$. Combining Eq. (49) with the term $\frac{\gamma}{4} S_t(\pi_t)$ in Eq. (48), pair by pair we get the upper bound

$$\frac{\gamma}{2} z^2 + \frac{\gamma}{4} u^2 - \frac{\gamma}{2} (u - z)^2 = \gamma u z - \frac{\gamma}{4} u^2 \le \gamma z^2,$$

where the last inequality follows from $uz - u^2/4 \le z^2$. Therefore

$$\frac{1}{m_t} \sum_{i=1}^{m_t} [\log \pi_t^\star(\widetilde{\tau}_{t-1,i}) - \log \pi_t(\widetilde{\tau}_{t-1,i})] + \frac{\gamma}{4} S_t(\pi_t) \le \gamma\eta^2 \sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} \big(e_t(\tau_{s,i}) - e_t(\tau'_{s,i})\big)^2.$$

By Eq. (46), each squared difference is at most $4\epsilon_{p,t}^2$. Since

$$\sum_{s=0}^{t-1} \frac{1}{n_{s+1}} \sum_{i=1}^{n_{s+1}} 1 = t,$$

the last display is bounded by $4\gamma\eta^2 t \epsilon_{p,t}^2$. Substituting into Eq. (48) proves Eq. (47). □

Combining the extrapolation bound with the local bound cancels the shared quadratic remainder. This gives a uniform bound on the accumulated inexact-update error $E_T(\pi)$.

Proposition B.20 *Assume the events Eqs.* (41) *and* (44) *to* (46) *hold for every $t \le T$. Then for every fixed comparator $\pi \in \Pi$,*

$$\begin{aligned}
E_T(\pi) &:= \sum_{t=1}^{T} \big(D_{\mathrm{KL}}(\pi \| \pi_t) - D_{\mathrm{KL}}(\pi \| \pi_t^\star)\big) \\
&\le \frac{16}{\gamma} \mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda) + \frac{3\gamma\lambda T}{8} + \sum_{t=1}^{T} \epsilon_{m,t} + \frac{18}{\gamma\lambda} \sum_{t=1}^{T} \epsilon_{m,t}^2 + 4\gamma\eta^2 \sum_{t=1}^{T} t\epsilon_{p,t}^2. \qquad (50)
\end{aligned}$$

Proof. Fix a comparator $\pi \in \Pi$. Throughout this proof, for each t, write

$$\Delta_t(\tau, \tau') := \delta_t(\tau) - \delta_t(\tau').$$

Also define the accumulated population quadratic term

$$Q_t := \sum_{s=0}^{t-1} \mathbb{E}_{\tau\sim\pi_{s-1},\,\tau'\sim\widehat{\pi}_s}\left[\Delta_t(\tau,\tau')^2\right],$$

and

$$Q_T^{\mathrm{tot}} := \sum_{t=1}^{T} Q_t = \sum_{t=1}^{T}\sum_{s=0}^{t-1} \mathbb{E}_{\tau\sim\pi_{s-1},\,\tau'\sim\widehat{\pi}_s}\left[\Delta_t(\tau,\tau')^2\right].$$

We start from the definition of $E_T(\pi)$. For every t,

$$D_{\mathrm{KL}}(\pi\|\pi_t) - D_{\mathrm{KL}}(\pi\|\pi_t^\star) = \mathbb{E}_{\tau\sim\pi}\left[\log\frac{\pi_t^\star(\tau)}{\pi_t(\tau)}\right].$$

Therefore,

$$E_T(\pi) = \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right].$$

Next we use the residual decomposition from Eq. (28). For completeness, we spell out the identity. By the residual identity,

$$\log\pi_t^\star(\tau) - \log\pi_t(\tau) = \delta_t^\star(\tau) - \delta_t(\tau),$$

and by Proposition B.14, $\delta_t^\star(\cdot)$ is constant in τ. Hence

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right] &= \mathbb{E}_{\tau\sim\pi}\left[\delta_t^\star(\tau) - \delta_t(\tau)\right] \\
&= \mathbb{E}_{\tau\sim\pi_{t-1}}\left[\delta_t^\star(\tau) - \delta_t(\tau)\right] + \mathbb{E}_{\tau\sim\pi_{t-1}}\left[\delta_t(\tau)\right] - \mathbb{E}_{\tau'\sim\pi}\left[\delta_t(\tau')\right] \\
&= \mathbb{E}_{\tau\sim\pi_{t-1}}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right] + \mathbb{E}_{\tau\sim\pi_{t-1},\,\tau'\sim\pi}\left[\delta_t(\tau) - \delta_t(\tau')\right] \\
&= \mathbb{E}_{\tau\sim\pi_{t-1}}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right] + \mathbb{E}_{\tau\sim\pi_{t-1},\,\tau'\sim\pi}\left[\Delta_t(\tau,\tau')\right].
\end{aligned}$$

Summing this equality over $t = 1,\dots,T$ gives

$$E_T(\pi) = \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_{t-1}}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right]}_{=:L_T} + \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_{t-1},\,\tau'\sim\pi}\left[\Delta_t(\tau,\tau')\right]}_{=:X_T}. \tag{51}$$

We now bound the extrapolation part X_T. Since the event Eq. (41) holds for every $t \le T$, the hypotheses of Lemma B.18 are satisfied. Applying Lemma B.18 with the fixed comparator π gives

$$\begin{aligned}
X_T &= \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_{t-1},\,\tau'\sim\pi}\left[\Delta_t(\tau,\tau')\right] \\
&\le \frac{16}{\gamma}\mathrm{SEC}_{\mathrm{NLHF}}(\Psi,\lambda) + \frac{\gamma\lambda T}{4} + \frac{18}{\gamma\lambda}\sum_{t=1}^{T}\epsilon_{m,t}^2 + \frac{\gamma}{4}Q_T^{\mathrm{tot}}.
\end{aligned} \tag{52}$$

We next bound the local part L_T. Since the events Eqs. (44) to (46) hold for every $t \le T$, the hypotheses of Lemma B.19 are satisfied at every round t. Thus, for each t,

$$\mathbb{E}_{\tau\sim\pi_{t-1}}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right] + \frac{\gamma}{4}Q_t \le \epsilon_{m,t} + \frac{\gamma\lambda}{8} + 4\gamma\eta^2 t\epsilon_{p,t}^2. \tag{53}$$

Rearranging Eq. (53) gives

$$\mathbb{E}_{\tau\sim\pi_{t-1}}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right] \le \epsilon_{m,t} + \frac{\gamma\lambda}{8} + 4\gamma\eta^2 t\epsilon_{p,t}^2 - \frac{\gamma}{4}Q_t.$$

Summing this inequality over $t = 1,\dots,T$, we obtain

$$\begin{aligned}
L_T &= \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_{t-1}}\left[\log\pi_t^\star(\tau) - \log\pi_t(\tau)\right] \\
&\le \sum_{t=1}^{T}\epsilon_{m,t} + \frac{\gamma\lambda T}{8} + 4\gamma\eta^2\sum_{t=1}^{T} t\epsilon_{p,t}^2 - \frac{\gamma}{4}\sum_{t=1}^{T} Q_t \\
&= \sum_{t=1}^{T}\epsilon_{m,t} + \frac{\gamma\lambda T}{8} + 4\gamma\eta^2\sum_{t=1}^{T} t\epsilon_{p,t}^2 - \frac{\gamma}{4}Q_T^{\mathrm{tot}}.
\end{aligned} \tag{54}$$

Finally, substitute Eq. (52) and Eq. (54) into the decomposition Eq. (51). We get

$$\begin{aligned} E_T(\pi) &= L_T + X_T \\ &\le \left[\sum_{t=1}^{T} \epsilon_{m,t} + \frac{\gamma\lambda T}{8} + 4\gamma\eta^2 \sum_{t=1}^{T} t\epsilon_{p,t}^2 - \frac{\gamma}{4} Q_T^{\mathrm{tot}}\right] \\ &\quad + \left[\frac{16}{\gamma}\mathrm{SEC}_{\mathrm{NLHF}}(\Psi,\lambda) + \frac{\gamma\lambda T}{4} + \frac{18}{\gamma\lambda}\sum_{t=1}^{T}\epsilon_{m,t}^2 + \frac{\gamma}{4} Q_T^{\mathrm{tot}}\right]. \end{aligned}$$

Therefore,

$$E_T(\pi) \le \frac{16}{\gamma}\mathrm{SEC}_{\mathrm{NLHF}}(\Psi,\lambda) + \frac{3\gamma\lambda T}{8} + \sum_{t=1}^{T}\epsilon_{m,t} + \frac{18}{\gamma\lambda}\sum_{t=1}^{T}\epsilon_{m,t}^2 + 4\gamma\eta^2\sum_{t=1}^{T} t\epsilon_{p,t}^2,$$

which is exactly Eq. (50). □

The deterministic regret bound still contains the total movement of the exact mirror targets. The next lemma controls this movement using the contraction of the population mirror map.

Lemma B.21 *Let*

$$E_T(\pi^\star) := \sum_{t=1}^{T} \left(D_{\mathrm{KL}}(\pi^\star\|\pi_t) - D_{\mathrm{KL}}(\pi^\star\|\pi_t^\star)\right).$$

Under Eq. (31)*,*

$$\sum_{t=1}^{T} D_{\mathrm{KL}}(\pi_{t-1}\|\pi_t^\star) \le \frac{e^{R_\eta}}{c_\eta}\left(D_{\mathrm{KL}}(\pi^\star\|\pi_0) + E_T(\pi^\star)\right), \tag{55}$$

where $c_\eta = 1 - \eta\beta - \eta/(4\beta)$ *and* $R_\eta = \eta(2V+1)$.

