PREFERENCE LEARNING BEYOND CONFIDENCE: EFFICIENT GRADIENT-BASED RLHF VIA ADVERSARIAL PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in Large Language Models (LLMs) and robotics have intensified interest in Reinforcement Learning from Human Feedback (RLHF), where reward functions are learned from pairwise human preferences modeled via sigmoid-based utility functions. Despite their widespread adoption, existing RLHF frameworks either lack theoretical guarantees or are computationally intractable — relying on confidence-set-based methods that cannot scale to continuous, high-dimensional decision spaces. We address this gap by developing `Mirror-Duel`, an efficient gradient descent-based algorithm with provably optimal regret guarantees under adversarially changing preferences, without assuming a fixed stochastic reward model. Formalizing the problem as an adversarial online convex optimization (OCO) problem with only weak pairwise preference feedback, our online mirror descent (OMD) approach achieves an optimal $\tilde{O}(d\sqrt{T})$ regret bound — matching a lower bound — with only $O(dT)$ runtime. Beyond pairwise preferences, we extend `Mirror-Duel` to B-batched and top-m partial ranking feedback, respectively yielding improved $\tilde{O}(d\sqrt{T/\min\{B,d\}})$ and $\tilde{O}(d\sqrt{T/\min\{m,d\}})$ regret guarantees. We validate our theoretical findings extensively on synthetic adversarial instances as well as real-world RL environments (`CartPole-v1`, `Acrobot-v1`), demonstrating that `Mirror-Duel` consistently outperforms UCB-based baselines, including `Online-IPO`, `APO`, and `Nash-MD` in regret performance. Our work lays a theoretical and algorithmic foundation for scalable, gradient-based reward learning from human preferences, contributing to the broader effort of improving AI alignment with preferences.

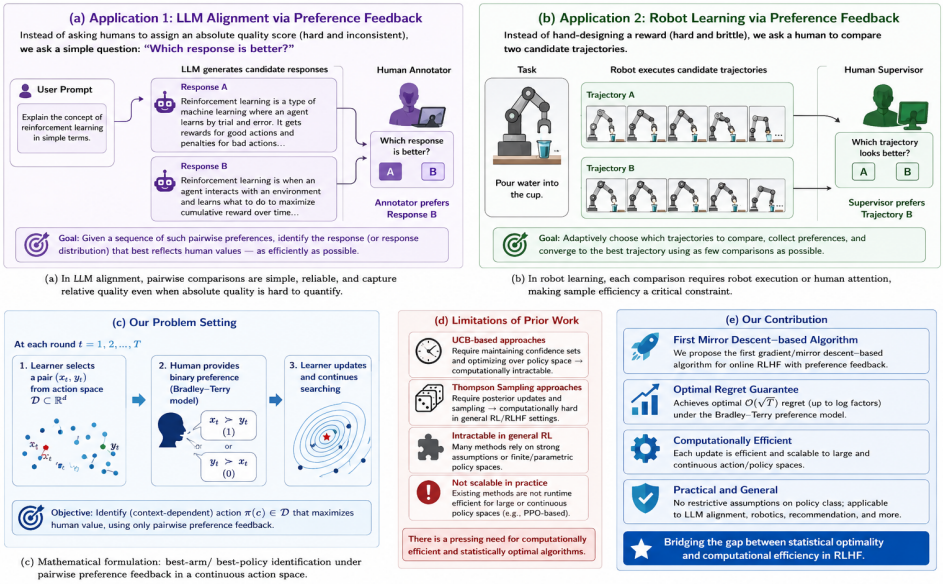
1 INTRODUCTION

The rapidly advancing field of AI has sparked intense interest in Reinforcement Learning from Human Feedback (RLHF), which incorporates human input to refine AI systems, mitigating risks in autonomous decision-making and fostering systems that act aligned with users' best interests. A central challenge in this pipeline is *reward design*: rather than hand-crafting a reward function — a notoriously brittle and expensive process — one seeks to *learn* what humans value directly from their feedback. We explore the theoretical foundations of this reward learning problem under preference feedback, emphasizing its potential towards scalable and sample-efficient alignment.