Proof. By Lemma B.16 applied with $\pi = \pi_{t-1}$ and $\pi^+ = \pi_t^\star$,

$$c_\eta D_{\mathrm{KL}}(\pi_t^\star\|\pi_{t-1}) \le (1-\eta\beta) D_{\mathrm{KL}}(\pi^\star\|\pi_{t-1}) - D_{\mathrm{KL}}(\pi^\star\|\pi_t^\star).$$

Dropping the nonpositive term $-\eta\beta D_{\mathrm{KL}}(\pi^\star\|\pi_{t-1})$ and adding then subtracting $D_{\mathrm{KL}}(\pi^\star\|\pi_t)$ gives

$$c_\eta D_{\mathrm{KL}}(\pi_t^\star\|\pi_{t-1}) \le D_{\mathrm{KL}}(\pi^\star\|\pi_{t-1}) - D_{\mathrm{KL}}(\pi^\star\|\pi_t) + D_{\mathrm{KL}}(\pi^\star\|\pi_t) - D_{\mathrm{KL}}(\pi^\star\|\pi_t^\star).$$

Sum over $t = 1, \dots, T$ and telescope:

$$c_\eta \sum_{t=1}^{T} D_{\mathrm{KL}}(\pi_t^\star\|\pi_{t-1}) \le D_{\mathrm{KL}}(\pi^\star\|\pi_0) + E_T(\pi^\star).$$

Finally apply Lemma B.17 to each term to obtain the desired result. □

We now turn to concentration. The following martingale Hoeffding bound handles adaptively chosen but conditionally bounded samples.

Lemma B.22 *Let* $(\mathcal{F}_i)_{i=0}^N$ *be a filtration and let* $(X_i)_{i=1}^N$ *be adapted. Suppose deterministic constants* $a_i \le b_i$ *satisfy* $a_i \le X_i \le b_i$ *almost surely. Then, for every* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\left|\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]\right| \le \sqrt{\frac{1}{2}\left(\sum_{i=1}^{N}(b_i - a_i)^2\right)\log\left(\frac{2}{\delta}\right)}.$$

Proof. Let $D_i := X_i - \mathbb{E}[X_i|\mathcal{F}_{i-1}]$ and $S_n := \sum_{i=1}^{n} D_i$. Then (S_n) is a martingale. Conditional on $\mathcal{F}_{i-1}$, D_i has mean zero and lies in an interval of length at most $c_i := b_i - a_i$. Hoeffding's lemma applied conditionally gives

$$\mathbb{E}[e^{\theta D_i} \mid \mathcal{F}_{i-1}] \le \exp\left(\frac{\theta^2 c_i^2}{8}\right), \qquad \theta \in \mathbb{R}.$$

Thus $M_n := \exp(\theta S_n - \theta^2 \sum_{i=1}^n c_i^2/8)$ is a nonnegative supermartingale. For $\theta > 0$ and $\varepsilon > 0$, Markov's inequality gives

$$P(S_N \geq \varepsilon) \leq \exp\left(-\theta\varepsilon + \frac{\theta^2}{8}\sum_{i=1}^N c_i^2\right).$$

Optimizing over θ gives

$$P(S_N \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_i c_i^2}\right).$$

Applying the same bound to $-S_N$ and union bounding gives

$$P(|S_N| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_i c_i^2}\right).$$

The claim follows by taking $\varepsilon = \sqrt{(\sum_i c_i^2)\log(2/\delta)/2}$. □

This corollary records the conditional i.i.d. form used for the per-round batches in Algorithm 3.

Corollary B.23 *Let $\mathcal{G}$ be a sigma-field. Conditional on $\mathcal{G}$, let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a, b]$ almost surely and $\mathbb{E}[X_i \mid \mathcal{G}] = \mu$ for every i. Then, for every $\epsilon > 0$,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \epsilon \,\middle|\, \mathcal{G}\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Equivalently, for every $\alpha \in (0, 1)$, with conditional probability at least $1 - \alpha$,

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq (b-a)\sqrt{\frac{\log(2/\alpha)}{2n}}.$$

Proof. Apply Lemma B.22 to the filtration $\mathcal{F}_i := \mathcal{G} \vee \sigma(X_1, \ldots, X_i)$, $i = 0, \ldots, n$. Since the variables are conditionally independent given $\mathcal{G}$ and have conditional mean μ, we have $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = \mathbb{E}[X_i \mid \mathcal{G}] = \mu$. Moreover, the range length is $b - a$ for every i. Lemma B.22 gives the displayed high-probability inequality conditionally on $\mathcal{G}$; rewriting it in tail form gives the exponential bound. □

Here, we restate the batch sample sizes defined in Theorem B.10.

For $T \geq 1$ and $\delta \in (0, 1)$ define the policy-class and active-set logarithmic factors as follows:

$$m_t := t, \qquad n_t := t, \qquad k_t := t, \qquad t = 1, \ldots, T. \tag{56}$$

The logarithmic factors are moved into the ridge parameter and the final regret bound. Set

$$L_\Pi := \log\left(\frac{16T|\Pi|^2}{\delta}\right). \tag{57}$$

Because $n_r = r$, the total number of active atoms across all rounds is bounded by

$$M_T := 2\sum_{r=1}^T (T - r + 1)r = \frac{T(T+1)(T+2)}{3}. \tag{58}$$

Define

$$L_P := \log\left(\frac{16M_T}{\delta}\right), \qquad L_T := \max\{1, L_\Pi, L_P, \log(eT)\}. \tag{59}$$

Then $\sum_{t=1}^T |\mathcal{Y}_t^{\mathrm{act}}| \leq M_T$. Define

$$\overline{H} := 1 + B^2 + \beta^{-2}, \qquad \lambda := 16\overline{H}L_T^2, \qquad S_T \geq 1 \vee \mathrm{SEC}_{\mathrm{NLHF}}(\Psi, \lambda), \qquad \gamma := \sqrt{\frac{S_T}{\lambda T}}. \tag{60}$$

Use the learner coefficient

$$\alpha_t = \frac{2}{\gamma m_t} = \frac{2}{\gamma t}. \tag{61}$$

Finally, fix the constant stepsize

$$\eta := \frac{2\beta}{1+4\beta^2}. \tag{62}$$

Then $\eta \le 4\beta/(1+4\beta^2)$, $\eta\beta < 1$, $\eta \le 1$, and

$$c_\eta = 1 - \eta\beta - \frac{\eta}{4\beta} = \frac{1}{2}. \tag{63}$$

The final lemma verifies all deterministic events required above hold with high probability. It controls the explorer numerator, the explorer denominator, and the active-set preference estimates simultaneously.

Lemma B.24 *Under the sampling rules of Algorithm 3 and the choices Eqs.* (56) *and* (57)*, with probability at least $1-\delta$, all of the following statements hold simultaneously for all $t \le T$.*

(i) For all $\rho', \pi \in \Pi$ and all $\nu \in \Pi$,

$$|N_t^{\rho',\pi}(\nu) - \widehat{N}_t^{\rho',\pi}(\nu)| \le \epsilon_{m,t}, \qquad \epsilon_{m,t} := B\sqrt{\frac{2L_\Pi}{m_t}}. \tag{64}$$

Although the event is stated for all $\nu \in \Pi$, the random deviation depends only on the on-policy empirical average, so no union bound over ν is needed. In particular, Eq. (44) *holds with the same $\epsilon_{m,t}$.*

(ii) For all $\rho', \pi \in \Pi$,

$$|\widehat{H}_t^{\rho',\pi} - H_t^{\rho',\pi}| \le \frac{\lambda}{2}. \tag{65}$$

Consequently, for all $\pi \in \Pi$,

$$\frac{1}{2}\big(\lambda + H_t(\pi)\big) \le \lambda + \widehat{H}_t(\pi) \le 2\big(\lambda + H_t(\pi)\big), \tag{66}$$

and Eq. (45) *holds.*

(iii) For every active prompt-response atom in the empirical historical loss at round t,

$$|\widehat{P}_t(a \succ \pi_{t-1}) - P^\star(a \succ \pi_{t-1})| \le \epsilon_{p,t}, \qquad \epsilon_{p,t}^2 := \frac{L_P}{2k_t}. \tag{67}$$

Proof. We prove the three claims and then union bound.

(i) Fix t, $\rho', \pi \in \Pi$, and $\nu \in \Pi$. Let $\mathcal{G}_t^m$ be the sigma-field containing all randomness before drawing the on-policy batch $\widetilde{\tau}_{t-1,1:m_t}$. Conditional on $\mathcal{G}_t^m$, the variables

$$X_i := \log \frac{\rho'(\widetilde{\tau}_{t-1,i})}{\pi(\widetilde{\tau}_{t-1,i})}, \qquad i = 1, \dots, m_t,$$

are independent because $\widetilde{\tau}_{t-1,i}$ are sampled independently from π_{t-1}. Assumption B.5 gives $X_i \in [-B, B]$ almost surely. Moreover,

$$\mathbb{E}[X_i \mid \mathcal{G}_t^m] = \mathbb{E}_{\tau\sim\pi_{t-1}}\left[\log \frac{\rho'(\tau)}{\pi(\tau)}\right].$$

The ν-dependent term $\mathbb{E}_{\tau'\sim\nu}[\log(\rho'(\tau')/\pi(\tau'))]$ is deterministic conditional on $\mathcal{G}_t^m$ and appears in both $N_t^{\rho',\pi}(\nu)$ and $\widehat{N}_t^{\rho',\pi}(\nu)$, so it cancels. Hence

$$N_t^{\rho',\pi}(\nu) - \widehat{N}_t^{\rho',\pi}(\nu) = \mathbb{E}[X_1 \mid \mathcal{G}_t^m] - \frac{1}{m_t}\sum_{i=1}^{m_t} X_i.$$

Apply Corollary B.23 with $a = -B$, $b = B$, $n = m_t$, and $\epsilon = B\sqrt{2L_\Pi/m_t}$. Since $b-a = 2B$, this gives

$$P\left(|N_t^{\rho',\pi}(\nu) - \widehat{N}_t^{\rho',\pi}(\nu)| > \epsilon \;\middle|\; \mathcal{G}_t^m\right) \le 2e^{-L_\Pi}.$$

This bound is uniform in ν because the random deviation does not depend on ν. Therefore no union bound over ν is needed. Union bounding over $t \le T$ and $(\rho', \pi) \in \Pi^2$ gives total failure probability at most $2T|\Pi|^2 e^{-L_\Pi} = \delta/8$. The local event Eq. (44) follows by substituting $(\rho', \pi) = (\pi_t^\star, \pi_t)$; this is legitimate because Assumption B.4 implies $\pi_t^\star \in \Pi$ and Algorithm 3 outputs $\pi_t \in \Pi$.

(ii) Fix t and $(\rho', \pi) \in \Pi^2$. For $s = 0, \dots, t-1$ and $i = 1, \dots, n_{s+1}$ define

$$Z_{s,i} := \frac{1}{n_{s+1}} \psi_{\rho',\pi}(\tau_{s,i}, \tau'_{s,i})^2.$$

Order the pairs (s, i) lexicographically and let $\mathcal{F}_{s,i}$ be the filtration generated by all samples revealed up to pair (s, i). Conditional on the history before drawing the s-th exploration batch, the pairs $(\tau_{s,i}, \tau'_{s,i})$ are independent with law $\pi_{s-1} \otimes \widehat{\pi}_s$. Hence $Z_{s,i}$ is adapted and

$$\mathbb{E}[Z_{s,i} \mid \mathcal{F}_{s,i-1}] = \frac{1}{n_{s+1}} \mathbb{E}_{\tau \sim \pi_{s-1}, \tau' \sim \widehat{\pi}_s}\left[\psi_{\rho',\pi}(\tau, \tau')^2\right].$$

By Assumption B.5, $|\log(\rho'(\tau)/\pi(\tau))| \le B$, so $|\psi_{\rho',\pi}(\tau, \tau')| \le 2B$. Therefore $0 \le Z_{s,i} \le 4B^2/n_{s+1}$, and the range length in Lemma B.22 is $4B^2/n_{s+1}$. Applying Lemma B.22 to the ordered collection $\{Z_{s,i}\}_{s<t, i \le n_{s+1}}$ with confidence parameter $\alpha = 2e^{-L_\Pi}$ gives, for this fixed t, ρ', π,

$$|\widehat{H}_t^{\rho',\pi} - H_t^{\rho',\pi}| \le \sqrt{\frac{1}{2}\sum_{s=0}^{t-1} n_{s+1}\left(\frac{4B^2}{n_{s+1}}\right)^2 L_\Pi} = 4B^2\sqrt{\frac{1}{2}\left(\sum_{s=0}^{t-1}\frac{1}{n_{s+1}}\right) L_\Pi}$$

with failure probability at most $2e^{-L_\Pi}$. Union bounding over $t \le T$ and $(\rho', \pi) \in \Pi^2$ costs at most another $2T|\Pi|^2 e^{-L_\Pi} = \delta/8$. Under the simple schedule $n_{s+1} = s + 1$,

$$\sum_{s=0}^{t-1}\frac{1}{n_{s+1}} \le \sum_{r=1}^{T}\frac{1}{r} \le \log(eT) \le L_T, \qquad L_\Pi \le L_T.$$

Thus

$$|\widehat{H}_t^{\rho',\pi} - H_t^{\rho',\pi}| \le 4B^2\frac{L_T}{\sqrt{2}} \le 4\overline{H}L_T \le 8\overline{H}L_T^2 = \frac{\lambda}{2},$$

where $B^2 \le \overline{H}$, $L_T \ge 1$, and $\lambda = 16\overline{H}L_T^2$. The comparison Eq. (66) follows by applying this bound to $(\rho', \pi) = (\pi_t, \pi)$ and using $|\widehat{H} - H| \le \lambda/2$.

(iii) Fix t and condition on the history $\mathcal{G}_t^p$ just before drawing the comparison opponents for preference estimation. For a fixed active prompt-response atom $a = (\mathbf{x}, \mathbf{y}) \in \mathcal{Y}_t^{\mathrm{act}}$, Algorithm 3 draws $c_{t,a,j} \sim \pi_{t-1}(\cdot \mid \mathbf{x})$ independently and then observes

$$Y_{t,a,j} \sim \mathrm{Bern}\big(P^\star(\mathbf{y} \succ c_{t,a,j} \mid \mathbf{x})\big), \qquad j = 1, \dots, k_t.$$

The variables $Y_{t,a,j}$ are conditionally independent and lie in $[0, 1]$. By the tower property,

$$\mathbb{E}[Y_{t,a,j} \mid \mathcal{G}_t^p] = \mathbb{E}_{c \sim \pi_{t-1}(\cdot|\mathbf{x})}[P^\star(\mathbf{y} \succ c \mid \mathbf{x})] = P^\star(a \succ \pi_{t-1}).$$

Apply Corollary B.23 with $a = 0$, $b = 1$, $n = k_t$, and $\epsilon = \sqrt{L_P/(2k_t)}$. For this fixed active atom, the conditional failure probability is at most $2e^{-L_P}$. Since the active sets are history-measurable before the labels at round t are drawn and $\sum_{t=1}^{T} |\mathcal{Y}_t^{\mathrm{act}}| \le M_T$, union bounding over all realized active atoms and all times gives failure probability at most $2M_T e^{-L_P} = \delta/8$.

Combining the three bounds gives total failure probability at most $3\delta/8 < \delta$. □

Algorithm 4: Bonus-explorer Exploratory Nash Preference Optimization (BENPO)

Input: horizon T, reference policy π_{ref}, confidence-radius schedule $\{\gamma_t(\delta)\}_{t\geq 1}$, coefficient $\lambda > 0$;

Initialize $D^0 \leftarrow \varnothing$;

for $t = 1, 2, \ldots, T$ **do**

Compute MLE $\widehat{P}_t$ via Eq. (68);

Compute a Nash policy under $\widehat{P}_t$ by solving

$$\pi_t \in \arg\max_{\pi_1} \min_{\pi_2} J_\beta\big(\pi_1, \pi_2; \widehat{P}_t\big);$$

Generate a prompt-response pair $(\mathbf{x}_t, \mathbf{y}_t)$ via $\mathbf{x}_t \sim \rho$ and $\mathbf{y}_t \sim \pi_t(\cdot \mid \mathbf{x}_t)$;

Construct UCB bonus at $(\mathbf{x}_t, \mathbf{y}_t)$ for any $\mathbf{y}' \in \mathcal{Y}_{\mathbf{x}_t}$: $\mathfrak{b}_t(\mathbf{y}'; \mathbf{x}_t, \mathbf{y}_t)$ by

$$\min\left\{1, \sqrt{\gamma_t(\delta)} \sup_{P,P'\in\mathcal{P}} \frac{\big|P(\mathbf{y}_t \succ \mathbf{y}'|\mathbf{x}_t) - P'(\mathbf{y}_t \succ \mathbf{y}'|\mathbf{x}_t)\big|}{\sqrt{\lambda + \sum_{s=1}^{t-1}\big(P(\mathbf{y}_s \succ \hat{\mathbf{y}}_s|\mathbf{x}_s) - P'(\mathbf{y}_s \succ \hat{\mathbf{y}}_s|\mathbf{x}_s)\big)^2}}\right\};$$

Sample from the exploiter policy:

$$\hat{\mathbf{y}}_t \in \arg\max_{\mathbf{y}'\in\mathcal{Y}_{\mathbf{x}_t}} \mathfrak{b}_t(\mathbf{y}'; \mathbf{x}_t, \mathbf{y}_t);$$

Query preference label $I_t \sim \mathrm{Bern}(P^\star(\mathbf{y}_t \succ \hat{\mathbf{y}}_t|\mathbf{x}_t))$;

Update dataset $D^t \leftarrow D^{t-1} \cup \{(\mathbf{x}_t, \mathbf{y}_t, \hat{\mathbf{y}}_t, I_t)\}$;

Output: $\{\pi_t\}_{t=1}^T$;

B.4 Proofs of improved sample efficiency of BENPO

B.4.1 Algorithm

We first provide the high-level idea of Bonus-Explorer ENPO. At round t, BENPO computes the MLE $\widehat{P}_t$ of the preference function from the data collected so far:

$$\widehat{P}_t \in \arg\max_{P\in\mathcal{P}} \sum_{s=1}^{t-1}\Big[I_s \log P(\mathbf{y}_s \succ \hat{\mathbf{y}}_s|\mathbf{x}_s) + (1 - I_s)\log\big(1 - P(\mathbf{y}_s \succ \hat{\mathbf{y}}_s|\mathbf{x}_s)\big)\Big], \tag{68}$$

where $\mathbf{x}_s, \mathbf{y}_s, \hat{\mathbf{y}}_s$ are sampled prompts and responses at round s, and $I_s \sim \mathrm{Bern}(P^\star(\mathbf{y}_s \succ \hat{\mathbf{y}}_s|\mathbf{x}_s))$ is the preference label collected at round t. Then, it finds a regularized Nash policy $\hat{\pi}_t$ by querying a minimax oracle:

$$\pi_t \in \arg\max_{\pi_1} \min_{\pi_2} \widehat{P}_t(\pi_1 \succ \pi_2) - \beta\mathrm{KL}(\pi_1\|\pi_{\mathrm{ref}}) + \beta\mathrm{KL}(\pi_2\|\pi_{\mathrm{ref}}).$$