Human preference feedback is critical in machine learning (ML) for designing predictive AI models, which include demonstrations Hussein et al. (2017); Swamy et al. (2023); Torabi et al. (2018), gradient-based feedback Zinkevich (2003); Boyd et al. (2004); Fletcher (2013), and value-based feedback Flaxman et al. (2005); Cesa-Bianchi & Lugosi (2006b); Shamir (2015); Saha (2021a). Preference feedback is a *weak* form of feedback that receives only the relative desirability (a.k.a. preference) of different outcomes or actions for a given task. Despite its weakness, preference feedback captures a more nuanced understanding of human values and priorities than scalar reward signals, by directly tallying the relative desirability of different outcomes. Studies in psychology and cognitive neuroscience corroborate that humans are naturally more comfortable providing relative feedback than absolute scores (Musallam et al., 2004; Kahneman & Tversky, 1982) — making preference-labeled data less biased, cheaper to collect, and more resource-efficient at scale. This form of feedback therefore enables AI systems to learn more complex and subtle aspects of human intentions that are difficult to encode through demonstrations or explicit reward engineering.

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.

Motivating Applications. While preference-based reward learning arises broadly across AI — including recommendation systems (Christiano et al., 2017; Sadigh et al., 2017), dialogue systems (Ouyang et al., 2022; Rafailov et al., 2024), autonomous driving (Sadigh et al., 2017; Katz et al., 2021), and robot learning (Lee et al., 2021; Christiano et al., 2017) — We highlight two domains where the need for computational efficiency is especially acute, and the connection to our theoretical setting is most direct.



– LLM Alignment. Fine-tuning an LLM to produce helpful, harmless, and honest responses via scalar quality scores is cognitively demanding, inconsistent, and does not scale. Instead, modern RLHF pipelines (Ouyang et al., 2022; Rafailov et al., 2024) ask annotators a simpler question: *given two responses, which is better?* Pairwise comparisons are fast, reliable, and naturally capture relative quality even when absolute quality is hard to quantify. The reward learning problem then becomes: given a sequence of pairwise preferences, identify the response distribution best reflecting human values — efficiently and with provable guarantees on quality loss.

– Training Robots. Specifying reward functions for robotic manipulation tasks — pouring liquid, assembling parts, handing objects — is notoriously difficult: translating human intuitions about "good motion" into mathematical rewards is hard. Instead, a human supervisor can compare two candidate trajectories and indicate which looks better (Christiano et al., 2017; Lee et al., 2021). The reward learning problem is then identical in structure: adaptively select trajectory pairs, collect preference feedback, and converge to the best trajectory using as few comparisons as possible — a critical concern since each comparison requires physical execution or human attention.

In both settings, the underlying mathematical problem is the same: an agent must identify the best arm in a continuous decision space $\mathcal{D} \subset \mathbb{R}^d$ using only pairwise preference feedback, under a logistic (Bradley-Terry) preference model. This is precisely the problem we study.

Problem Setting. To understand the RLHF with preference feedback problem Wu & Sun (2024); Xie et al. (2024); Xiong et al. (2024); Rafailov et al. (2024) more formally: in the online/active exploration RLHF setting, the learner sequentially queries a pair of actions and receives binary 0–1 preference feedback indicating the preferred item. The objective is to find a good (value-maximizing) policy $\pi : \mathcal{C} \mapsto \mathcal{D}$, a mapping from the context space $\mathcal{C}$ to decision (action) space $\mathcal{D}$, as efficiently as possible. The decision space $\mathcal{D}$ represents the set of actions/alternatives to learn from — for language models, $\mathcal{D}$ could be the class of all candidate responses, for robot learning it could be the space of motion trajectories, and for recommendation systems it could be the set of available items.