This main-policy step plays the same role as ENPO's policy update in Eq. (14)—it tracks the regret-relevant policy, the one whose suboptimality is the object of analysis. We then draw one sample from $\mathbf{y}_t \sim \pi_t$, and based on this sample, we solve the following maximization problem:

$$\max_{\mathbf{y}'\in\mathcal{Y}_{\mathbf{x}_t}} \left\{1 \wedge \sqrt{\gamma_t(\delta)} \sup_{P,P'\in\mathcal{P}} \frac{\big|P(\mathbf{y}_t \succ \mathbf{y}'|\mathbf{x}_t) - P'(\mathbf{y}_t \succ \mathbf{y}'|\mathbf{x}_t)\big|}{\sqrt{\lambda + \sum_{s=1}^{t-1}\big(P(\mathbf{y}_s \succ \hat{\mathbf{y}}_s|\mathbf{x}_s) - P'(\mathbf{y}_s \succ \hat{\mathbf{y}}_s|\mathbf{x}_s)\big)^2}}\right\}, \tag{69}$$

and denote this adversarial sample by $\hat{\mathbf{y}}_t$. We then query a preference label I_t between $\mathbf{y}_t$ and $\hat{\mathbf{y}}_t$.

See Algorithm 4 for the full statement of the algorithm.

As a counterpart of ENPO with stronger oracles, BENPO aligns with ENPO in a clear manner: *maintain a main policy $\hat{\pi}_t$ whose regret is the object of analysis, and pair that main policy with a second mechanism whose only role is to guide exploration*.

B.4.2 ASSUMPTIONS AND ADDITIONAL NOTATIONS

Before we analyze the regret of BENPO, we first introduce some assumptions and notations.

Let $\mathcal{P}$ be a finite realizable class of preference models: $P^\star \in \mathcal{P}$. Furthermore, every $P \in \mathcal{P}$ satisfies

$$P(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}) + P(\mathbf{y}' \succ \mathbf{y} \mid \mathbf{x}) = 1, \text{ thus } P(\mathbf{y} \succ \mathbf{y} \mid \mathbf{x}) = \tfrac{1}{2}.$$

We define the KL-regularized game with a t-round empirical preference as follows:

$$\max_{\pi_1} \min_{\pi_2} J_\beta\big(\pi_1, \pi_2; \widehat{P}_t\big) := \widehat{P}_t(\pi_1 \succ \pi_2) - \beta \mathrm{KL}(\pi_1 \| \pi_{\text{ref}}) + \beta \mathrm{KL}(\pi_2 \| \pi_{\text{ref}}) \tag{70}$$

Note that the maximum and the minimum are taken over all policies with the same support set as π_{ref}, which is the class over which the Nash equilibrium $\pi^\star$ is defined in Section 2.

We denote $J_\beta(\pi_1, \pi_2) := J_\beta(\pi_1, \pi_2; P^\star)$. For each $P \in \mathcal{P}$, let $\pi_P^\star$ denote the unique symmetric Nash policy of the KL-regularized game under P, i.e.,

$$J_\beta(\pi_1, \pi_P^\star; P) \le J_\beta(\pi_P^\star, \pi_P^\star; P) \le J_\beta(\pi_P^\star, \pi_2; P), \qquad \forall\, \pi_1, \pi_2. \tag{71}$$

By anti-symmetry of the preference model P, $J_\beta(\pi_P^\star, \pi_P^\star; P) = 1/2$.

We write $\mathrm{gap}_\beta(\pi) := \max_{\pi_1} J_\beta(\pi_1, \pi) - \min_{\pi_2} J_\beta(\pi, \pi_2)$ for the duality gap over this class. By Proposition B.1, whose proof applies verbatim, $\mathrm{gap}_\beta(\pi) = 2\big(\max_{\pi_1} J_\beta(\pi_1, \pi) - \frac{1}{2}\big)$.

Throughout Appendix B.4, all maxima and minima over policies, including those defining gap_β, are taken over this class rather than the finite class of Assumption B.3. This is the natural choice for BENPO: its goal is to learn this Nash equilibrium, and at each round it solves the same KL-regularized minimax problem with $P^\star$ replaced by the estimate $\hat{P}_t$, so that $\pi_t := \pi^\star_{\hat{P}_t}$ is a plug-in estimate of $\pi^\star$.

We specify that the statistical domain is

$$\mathcal{Z} := \{(\mathbf{x}, \mathbf{y}, \mathbf{y}') : \mathbf{x} \in \mathcal{X},\ \mathbf{y}, \mathbf{y}' \in \mathcal{Y}_x\}.$$

Following Zhao et al. (2025) (see also Ye et al. (2024)), we measure the complexity of $\mathcal{P}$ by the eluder coefficient defined as follows:

$$d_{\mathcal{P}} := \sup_{z_1, \ldots, z_T \in \mathcal{Z}} \sum_{t=1}^{T} \min\left\{1, \sup_{P, P' \in \mathcal{P}} \frac{\big(P(\mathbf{y}_t \succ \mathbf{y}_t' | \mathbf{x}_t) - P'(\mathbf{y}_t \succ \mathbf{y}_t' | \mathbf{x}_t)\big)^2}{\lambda + \sum_{s<t} \big(P(\mathbf{y}_s \succ \mathbf{y}_s' | \mathbf{x}_s) - P'(\mathbf{y}_s \succ \mathbf{y}_s' | \mathbf{x}_s)\big)^2}\right\},$$

where $z_t := (\mathbf{x}_t, \mathbf{y}_t, \mathbf{y}_t')$ for all $1 \le t \le T$.

B.4.3 MAIN PROOF OF SAMPLE EFFICIENCY

The following squared ℓ^2 bound for the regret bound or equivalently duality gap, is the key to access the logarithmic regret bound. The proof strategy follows that in the proof of (Lee et al., 2026, Theorem 3.1.).

Lemma B.25 (Squared ℓ^2 bound) *Let $\widehat{P}$ be any antisymmetric preference function, and let $\pi^\star_{\widehat{P}}$ denote the symmetric regularized Nash policy under $\widehat{P}$:*

$$\pi^\star_{\widehat{P}} \in \arg\max_{\pi_1} \min_{\pi_2} J_\beta(\pi_1, \pi_2; \widehat{P}).$$

Then, we have

$$\mathrm{gap}_\beta(\pi^\star_{\widehat{P}}) \le \frac{4}{\beta} \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y} \sim \pi^\star_{\widehat{P}}(\cdot|\mathbf{x})} \left[\sup_{\mathbf{y}' \in \mathcal{Y}} \big(P^\star(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}) - \widehat{P}(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x})\big)^2 \right].$$

Proof. Denote

$$\widetilde{\pi} := \arg\max_{\pi_1} J_\beta(\pi_1, \pi^\star_{\widehat{P}}; P^\star).$$

First, note that

$$\begin{aligned}\frac{1}{2}\mathrm{gap}_\beta(\pi^\star_{\widehat{P}}) &= \max_{\pi_1} J_\beta(\pi_1, \pi^\star_{\widehat{P}}; P^\star) - \frac{1}{2} \\ &= J_\beta(\widetilde{\pi}, \pi^\star_{\widehat{P}}; P^\star) - J_\beta(\widetilde{\pi}, \pi^\star_{\widehat{P}}; \widehat{P}) + J_\beta(\widetilde{\pi}, \pi^\star_{\widehat{P}}; \widehat{P}) - \frac{1}{2} \\ &\le J_\beta(\widetilde{\pi}, \pi^\star_{\widehat{P}}; P^\star) - J_\beta(\widetilde{\pi}, \pi^\star_{\widehat{P}}; \widehat{P}) \\ &= \mathbb{E}_{\mathbf{x}\sim\rho}\, \mathbb{E}_{\mathbf{y}\sim\widetilde{\pi}(\cdot|\mathbf{x}),\, \mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\big[P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big],\end{aligned}$$

where the inequality follows by the fact that $\pi^\star_{\widehat{P}}$ is a Nash policy with preference $\widehat{P}$.

Since, by the antisymmetry of $P^\star$ and $\widehat{P}$, for any π and every $\mathbf{x}$, $\mathbb{E}_{\mathbf{y},\mathbf{y}'\sim\pi(\cdot|\mathbf{x})}\big[P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big] = 0$, we have

$$\begin{aligned}&\frac{1}{2}\mathrm{gap}_\beta(\pi^\star_{\widehat{P}}) \\ \le&\mathbb{E}_{\mathbf{x}\sim\rho}\Big[\mathbb{E}_{\mathbf{y}\sim\widetilde{\pi}(\cdot|\mathbf{x}),\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\big[P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big] \\ &\qquad - \mathbb{E}_{\mathbf{y},\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\big[P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big]\Big] \\ =\ &\mathbb{E}_{\mathbf{x}\sim\rho}\, \mathbb{E}_{\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\Big[\big\langle P^\star(\cdot\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\cdot\succ\mathbf{y}'|\mathbf{x}),\ \widetilde{\pi}(\cdot|\mathbf{x}) - \pi^\star_{\widehat{P}}(\cdot|\mathbf{x})\big\rangle\Big] \\ \le\ &\mathbb{E}_{\mathbf{x}\sim\rho}\Big[\mathbb{E}_{\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\Big[\sup_{\mathbf{y}\in\mathcal{Y}}\big|P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big|\Big]\cdot\big\|\widetilde{\pi}(\cdot|\mathbf{x}) - \pi^\star_{\widehat{P}}(\cdot|\mathbf{x})\big\|_1\Big] \\ \le\ &\Big(\mathbb{E}_{\mathbf{x}\sim\rho}\Big[\Big(\mathbb{E}_{\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\sup_{\mathbf{y}\in\mathcal{Y}}\big|P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big|\Big)^2\Big]\Big)^{1/2}\Big(\mathbb{E}_{\mathbf{x}\sim\rho}\big\|\widetilde{\pi}(\cdot|\mathbf{x}) - \pi^\star_{\widehat{P}}(\cdot|\mathbf{x})\big\|_1^2\Big)^{1/2},\end{aligned} \tag{72}$$

where the last two inequalities follow from Hölder's inequality for each $\mathbf{x}$ and the Cauchy–Schwarz inequality over $\mathbf{x}\sim\rho$.