Limitations of Existing Online RLHF with Preference Feedback Algorithms. The well-cited works of Rafailov et al. (2024); Ouyang et al. (2022); Chen et al. (2024) use offline data in nature which do not allow active exploration and lack convergence guarantees. There has been a recent surge in online RLHF with preference feedback (Xu et al., 2020; Chatterji et al., 2021; Saha et al.,

2023; Saha, 2021a; Kausik et al., 2024; Das et al., 2024), but these algorithms are based on the optimism in the face of the uncertainty (UCB) based principle, which requires maintaining confidence sets and optimizing over the policy space which can be computationally intractable. Few studies (Efroni et al., 2021; Li et al., 2024; Wu & Sun, 2024) have also considered Thompson Sampling (TS) approaches as an alternative, but again updating and sampling from the posterior distribution can be computationally hard, making the approaches impractical for real-world applications. Specifically, quoting from Xie et al. (2024), "However, the most powerful approaches in this space are computationally intractable in the general reinforcement learning (RL) setting (Jiang et al., 2017; Jin et al., 2021; Foster et al., 2021), and prior attempts to adapt them to RLHF either make unrealistic modeling assumptions (i.e., do not allow for general function approximation) (Xu et al., 2020; Novoseller et al., 2020; Saha et al., 2023; Wu & Sun, 2024; Zhan et al., 2023; Du et al., 2024; Das et al., 2024), or are computationally inefficient and not feasible to faithfully implement (Chen et al., 2022; Wang et al., 2023; Ye et al., 2024). In fact, the computational efficiency of Xie et al. (2024) itself is in question since they require optimizing in the policy space using methods like PPO (Schulman et al., 2017) which may not be runtime efficient unless the policy space is finite or parameterized under some restrictive assumptions.

Consequently, there is a pressing need for the development of computationally efficient algorithms that can harness preference feedback in a timely and resource-effective manner, thereby enhancing the feasibility and responsiveness of AI alignment strategies. In this work, we present the first mirror (gradient) descent-based algorithm for the problem with an optimal performance guarantee.

1.1 Contributions

We frame the RLHF problem as a best-arm identification problem in the online linear optimization (OLO) framework Shamir (2015); Hazan (2019) in dimension d, with pairwise preference feedback (see Sec. 2). Note this is the first step in designing gradient descent-based approaches for RLHF which certainly will play a critical role in extending these methods for more complex policy optimization settings.

Key Contribution. Our contributions are multifaceted. At a technical level, we are the first to address the problem of adversarial online linear optimization with preference feedback, which has critical implications in the RLHF with preference feedback literature. Our specific contributions are:
•(1) **Algorithmic Contribution.** Our key contribution is to design an online mirror-descent based (OMD) algorithm to obtain a near-optimal $\tilde{O}(d\sqrt{T})$ regret algorithm for online optimization with adversarial preferences (Alg. 1,Thm. 1).

We urge interested readers to check Sec. A to understand the advantages of gradient descent techniques in-depth. One of our primary contributions in our proposed algorithm *Mirror-Duel*is to estimate the loss function per round from the *weaker preference feedback*. Consequently, we adapt to a different proof analysis to incorporate the changes which lead to near-optimal (up to log factors) $\tilde{O}(d\sqrt{T})$ regret bounds of *Mirror-Duel* (Thm. 1). •(2) **Performance Limit Analysis.** To understand the effectiveness of our analysis, we further prove a matching lower bound of $\Omega(d\sqrt{T})$ to show that our algorithm is within a logarithmic factor of the optimal performance limit (Thm. 3). Deriving the lower bound for this problem was non-trivial, we derive this from the first principle lifting tools from the classical literature of information theory.

Additional Contributions. We extended our above result with pairwise preferences as follows:

1. In Sec. 4, we first generalize the above algorithm to multi-wise (batched) preference feedback, where the learner can query a set of B pairwise preferences in one go. In many settings, it is not feasible to actively update the model's prediction after every round, perhaps due to communication delays, parallel processing or time/ cost overhead. Instead, the system may prefer to collect a batch of comparison queries and then update its model to generate the next set of queries. In such settings, batched RLHF is a natural model to consider. Our improved analysis of Alg. 2, the batched variant of *Mirror-Duel* shows that one can achieve a faster $\tilde{O}(\frac{d}{\sqrt{\min\{B,d\}}}\sqrt{T})$ regret learning rate for this case (Thm. 4).
2. Next we consider another interesting feedback model in Sec. 5 which generalizes pairwise preference feedback to subsets of size k (for any $k \geq 2$), and allows the learners to query partial rank-ordered feedback of length $m \in \{1, \dots k\}$. The objective was to understand if the learning algorithm is allowed to query from a larger set of k alternatives and obtain a richer m-length

ranking feedback, can it learn faster? What is the optimal trade-off of the learning rate with m and k? Our proposed algorithm *Mirror-Rank* addresses this setting with a regret guarantee of $\tilde{O}(\frac{d}{\sqrt{\min\{m,d\}}}\sqrt{T})$. The k (subset size) independence of the result could be surprising to many, as one may expect that larger subsets may lead to faster convergence! However, this is not the case we explained in Rem. 5. On the other hand, as expected the regret indeed improves with the increasing length of the rank-ordered feedback m. Our algorithm *Mirror-Rank* actually exploits the key ideas of our batched algorithm Alg. 2 by *cleverly extracting m-batched pairwise preference information from the top-m ranking feedback $\sigma_{m,t}$*. We describe the algorithm in Sec. 5 and its regret performance follows almost immediately from Thm. 4.

3. Same as Thm. 3, we corroborate the performance analysis of all our algorithms with their corresponding lower bound analysis to understand the tightness of our algorithmic guarantees. Precisely, for the B-batched feedback setting we show that our regret guarantee of *Mirror-BPref* is optimal (up to log factors) with a matching lower bound analysis Thm. 6. Similarly, our top-m ranking feedback lower bound of Thm. 9 justifies the tightness of regret performance of *Mirror-Rank*. All our lower bounds are derived from the first principles of information theory.
4. Finally, we also corroborate our results empirically in Sec. 6 on both synthetic and real datasets. We especially compare the performance of our proposed method with multiple existing baselines (App. G.4) on Real-RL Environments and validated our superior performance (App. G.3).

2 Problem Setup

Notation. Let $[n] = \{1, \dots n\}$, for any $n \in \mathbb{N}$. Given a set S and two items $x, y \in S$, we denote by $x \succ y$ the event x is preferred over y. For any $r > 0$, let $\mathcal{B}_d(r)$ and $\mathcal{S}_d(r)$ denote the ball and the surface of the sphere of radius r in d dimensions, respectively. $\mathbf{I}_d$ denotes the $d \times d$ identity matrix, $\mathbf{1}_d$ denotes the d-dimensional vector of all 1s. For any vector $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_2$ denotes the ℓ_2 norm of vector $\mathbf{x}$. $\mathbf{1}(\varphi)$ is generically used to denote an indicator variable that takes the value 1 if the predicate φ is true and 0 otherwise. $\text{Unif}(S)$ denotes a uniform distribution over any set S. We write $\tilde{O}$ for the big O notation up to logarithmic factors. For any set $\Omega \subset \mathbb{R}^d$, $\text{int}(\Omega)$ denotes the interior of the set Ω. $\text{Ber}(p)$ defines *Bernoulli* distribution with parameter $p \in [0, 1]$.