On the other hand, since $-\beta\,\mathrm{KL}(\pi_1\|\pi_{\mathrm{ref}}) = -\beta\,\mathbb{E}_{\mathbf{x}\sim\rho}\mathrm{KL}(\pi_1(\cdot|\mathbf{x})\|\pi_{\mathrm{ref}}(\cdot|\mathbf{x}))$ is β-strongly concave with respect to $\|\pi\| := (\mathbb{E}_{\mathbf{x}\sim\rho}\|\pi(\cdot|\mathbf{x})\|_1^2)^{1/2}$ (by Pinsker's inequality, prompt by prompt), and $\tilde{\pi}$ maximizes the concave map $\pi_1 \mapsto J_\beta(\pi_1, \pi^\star_{\widehat{P}}; P^\star)$ over this (convex) policy class, so that $J_\beta(\pi_1, \pi^\star_{\widehat{P}}; P^\star)$ is β-strongly concave in π_1 and the first-order optimality condition holds at $\tilde{\pi}$. Also, we have $J_\beta(\pi^\star_{\widehat{P}}, \pi^\star_{\widehat{P}}; P^\star) = \frac{1}{2}$, therefore

$$\frac{1}{2}\mathrm{gap}_\beta(\pi^\star_{\widehat{P}}) = J_\beta(\widetilde{\pi}, \pi^\star_{\widehat{P}}; P^\star) - J_\beta(\pi^\star_{\widehat{P}}, \pi^\star_{\widehat{P}}; P^\star) \ge \frac{\beta}{2}\big\|\widetilde{\pi} - \pi^\star_{\widehat{P}}\big\|^2. \tag{73}$$

By rearranging Eq. (73), we have

$$\Big(\mathbb{E}_{\mathbf{x}\sim\rho}\Big[\big\|\widetilde{\pi}(\cdot|\mathbf{x}) - \pi^\star_{\widehat{P}}(\cdot|\mathbf{x})\big\|_1^2\Big]\Big)^{1/2} \le \sqrt{\mathrm{gap}_\beta(\pi^\star_{\widehat{P}})/\beta}. \tag{74}$$

Substituting Eq. (74) into Eq. (72) yields the quadratic inequality

$$\frac{1}{2}\mathrm{gap}_\beta(\pi^\star_{\widehat{P}}) \le \Big(\mathbb{E}_{\mathbf{x}\sim\rho}\Big[\Big(\mathbb{E}_{\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\sup_{\mathbf{y}\in\mathcal{Y}}\big|P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big|\Big)^2\Big]\Big)^{1/2}\sqrt{\mathrm{gap}_\beta(\pi^\star_{\widehat{P}})/\beta},$$

which solves to

$$\mathrm{gap}_\beta(\pi^\star_{\widehat{P}}) \le \frac{4}{\beta}\,\mathbb{E}_{\mathbf{x}\sim\rho}\Big[\Big(\mathbb{E}_{\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\sup_{\mathbf{y}\in\mathcal{Y}}\big|P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big|\Big)^2\Big].$$

By Jensen's inequality, we obtain

$$\mathrm{gap}_\beta(\pi^\star_{\widehat{P}}) \le \frac{4}{\beta}\,\mathbb{E}_{\mathbf{x}\sim\rho}\,\mathbb{E}_{\mathbf{y}'\sim\pi^\star_{\widehat{P}}(\cdot|\mathbf{x})}\Big[\sup_{\mathbf{y}\in\mathcal{Y}}\big(P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big)^2\Big]. \tag{75}$$

Note that, by antisymmetric property of the preferences, we further have

$$\big(P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big)^2 = \big(P^\star(\mathbf{y}'\succ\mathbf{y}|\mathbf{x}) - \widehat{P}(\mathbf{y}'\succ\mathbf{y}|\mathbf{x})\big)^2,$$

By swapping the names of $\mathbf{y}$ and $\mathbf{y}'$ in Eq. (75), we complete the proof. □

Theorem B.26 (Logarithmic regret bound of Algorithm 4) *Suppose $\mathcal{P}$ is a finite class of antisymmetric preference functions with $P^\star \in \mathcal{P}$. In Algorithm 4, set $\lambda \in (0, 1]$ and*

$$\gamma_t(\delta) = 4\log(2T|\mathcal{P}|/\delta) + \lambda, \quad \forall\, 1 \le t \le T.$$

Then, the iterates $\{\pi_t\}_{t=1}^T$ produced by Algorithm 4 satisfy

$$\sum_{t=1}^{T} \mathrm{gap}_\beta(\pi_t) \le \widetilde{O}\Big(\frac{d_{\mathcal{P}}\ \log(T|\mathcal{P}|/\delta)\ \log T}{\beta}\Big)$$

with probability at least $1-\delta$.

Proof of Theorem B.26. We begin by reducing the per-round suboptimality to a *squared* preference error via Lemma B.25. Since $\pi_t = \pi^\star_{\widehat{P}_t}$ by construction, by Lemma B.25 we have

$$\mathrm{gap}_\beta(\pi_t) \le \frac{4}{\beta}\mathbb{E}_{\mathbf{x}\sim\rho,\,\mathbf{y}\sim\pi_t(\cdot|\mathbf{x})}\left[\sup_{\mathbf{y}'\in\mathcal{Y}}\big(P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}_t(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})\big)^2\right]. \tag{76}$$

Throughout this proof, we abbreviate $\Delta_t(\mathbf{y},\mathbf{y}'|\mathbf{x}) := P^\star(\mathbf{y}\succ\mathbf{y}'|\mathbf{x}) - \widehat{P}_t(\mathbf{y}\succ\mathbf{y}'|\mathbf{x})$, and denote $\mathcal{F}_{t-1} := \sigma\big(\{\mathbf{x}_s,\mathbf{y}_s,\hat{\mathbf{y}}_s,I_s\}_{s<t}\big)$ and $\mathbb{E}_t[\,\cdot\,] := \mathbb{E}[\,\cdot \mid \mathcal{F}_{t-1}]$.

By Lemma B.27, there exists an event $\mathcal{E}_1$ with $\mathbb{P}(\mathcal{E}_1) \ge 1-\delta/2$ on which

$$\sum_{s=1}^{t-1}\Big(P^\star(\mathbf{y}_s\succ\hat{\mathbf{y}}_s|\mathbf{x}_s) - \widehat{P}_t(\mathbf{y}_s\succ\hat{\mathbf{y}}_s|\mathbf{x}_s)\Big)^2 \le 4\log(2T|\mathcal{P}|/\delta) \qquad \forall\, t\in[T].$$

Because $\gamma_t(\delta) := 4\log(2T|\mathcal{P}|/\delta)+\lambda$, taking $P = P^\star$, $P' = \widehat{P}_t$ inside the supremum that defines $\mathfrak{b}_t$, on the event $\mathcal{E}_1$, we have $\lambda + \sum_{s<t}\Delta_t(\mathbf{y}_s,\hat{\mathbf{y}}_s|\mathbf{x}_s)^2 \le \gamma_t(\delta)$. Note that $|\Delta_t(\mathbf{y}_s,\mathbf{y}'_s|\mathbf{x}_s)| \le 1$ holds trivially. Then,

$$\Delta_t(\mathbf{y},\mathbf{y}'|\mathbf{x})^2 = \min\big\{\Delta_t(\mathbf{y},\mathbf{y}'|\mathbf{x})^2, 1\big\} \le \min\left\{\frac{\gamma_t(\delta)\Delta_t(\mathbf{y},\mathbf{y}'|\mathbf{x})^2}{\lambda+\sum_{s=1}^{t-1}\Delta_t(\mathbf{y}_s,\hat{\mathbf{y}}_s|\mathbf{x}_s)^2}, 1\right\} \le \mathfrak{b}_t(\mathbf{y}';\mathbf{x},\mathbf{y})^2,$$

therefore $|\Delta_t(\mathbf{y},\mathbf{y}'|\mathbf{x})| \le \mathfrak{b}_t(\mathbf{y}';\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y},\mathbf{y}') \in \mathcal{X}\times\mathcal{Y}^2$.