2.1 Problem: Adversarial Logistic Dueling Bandits (Pref-DB):

We consider online T round sequential decision-making game on a decision space $\mathcal{D} \subset \mathbb{R}^d$ in the *Adversarial Online Linear Optimization (Bandits)* Hazan (2019); Abernethy et al. (2008) framework. At every round, the algorithm plays $\mathbf{x}_t, \mathbf{y}_t \in \mathcal{D}$ and observes a binary feedback o_t s.t. $o_t \sim \text{Ber}\Big(\sigma\big({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}_t - \mathbf{y}_t)\big)\Big)$. $\forall \mathbf{x}, \mathbf{y} \in \mathcal{D}$, let the probability of arm $\mathbf{x}$ preferred over arm $\mathbf{y}$ is:

$$P_t(\mathbf{x}, \mathbf{y}) = \sigma\big({\boldsymbol{\theta}_t^*}^\top(\mathbf{x} - \mathbf{y})\big) = \frac{\exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x})}{\exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}) + \exp({\boldsymbol{\theta}_t^*}^\top \mathbf{y})}. \tag{1}$$

Note we call the problem Pref-DB since the preference relation P_t follows a logistic model, as $\sigma : \mathbb{R} \mapsto [0, 1]$ is the logistic link function, i.e. $\sigma(x) = (1 + e^{-x})^{-1}$.

Objective-I: Regret Minimization w.r.t. the Best Choice. The goal of the algorithm is to minimize the cumulative regret, defined as;

$$\text{Reg}_T^{\text{Logit-DB}} := \sum_{t=1}^{T} \frac{(P_t(\mathbf{x}^*, \mathbf{x}_t) - 1/2) + (P_t(\mathbf{x}^*, \mathbf{y}_t) - 1/2)}{2}, \tag{2}$$

assuming $\mathbf{x}^* \leftarrow \arg\max_{\mathbf{x} \in \mathcal{D}} \sum_{t=1}^{T} {\boldsymbol{\theta}_t^*}^\top \mathbf{x}$ the best (highest scoring) arm in the hindsight.

Remark 1. *For any $x \in \mathcal{D}$, note then $\frac{{\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})}{4} \le P_t(\mathbf{x}^*, \mathbf{x}) - 1/2 \le {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})$, when $\mathcal{D} \subseteq \mathcal{B}_d(1)$. We prove this in Sec. B. Consequently, we address the regret $\widehat{\text{Reg}}_T^{\text{Logit-DB}} := \sum_{t=1}^{T} \left[\frac{{\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x}_t) + {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{y}_t)}{2}\right]$, noting $\text{Reg}_T^{\text{Logit-DB}} \le \widehat{\text{Reg}}_T^{\text{Logit-DB}}$ from Rem. 1, thus designing algorithm to bound $\widehat{\text{Reg}}_T^{\text{Logit-DB}}$ would suffice to bound $\text{Reg}_T^{\text{Logit-DB}}$.*

Objective-II: Sample Complexity. One can also consider a different learning objective where instead of regret minimization the goal is to find 'ϵ-best arm'. Precisely, we define an arm $\hat{\mathbf{x}} \in \mathcal{D}$ to be an *ϵ-best arm* if $\frac{1}{T}\sum_{t=1}^{T} {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \hat{\mathbf{x}}) < \epsilon$. The goal could be find any such $\hat{\mathbf{x}} \in \mathcal{D}$ with the least number of samples.

Remark 2. *An attentive reader might have already noticed that for our problem setting,* regret minimization *is a stronger objective than the latter as a regret guarantee of a learning algorithm immediately yields a valid* sample complexity *bound for the same setting* $\hat{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T} \frac{(\mathbf{x}_t + \mathbf{y}_t)}{2}$.

2.2 Problem Setup: Adversarial Batched Pref-DB

As motivated in the introduction, from a practical viewpoint it might be hard to actively update the parameters of the learning algorithm at each round t due to time or cost overhead. Distributed deployment of the system might also hinder an active sequential adaptation of the learning algorithm due to parallel processing. Thus a natural variant of Pref-DB problem is to consider a *batched feedback model* where the learner actively queries B-pairwise queries in a batched fashion: Precisely, at each round t, the learner queries B-pairs $\{(\mathbf{x}_t^1, \mathbf{y}_t^1), (\mathbf{x}_t^2, \mathbf{y}_t^2), \ldots, (\mathbf{x}_t^B, \mathbf{y}_t^B)\}$ together and observes the corresponding B-pairwise preferences $o_t^1, o_t^2, \ldots o_t^B$, where $o_t^i \sim \text{Ber}\big(\sigma\big({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}_t^i - \mathbf{y}_t^i)\big)\big)$.

Regret Objective. Same as $\widehat{\text{Reg}}_T^{\text{BPref-DB}}$, the objective of the learner, in this case, is to minimize the regret over T rounds, defined as $\widehat{\text{Reg}}_T^{\text{BPref-DB}} := \sum_{t=1}^{T} \frac{1}{B}\left[\sum_{i=1}^{B} \frac{{\boldsymbol{\theta}_t^*}^\top(\mathbf{x}_t^* - \mathbf{x}_t^i) + {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}_t^* - \mathbf{y}_t^i)}{2}\right]$. Note we can also consider the exact same *Sample complexity* objective in this setting as well, as defined above for Pref-DB in Sec. 2.1.

2.3 Problem Setup: Adversarial Multinomial Logit Bandits (Rank-MNL):

In this setting, we generalize the Pref-DB problem (Sec. 3) to the subsetwise case, where the learner can observe preference feedback over a subset of items. More formally, as before, we consider a decision space $\mathcal{D} \subset \mathbb{R}^d$ and at every round the algorithm plays a k-subset $S_t := \{\mathbf{x}_t^1, \mathbf{x}_t^2, \ldots \mathbf{x}_t^k\} \subseteq \mathcal{D}$ and gets to see *Categorial* feedback $o_t \in [k]$ indicating the index of the winning arm in S_t s.t. $o_t \sim \text{Categorial}\big(p_1^t, p_2^t, \ldots, p_k^t\big)$, where $p_i^t = \frac{\exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}_t^i)}{\sum_{j=1}^{k} \exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}_t^j)}$.

In fact, one could even consider a top-m ranking generalization of the above feedback model. Let, for any subset S, $\Sigma_S = \{\sigma \mid \sigma \text{ is a permutation over items of } S\}$, where for any permutation $\sigma \in \Sigma_S$, $\sigma(i)$ denotes the element at the i-th position in σ, $i \in [|S|]$. Then one can further define $\Sigma_S^m := \{\sigma_m := (\sigma(1), \ldots, \sigma(m)) \mid \sigma \in \Sigma_S\}$ and formally define the top-m ranking feedback as:

Generalized Top-m ranking of items (TR-m): In this case at every round t the environment returns a ranking of the top m items from S_t ($|S_t| = k$) by drawing a full ranking $\sigma_t \in \Sigma_{S_t}$ over S_t according to Plackett-Luce (PL) model without replacement, and returns the first m ranked elements of σ, i.e., $(\sigma(1), \ldots, \sigma(m))$. More precisely, if $\sigma_{m,t}$ is any random top-m ranking on a k-subset S_t drawn according to the multinomial (MNL) model (Saha & Gopalan, 2019), then for every position $i \in [m], 1 \le m \le k-1$, $\sigma_{m,t}(i) \sim \text{Categorial}\big(p_1^t, p_2^t, \ldots, p_k^t \setminus \{p^t_{\sigma_{m,t}(1)}, p^t_{\sigma_{m,t}(2)}, \ldots p^t_{\sigma_{m,t}(i-1)}\}\big)$, where $p_i^t = \frac{\exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}_t^i)}{\sum_{j=1}^{k} \exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}_t^j)}$, is drawn by successively sampling m winners from S according to the PL model, without replacement. Thus (equivalently) we have $\mathbf{P}_t(\sigma_{m,t}|S) =:$

$$\prod_{i=1}^{m} \frac{\exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}_t^{\sigma_{m,t}(i)})}{\sum_{j=i}^{m} \exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}_t^{\sigma_{m,t}(j)}) + \sum_{j \in S_{t\setminus m}} \exp({\boldsymbol{\theta}_t^*}^\top \mathbf{x}_t^{\sigma_{m,t}(j)})}, \quad \text{s.t. } S_{t\setminus m} = S_t \setminus \{\sigma_{m,t}(i)\}_{i=1}^{m}.$$

Note for $k = 2$, the top-m ranking feedback is equivalent to the dueling feedback in Sec. 2.1.