Squaring, taking $\sup_{\mathbf{y}'\in\mathcal{Y}}$, and noting that the exploiter $\hat{\mathbf{y}}_t = \arg\max_{\mathbf{y}'\in\mathcal{Y}_{\mathbf{x}_t}}\mathfrak{b}_t(\mathbf{y}';\mathbf{x}_t,\mathbf{y}_t)$ realizes the supremum, hence we have

$$\sup_{\mathbf{y}'\in\mathcal{Y}_{\mathbf{x}_t}}\Delta_t(\mathbf{y}_t,\mathbf{y}'|\mathbf{x}_t)^2 \le \mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2.$$

Substituting into Eq. (76) yields

$$\mathrm{gap}_\beta(\pi_t) \le \frac{4}{\beta}\mathbb{E}_t\big[\mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2\big]. \tag{77}$$

We next concentrate the squared bonus around its conditional expectation. Since $\mathfrak{b}_t \in [0,1]$ by construction, the sequence

$$X_t \;:=\; \mathbb{E}_t\big[\mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2\big] - \mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2$$

is a martingale difference sequence with $|X_t| \le 1$ and $\mathbb{E}_t[X_t^2] \le \mathbb{E}_t\big[\mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2\big]$. Applying Lemma B.28 with $Y_t = \mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)$ and confidence $\delta/2$ yields an event $\mathcal{E}_2$ with $\mathbb{P}(\mathcal{E}_2) \ge 1-\delta/2$ on which

$$\sum_{t=1}^{T}\mathbb{E}_t\big[\mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2\big] \le 2\sum_{t=1}^{T}\mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2 + 4\log(2/\delta). \tag{78}$$

To control the cumulative realized squared bonus, observe that by the definition of $\mathfrak{b}_t$ and $\gamma_t(\delta) \equiv \gamma_T(\delta) = 4\log(2T|\mathcal{P}|/\delta)+1 \ge 1$,

$$\mathfrak{b}_t(\hat{\mathbf{y}}_t;\mathbf{x}_t,\mathbf{y}_t)^2 \le \gamma_T(\delta)\cdot\min\left\{1,\ \sup_{P,P'\in\mathcal{P}}\frac{(P-P')^2(\mathbf{y}_t\succ\hat{\mathbf{y}}_t|\mathbf{x}_t)}{\lambda+\sum_{s=1}^{t-1}(P-P')^2(\mathbf{y}_s\succ\hat{\mathbf{y}}_s|\mathbf{x}_s)}\right\}.$$

By the definition of $d_{\mathcal{P}}$, we have

$$\sum_{t=1}^{T} \mathfrak{b}_t(\hat{\mathbf{y}}_t; \mathbf{x}_t, \mathbf{y}_t)^2 \le \gamma_T(\delta)\, d_{\mathcal{P}}. \tag{79}$$

Combining the pieces, on $\mathcal{E}_1 \cap \mathcal{E}_2$ (which has probability at least $1-\delta$ by a union bound), summing Eq. (77) over $t \in [T]$ and chaining Eqs. (78) and (79),

$$\begin{aligned}
\sum_{t=1}^{T} \mathrm{gap}_\beta(\pi_t) &\le \frac{4}{\beta} \sum_{t=1}^{T} \mathbb{E}_t\big[\mathfrak{b}_t(\widehat{\mathbf{y}}_t; \mathbf{x}_t, \mathbf{y}_t)^2\big] \\
&\le \frac{8}{\beta} \sum_{t=1}^{T} \mathfrak{b}_t(\widehat{\mathbf{y}}_t; \mathbf{x}_t, \mathbf{y}_t)^2 + \frac{16 \log(2/\delta)}{\beta} \\
&\le \frac{8}{\beta} \cdot \gamma_T(\delta) \cdot d_{\mathcal{P}} + \frac{16 \log(2/\delta)}{\beta} \\
&= \widetilde{O}\left(\frac{d_{\mathcal{P}} \cdot \log(|\mathcal{P}|/\delta) \cdot \log T}{\beta} \right).
\end{aligned}$$

□

B.4.4 Technical lemmas

Lemma B.27 (MLE concentration) *Let $\mathcal{P}$ be a finite class of preference functions $P : \mathcal{V}^\star \times \mathcal{V}^\star \times \mathcal{V}^\star \to [0,1]$ with $P^\star \in \mathcal{P}$. Suppose that at each round $s \in [T]$ the preference label is generated as $I_s \sim \mathrm{Bern}\big(P^\star(\mathbf{y}_s \succ \widehat{\mathbf{y}}_s \mid \mathbf{x}_s)\big)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{s=1}^{t-1} \big(\widehat{P}_t(\mathbf{y}_s \succ \widehat{\mathbf{y}}_s \mid \mathbf{x}_s) - P^\star(\mathbf{y}_s \succ \widehat{\mathbf{y}}_s \mid \mathbf{x}_s)\big)^2 \le 4 \log\big(T|\mathcal{P}|/\delta\big) \qquad \forall\, t \in [T].$$

Proof. Fix $t \in [T]$ and a candidate $P \in \mathcal{P}$. Write $p_s(P) := P(\mathbf{y}_s \succ \widehat{\mathbf{y}}_s \mid \mathbf{x}_s)$ for short, and let $\mathcal{F}_s := \sigma\big(\{\mathbf{x}_r, \mathbf{y}_r, \widehat{\mathbf{y}}_r, I_r\}_{r \le s}\big)$ be the natural filtration and $\mathcal{G}_s := \sigma\big(\mathcal{F}_{s-1}, \mathbf{x}_s, \mathbf{y}_s, \widehat{\mathbf{y}}_s\big)$, which contains the round-s query but not its label I_s. Define the half-log-likelihood-ratio process

$$M_t(P) := \prod_{s=1}^{t-1} \sqrt{\frac{p_s(P)^{I_s}(1-p_s(P))^{1-I_s}}{p_s(P^\star)^{I_s}(1-p_s(P^\star))^{1-I_s}}}.$$

A direct computation gives

$$\begin{aligned}
\mathbb{E}\big[M_t(P) \,\big|\, \mathcal{G}_{t-1}\big] &= M_{t-1}(P) \cdot \Big(\sqrt{p_{t-1}(P) p_{t-1}(P^\star)} + \sqrt{(1-p_{t-1}(P))(1-p_{t-1}(P^\star))} \Big) \\
&= M_{t-1}(P) \cdot \big(1 - \tfrac{1}{2} D_H^2(p_{t-1}(P), p_{t-1}(P^\star))\big),
\end{aligned}$$

where $D_H^2(p,q) := (\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2$ is the squared Hellinger distance between Bernoulli(p) and Bernoulli(q). Since $1 - u \le e^{-u}$, the process

$$W_t(P) := M_t(P) \exp\Big(\tfrac{1}{2} \sum_{s=1}^{t-1} D_H^2\big(p_s(P), p_s(P^\star)\big) \Big)$$

satisfies $\mathbb{E}\big[W_t(P) \mid \mathcal{G}_{t-1}\big] \le W_{t-1}(P)$. Hence $(W_t(P))_t$ is a non-negative supermartingale with $\mathbb{E}[W_t(P)] \le W_1(P) = 1$, and by Markov's inequality, for any $P \in \mathcal{P}$ and $\delta' \in (0,1)$,

$$\Pr\left[\tfrac{1}{2} \sum_{s=1}^{t-1} D_H^2(p_s(P), p_s(P^\star)) \;>\; -\log M_t(P) + \log(1/\delta') \right] \le \delta'. \tag{80}$$

Setting $\delta' = \delta/(T|\mathcal{P}|)$ and taking a union bound over $P \in \mathcal{P}$ and $t \in [T]$, the event in Eq. (80) fails for all (P,t) with probability at least $1-\delta$. On this event, plugging in $P = \widehat{P}_t$

and using $-\log M_t(\widehat{P}_t) \le 0$ by the MLE optimality $\sum_{s<t} \log p_s(\widehat{P}_t)^{I_s}(1-p_s(\widehat{P}_t))^{1-I_s} \ge \sum_{s<t} \log p_s(P^\star)^{I_s}(1-p_s(P^\star))^{1-I_s}$,

$$\sum_{s=1}^{t-1} D_H^2\big(p_s(\widehat{P}_t), p_s(P^\star)\big) \le 2\log\big(T|\mathcal{P}|/\delta\big) \qquad \forall\, t \in [T].$$

Finally, the standard Hellinger-vs.-TV comparison for Bernoulli distributions gives $D_H^2(p,q) \ge \frac{1}{2}(p-q)^2$ for all $p, q \in [0,1]$, so

$$\sum_{s=1}^{t-1} \big(\widehat{P}_t - P^\star\big)^2(\mathbf{y}_s \succ \widehat{\mathbf{y}}_s \mid \mathbf{x}_s) \le 2\sum_{s=1}^{t-1} D_H^2\big(p_s(\widehat{P}_t), p_s(P^\star)\big) \le 4\log\big(T|\mathcal{P}|/\delta\big).$$

□

Lemma B.28 (Freedman's inequality Freedman (1975)) *Let $\{X_t\}_{t=1}^T$ be a sequence of real-valued random variables adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^T$ satisfying $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$ and $|X_t| \le R$ almost surely. Let $V_T := \sum_{t=1}^T \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}]$ denote the predictable quadratic variation. Then for any $\delta \in (0,1)$ and any $\eta \in (0, 1/R)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^T X_t \le \eta V_T + \frac{\log(1/\delta)}{\eta}.$$

In particular, applying this with $X_t = \mathbb{E}_t[Y_t] - Y_t$ for $Y_t \in [0,1]$ adapted, and noting that $\mathbb{E}_t[X_t^2] \le \mathbb{E}_t[Y_t]$, with probability at least $1-\delta$,

$$\sum_{t=1}^T \mathbb{E}_t[Y_t] \le 2\sum_{t=1}^T Y_t + 4\log(1/\delta).$$

C Additional details in DENPO and numerical experiments

C.1 Details on Direct ENPO (DENPO)

In this section, we discuss a simplified version of ENPO: Direct ENPO (DENPO).

For ease of presentation, we summarize the notation for policies and datasets. At iteration t, let π_t be the output policy learner and $\hat{\pi}_t$ be the output adversarial explorer. We use D_1^t for the SFT dataset sampled from π_{t-1}, and D_2^t for the contrastive dataset of response pairs from π_{t-2} and $\hat{\pi}_{t-1}$ with preference labels. Finally, $\tilde{D}_2^t$ denotes the re-labeled dataset used to train the adversarial explorer via the DPO-style objective.