Regret Objective. Following the objective from Pref-DB from Sec. 2.1 and Rem. 1, one can similarly define the regret for this setup as: $\widehat{\text{Reg}}_T^{\text{MNL}} := \sum_{t=1}^{T}\left(\frac{1}{k}\sum_{\mathbf{x} \in S_t} {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})\right)$,again assuming $\mathbf{x}^* := \arg\max_{\mathbf{x} \in \mathcal{D}} \sum_{t=1}^{T} {\boldsymbol{\theta}_t^*}^\top \mathbf{x}$ the best arm in the hindsight.

3 DUELING CASE: ALGORITHM FOR PREF-DB PROBLEM

Algorithm description: Our algorithm is motivated by the Scrible algorithm from Abernethy et al. (2008); Hazan (2019), which is a variant of the online mirror descent algorithm with a self-concordant barrier Boyd et al. (2004) as the regularizer.[1] The algorithm iteratively updates the decision variable $\mathbf{w}_t$ by minimizing the sum of the ψ-regularized linearized loss within the δ-contracted decision set $\mathcal{D}_\delta := \{\mathbf{x} \mid \frac{1}{1-\delta}\mathbf{x} \in \mathcal{D}\}$. Precisely at each step t, we compute $\mathbf{w}_t = \arg\min_{\mathbf{w}\in\mathcal{D}_\delta} \left\{\eta \sum_{\tau=1}^{t-1} \mathbf{g}_\tau^\top \mathbf{w} + \psi(\mathbf{w})\right\}$. We then perform eigendecomposition of the Hessian $\nabla^2\psi(\mathbf{w}_t)$, sample an index i_t uniformly at random from $[d]$, and generate perturbed solutions $\mathbf{x}_t = \mathbf{w}_t + \gamma_t \frac{1}{\sqrt{\lambda_{t,i_t}}}\mathbf{v}_{t,i_t}$ and $\mathbf{y}_t = \mathbf{w}_t - \gamma_t \frac{1}{\sqrt{\lambda_{t,i_t}}}\mathbf{v}_{t,i_t}$. It is important here to note that $\mathbf{x}_t, \mathbf{y}_t \in \mathcal{D}$ owing to the properties of self-concordant barrier functions, as argued in Lem. 10. By playing the pair $(\mathbf{x}_t, \mathbf{y}_t)$ and observing the outcome o_t, we construct the gradient estimator $\mathbf{g}_t = \frac{d}{\gamma_t}(o_t - \frac{1}{2})\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}$ for the next iteration and continue to the step iteration. Thm. 1 analyzes the regret performance of Alg. 1 yielding an optimal $O(\sqrt{T})$ regret for the problem, as justified in Rem. 3. The algorithm pseudocode is given in Alg. 1 (Sec. C.1) and the detailed regret analysis of *Mirror-Duel* (Alg. 1) is given in Sec. C.3.

Theorem 1 (Regret Analysis of Alg. 1)**.** *Consider the decision space $\mathcal{D}$, such that $\nabla^2\psi(\mathbf{w}) \geq H^2_{\mathcal{D},\psi}\mathbf{I}_d$, $\forall \mathbf{w} \in \mathcal{D}$. Then for the choice of $\eta = \frac{\sqrt{\nu}H_{\mathcal{D},\psi}}{d\sqrt{T\log T}}$, $\delta = \frac{1}{T}$ and $\gamma_t \leq 0.7H_{\mathcal{D},\psi}$, the Mirror-Duel (Alg. 1) guarantees a regret bound: $\widehat{\mathrm{Reg}}_T^{\textit{Mirror-Duel}} \leq O\left(\frac{d\sqrt{\nu T\log T}}{H_{\mathcal{D},\psi}}\right)$.*