SFT-type regularizer. The exploitation update drives the current policy in the direction of a mirror descent update. First, we sample n triples $(\mathbf{x}, \mathbf{y}, \mathbf{y}')$ by drawing $\mathbf{x}$ from a fixed prompt distribution, $\mathbf{y}$ from the policy learner π_{t-2}, and $\mathbf{y}'$ from the adversarial explorer $\hat{\pi}_{t-1}$, and use them to construct the dataset D_2^t for the contrastive objective. We then sample an i.i.d. comparator response $\hat{\mathbf{y}}$ from $\pi_{t-1}(\cdot|\mathbf{x})$ and query preference labels $I_t(\mathbf{y}) \sim \text{Bern}(P^\star(\mathbf{y} \succ \hat{\mathbf{y}}|\mathbf{x}))$ and $I_t(\mathbf{y}') \sim \text{Bern}(P^\star(\mathbf{y}' \succ \hat{\mathbf{y}}|\mathbf{x}))$ to estimate the mirror descent gradient. In addition, we generate or reuse m samples $(\mathbf{x}, \tilde{\mathbf{y}})$ from π_{t-1} to construct a dataset D_1^t for the SFT-type regularizer. We then minimize the loss function $\mathcal{L}_{\text{md}}(\pi, D_1^t, D_2^t) :=$

$$-\alpha_t \sum_{(\mathbf{x},\tilde{\mathbf{y}}) \in D_1^t} \log \pi(\tilde{\mathbf{y}}|\mathbf{x}) + \sum_{(\mathbf{x},\mathbf{y},\mathbf{y}', I_t(\mathbf{y}), I_t(\mathbf{y}')) \in D_2^t} \frac{1}{n}\Big(h_t(\pi, \mathbf{x}, \mathbf{y}, \mathbf{y}') - \eta\big(I_t(\mathbf{y}) - I_t(\mathbf{y}')\big)\Big)^2, \quad (81)$$

where

$$h_t(\pi, \mathbf{x}, \mathbf{y}, \mathbf{y}') := \log \frac{\pi(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{y}'|\mathbf{x})} - \eta\beta \log \frac{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}'|\mathbf{x})} - (1-\eta\beta)\log \frac{\pi_{t-1}(\mathbf{y}|\mathbf{x})}{\pi_{t-1}(\mathbf{y}'|\mathbf{x})}. \quad (82)$$

The validity of this loss function can be verified by an argument similar to those in Azar et al. (2024); Zhang et al. (2025b).

Algorithm 5: Direct Exploratory Nash Preference Optimization (DENPO)

Input : T, coefficients $\gamma, \lambda > 0$, stepsize $\eta > 0$, reference policy π_{ref}, α_t.
Initialize policies $\pi_{-1} \leftarrow \pi_{\text{ref}}$, $\pi_0 \leftarrow \pi_{\text{ref}}$, $\hat{\pi}_0 \leftarrow \pi_{\text{ref}}$;
Initialize datasets $D_1^0 \leftarrow \varnothing, D_2^0 \leftarrow \varnothing$;
for $t = 1, \ldots, T$ **do**
  Generate n response pairs $\{(\mathbf{x}, \mathbf{y}, \mathbf{y}')\}$ via $\mathbf{x} \sim \rho$, $\mathbf{y} \sim \pi_{t-2}(\cdot|\mathbf{x})$ and $\mathbf{y}' \sim \hat{\pi}_{t-1}(\cdot|\mathbf{x})$;
  Generate or reuse m responses $\{(\mathbf{x}, \tilde{\mathbf{y}})\}$ via $\mathbf{x} \sim \rho$ and $\tilde{\mathbf{y}} \sim \pi_{t-1}(\cdot|\mathbf{x})$;
  Sample response pairs from D_2^t, for each $(\mathbf{x}, \mathbf{y}, \mathbf{y}')$, generate or reuse $(\mathbf{x}, \hat{\mathbf{y}})$ via $\hat{\mathbf{y}} \sim \pi_{t-1}(\cdot \mid \mathbf{x})$;
  Query preference labels $I(\mathbf{y}) \sim \text{Bern}(P^\star(\mathbf{y} \succ \hat{\mathbf{y}} \mid \mathbf{x}))$ and $I(\mathbf{y}') \sim \text{Bern}(P^\star(\mathbf{y}' \succ \hat{\mathbf{y}} \mid \mathbf{x}))$;
  Update datasets $D_1^t \leftarrow \{(\mathbf{x}, \tilde{\mathbf{y}})\}$, $D_2^t \leftarrow D_2^{t-1} \cup \{(\mathbf{x}, \mathbf{y}, \mathbf{y}', I(\mathbf{y}), I(\mathbf{y}'))\}$;
  Compute $\pi_t \leftarrow \text{argmin}_{\pi \in \Pi} \mathcal{L}_{\text{md}}(\pi, D_1^t, D_2^t)$, where $\mathcal{L}_{\text{md}}(\pi, D_1^t, D_2^t)$ is defined in Eq. (81);
  Compute $\hat{\pi}_t \leftarrow \text{argmin}_{\tilde{\pi} \in \Pi} \mathcal{L}_{\text{exp}}(\tilde{\pi}, D_2^t)$, where $\mathcal{L}_{\text{exp}}(\tilde{\pi}, D_2^t)$ is defined in Eq. (87);
Output : $\text{best}(\{\pi^t\}_{t=1}^T)$.

The first term in Eq. (81) is an SFT-type bonus that strengthens the anchor to the current policy learner. This bonus differs from the more common exploration bonuses used in online RL. The additional regularization can improve the learning process by enhancing the proximal effect and mitigating the instability that arises when learning an equilibrium. This modification is also in the same spirit as Magnetic Mirror Descent (Sokota et al., 2023), which augments regularization for learning in games.

To guide the reader through the sequence of objectives below, we briefly summarize their roles. Eq. (81) defines the exploitation loss used to update the policy learner. Eq. (83) introduces the theoretical adversarial explorer objective. The subsequent derivations reformulate and simplify this objective, leading to the practical DPO-style loss in Eq. (87).

DPO-type explorer surrogate. After the exploitation update, we invoke an exploration oracle to identify an adversarial sampler by solving the following policy maximization problem without collecting additional samples:

$$\max_{\tilde{\pi} \in \Pi} \max_{\pi \in \Pi} \frac{\left(\frac{1}{m} \sum_{(\mathbf{x}, \tilde{\mathbf{y}}) \in D_1^t} \log \frac{\pi_t(\tilde{\mathbf{y}}|\mathbf{x})}{\pi(\tilde{\mathbf{y}}|\mathbf{x})} - \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y}' \sim \tilde{\pi}(\cdot|\mathbf{x})} \left[\log \frac{\pi_t(\mathbf{y}'|\mathbf{x})}{\pi(\mathbf{y}'|\mathbf{x})}\right]\right)^2}{\lambda + \frac{1}{n} \sum_{(\mathbf{x}, \mathbf{y}, \mathbf{y}') \in D_2^t} \left(\log \frac{\pi_t(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{y}|\mathbf{x})} - \log \frac{\pi_t(\mathbf{y}'|\mathbf{x})}{\pi(\mathbf{y}'|\mathbf{x})}\right)^2}. \tag{83}$$

One can use a surrogate of this oracle in practice. Applying the identity $a^2/b = \max_\omega \{2a\omega - b\omega^2\}$ for $b > 0$, we recast Eq. (83) as

$$\max_{\tilde{\pi} \in \Pi} \max_{\pi \in \Pi} \max_{\omega} 2\omega \Big(\mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y}' \sim \tilde{\pi}(\cdot|\mathbf{x})} \left[\log \pi_t(\mathbf{y}'|\mathbf{x}) - \log \pi(\mathbf{y}'|\mathbf{x})\right] - \frac{1}{m} \sum_{(\mathbf{x}, \tilde{\mathbf{y}}) \in D_1^t} \log \frac{\pi_t(\tilde{\mathbf{y}}|\mathbf{x})}{\pi(\tilde{\mathbf{y}}|\mathbf{x})} \Big)$$
$$- \omega^2 \Big(\lambda + \frac{1}{n} \sum_{(\mathbf{x}, \mathbf{y}, \mathbf{y}') \in D_2^t} \Big(\log \frac{\pi_t(\mathbf{y}|\mathbf{x})}{\pi_t(\mathbf{y}'|\mathbf{x})} - \log \frac{\pi(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{y}'|\mathbf{x})} \Big)^2 \Big). \tag{84}$$

The two empirical terms depending on π act as implicit anchors: each penalizes the log-density gap between π and π_t on observed data, mimicking the geometry of a KL regularizer. We then derive a tractable surrogate loss as follows. First, we approximate the implicit regularization terms using a simpler term:

$$\max_{\tilde{\pi} \in \Pi} \max_{\pi \in \Pi} \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y} \sim \tilde{\pi}(\cdot|\mathbf{x})} \left[\log \pi_t(\mathbf{y}|\mathbf{x}) - \log \pi(\mathbf{y}|\mathbf{x})\right] - c\, \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y} \sim \pi_{t-1}(\cdot|\mathbf{x})} \left[\log \frac{\pi_t(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{y}|\mathbf{x})}\right],$$

where c is a constant hyperparameter. This subproblem, up to a constant independent of π, is equivalent to

$$\max_{\tilde{\pi} \in \Pi} \max_{\pi \in \Pi} \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y} \sim \tilde{\pi}(\cdot|\mathbf{x})} \left[\log \pi_t(\mathbf{y}|\mathbf{x}) - \log \pi(\mathbf{y}|\mathbf{x})\right] - c\, \mathbb{E}_{\mathbf{x} \sim \rho} \left[D_{\text{KL}}(\pi_{t-1}(\cdot|\mathbf{x}) \| \pi(\cdot|\mathbf{x}))\right]. \tag{85}$$

Then, setting $\pi = \pi_{t-1}$ collapses the inner maximization and zeros out the KL anchor, leaving an objective in $\tilde{\pi}$ alone. The remaining linear term depends on $\tilde{\pi}$ only through its mass on $\text{supp}(\pi_t) \cap \text{supp}(\pi_{t-1})$—we pin this effect down with a KL anchor of strength $b > 0$ against π_{t-1}, which yields

$$\max_{\tilde{\pi} \in \Pi} \mathbb{E}_{\mathbf{x} \sim \rho, \mathbf{y}' \sim \tilde{\pi}} \left[\log \pi_t(\mathbf{y}'|\mathbf{x}) - \log \pi_{t-1}(\mathbf{y}'|\mathbf{x})\right] - b \mathbb{E}_{\mathbf{x} \sim \rho} \left[D_{\text{KL}}(\tilde{\pi}(\cdot|\mathbf{x}) \| \pi_{t-1}(\cdot|\mathbf{x}))\right]. \tag{86}$$

A DPO-style reparameterization of Eq. (86) yields the following surrogate loss:

$$-\sum_{(\mathbf{x},\mathbf{y}_+,\mathbf{y}_-)\in\tilde{D}_2^t} \log\sigma\left(b\log\frac{\tilde{\pi}(\mathbf{y}_+|\mathbf{x})}{\pi_{t-1}(\mathbf{y}_+|\mathbf{x})} - b\log\frac{\tilde{\pi}(\mathbf{y}_-|\mathbf{x})}{\pi_{t-1}(\mathbf{y}_-|\mathbf{x})}\right), \tag{87}$$

where each triple in $\tilde{D}_2^t$ is constructed from a triple $(\mathbf{x},\mathbf{y},\mathbf{y}')$ in D_2^t by assigning $(\mathbf{y}_+,\mathbf{y}_-)$ according to the ordering of $\log\pi_t(\mathbf{y}|\mathbf{x}) - \log\pi_{t-1}(\mathbf{y}|\mathbf{x})$ and $\log\pi_t(\mathbf{y}'|\mathbf{x}) - \log\pi_{t-1}(\mathbf{y}'|\mathbf{x})$. We summarize this simplified version of ENPO, called DENPO, in Algorithm 5.

We summarize the algorithm in Algorithm 5.

C.2 Computational cost and budget-matched comparisons

The comparison in the main texts fixes the number of prompts and iterations, but the methods have different costs per iteration. We therefore report sampling, preference feedback, generated tokens, elapsed time, and optimization work separately. All runs use 20,000 prompts per iteration on a node with eight H100 GPUs. We list the per-prompt cost for each methods in Table 2.

Table 2: Computational cost per prompt for different methods.

Method	Generated completions	Preference queries	Judge evaluations	Policy-training stages
XPO	16	N/A	16	1
INPO	8	10	20	1
DENPO	8	8	16	2

XPO generates 8 current-policy and 8 reference-policy completions. INPO generates 8 completions and applies its tournament procedure. DENPO generates 1 learner candidates, 1 explorer candidates, and 4 comparator responses, comparing the two candidates with the four comparators gives 8 pairwise queries. DENPO generates two fresh generations in the exploration phase. DENPO has two policy-training stages per iteration, one for the learner and one for the explorer (these are not counts of individual gradient updates).

Other computational-cost-per-iteration measures are listed in Table 3. We can see that DENPO is *not* substantially more costly than baselines in a variety of measures.

Table 3: Measured per-iteration cost on eight H100s in the seed-1 runs.

Method	Tokens	Wall-clock (h)	GPU-hours	Optimizer steps	Training stages
XPO	100M	1.567	12.54	217	1
INPO	54M	0.897	7.17	109	1
DENPO	68M	1.990	15.92	311	2

Sampling and feedback budgets. Table 4 compares checkpoints using the same evaluation protocol as Table 1. Its scores are means across the three training runs, while the completion and query counts are fixed by the sampling protocols. DENPO$_{1/8}$ after two iterations exceeds the strongest iteration-3 INPO variant and XPO on all four reported metrics, despite using 320K rather than 480K and 960K completions consumed by those baselines, respectively. It also requires 320K pairwise preference queries versus INPO's 600K. Thus, the quality gain is already visible before DENPO's third iteration.

Table 4: Quality versus cumulative sampling cost.

Checkpoint	Used comp.	Ordinal	AE2-mini	AE2-turbo	MT-Bench	Arena-Hard
XPO, iter. 3	960K	—	47.06	40.75	6.91	45.33
INPO, iter. 3	480K	600K	53.69	45.81	6.99	50.83
DENPO$_{1/8}$, iter. 2	**320K**	**320K**	**56.22**	**46.92**	**7.03**	**53.03**
DENPO$_{1/8}$, iter. 3	480K	480K	58.33	47.93	7.04	55.10

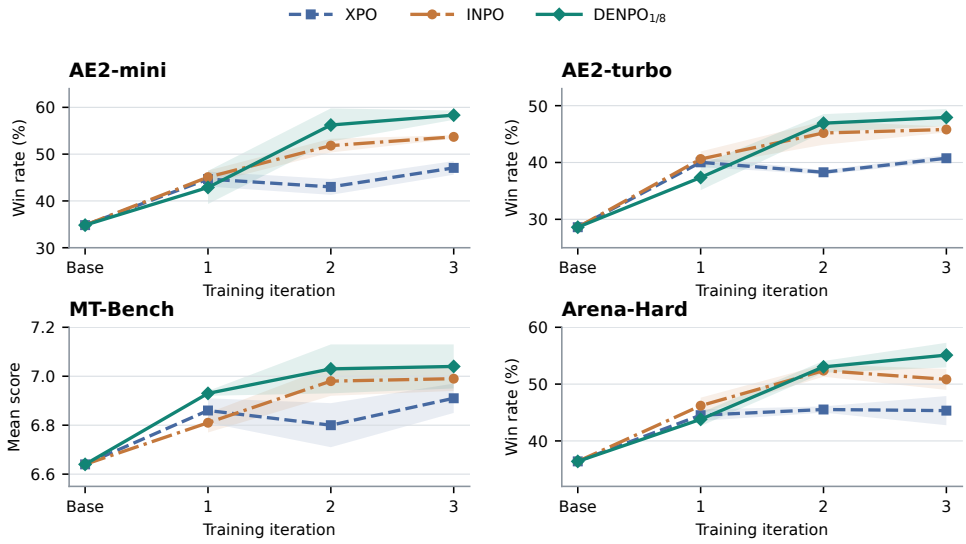

Figure 1: Evaluation performance over training iterations. AE2-mini, AE2-turbo, and Arena-Hard report win rates (%); MT-Bench reports a mean score on a 0–10 scale. Curves show three-run checkpoint means and shading shows one sample standard deviation at iterations 1–3. The common base-model point has no training-run variation. Higher is better.

Matched GPU-hour budgets. Table 5 uses the fully cost-tracked seed-1 run to compare checkpoints close in elapsed GPU-hours. DENPO after two iterations uses fewer GPU-hours and optimizer steps than XPO after three, while scoring higher on all four benchmarks. For INPO, we continued its strongest seed-1 variant from iteration 3 through iteration 6. This consumed 45.90 GPU-hours, within 4.2% of DENPO's 47.89 GPU-hours at iteration 3. The final DENPO checkpoint scores 57.88 versus 45.56 on AE2-mini and leads on the other three metrics. Even choosing INPO's *best earlier checkpoint* within those six iterations yields 54.04 AE2-mini, 45.22 AE2-turbo, 7.08 MT-Bench, and 51.8 Arena-Hard: DENPO remains ahead except for a tie on MT-Bench.

Table 5: Single-run, near-matched measured budgets.

Checkpoint	Wall (h)	GPU-h	Steps	AE2-mini	AE2-turbo	MT-Bench	Arena-Hard
XPO, iter. 3	4.710	37.680	651	45.65	40.35	6.97	42.6
DENPO$_{1/8}$, iter. 2	**3.912**	**31.296**	**621**	**57.40**	**45.84**	**7.08**	**52.2**
INPO, iter. 6	5.737	45.898	658	45.56	37.84	6.66	33.8
DENPO$_{1/8}$, iter. 3	**5.986**	**47.888**	**934**	**57.88**	**46.46**	**7.08**	**54.0**

The matched budget experiments demonstrates the convincing performances of DENPO.

C.3 Performance across training iterations

Figure 1 traces the four evaluation metrics from the base model through iterations 1-3. We display XPO, the default-parameter INPO variant, and DENPO$_{1/8}$, the strongest iteration-3 variants of INPO and DENPO in Table 1. The checkpoint values are means over three training runs. The shaded bands show their sample standard deviations. The base model is a single evaluation shared by the three trajectories.

The methods do not improve monotonically on every metric. XPO's two AlpacaEval-2 scores decline at iteration 2 before recovering at iteration 3; INPO's Arena-Hard mean falls from 52.37 at iteration 2 to 50.83 at iteration 3. DENPO$_{1/8}$ trails INPO on both AlpacaEval-2 metrics after the first iteration, then exceeds it on all four metrics at iterations 2 and 3. At iteration 2 its mean advantage over INPO is 4.40 points on AE2-mini, 1.73 points on AE2-turbo, 0.05 on MT-Bench, and 0.66 points on Arena-Hard. At iteration 3 the respective margins are 4.64, 2.12, 0.05, and 4.27.