It is worth noting that $H_{\mathcal{D},\psi}$ is generally a problem-dependent constant for bounded decision sets and most choices of ψ, as we explain in Rem. 6. Further note Thm. 1 holds assuming the minimal eigenvalue of $\nabla^2\psi(\mathbf{w})$ is larger than $H^2_{\mathcal{D},\psi}$. This assumption is equivalent to assuming $\psi(\mathbf{w})$ is strongly convex and may hold for different choices of decision sets: For example, for a decision set which is the interior of the unit ball $\mathcal{B}_d(1)$ and choosing $\psi(\mathbf{w}) = -\ln(1 - \|\mathbf{w}\|_2^2)$ is a 1-self concordant barrier and it is easy to check that $H^2_{\mathcal{D},\psi} = 2$. Another example could be $\psi(\mathbf{w}) = -\sum_{i=1}^d \ln w_i$, which is a d-self-concordant barrier in the unit ball $\mathcal{B}_d(1)$ and it is straightforward to verify that in this case $H^2_{\mathcal{D},\psi} = d$.

Remark 3 (Optimality of Thm. 1)**.** *The rate depicted in Thm. 1 is optimal (up to logarithmic factors), as follows from the existing lower bound of the Pref-DB problem (Saha (2021b)).*

Remark 4 (Advantage of Our Approach over Existing Algorithms for Pref-DB)**.** *(1) Prior works on online learning in the generalized linear bandit setting Li et al. (2017; 2024) assume a lower bound on the derivative of the sigmoid link function, incurring a multiplicative dependency on $\kappa = \min_{t\in[T]} \arg\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_t^*\|\leq 1} \sigma'(\boldsymbol{\theta}^\top(\mathbf{x}-\mathbf{y}))$ in Pref-DB Saha et al. (2023); Das et al. (2024). We avoid this assumption entirely by exploiting pairwise preferences of symmetrically opposite points $\mathbf{x}_t$ and $\mathbf{y}_t$ (see Lem. 11) — a clear advantage over UCB-based GLM-bandit approaches for Pref-DB. (2) Since our approach is gradient-based, it is computationally efficient with runtime $O(dT)$, unlike prior methods Saha et al. (2023); Kausik et al. (2024); Das et al. (2024).*

The complete proof of Thm. 1 is in Sec. C. We also note that the sample complexity bound of *Mirror-Duel* (Alg. 1) directly follows from Rem. 2, leading to the following result.

Corollary 2 (Sample Complexity Bound of *Mirror-Duel* (Alg. 1))**.** *Under the parameter settings of Thm. 1, the ϵ-sample complexity of Mirror-Duel (Alg. 1) is roughly $O\left(\frac{d^2\nu}{H^2_{\mathcal{D},\psi}\epsilon^2}\right)$.*

Theorem 3 (Regret Lower Bound for Pref-DB Problem)**.** *Consider any fixed time step T. Then for any algorithm $\mathcal{A}$ for the Pref-DB problem, there exists a decision space $\mathcal{D} \subset \mathbb{R}^d$ and a sequence of unknown linear functionals $\boldsymbol{\theta}_1^*, \ldots \boldsymbol{\theta}_T^* \in \mathbb{R}^d$, such that the regret of algorithm $\mathcal{A}$ in T rounds $\widehat{\mathrm{Reg}}_T^{\textit{Pref-DB}} \geq \frac{d\sqrt{T\log T}}{256}$.*

[1] Interested readers may check Luo (2017); Boyd et al. (2004); Hazan (2019) for the properties and examples of self-concordant barrier functions.

4 BATCHED FEEDBACK: ALGORITHM FOR B-BATCHED PREF-DB PROBLEM

This section can be considered a primer to our algorithm for the rank-ordered feedback setting (see Sec. 2.3) that we discuss in Sec. 5, as we will use a reduction of ranked-feedback setting to the batched-feedback setting.

4.1 ALGORITHM FOR BPREF-DB

Our proposed algorithm for this case is *Batched-Mirror-Duel* (*Mirror-BPref*) which is a variant of *Mirror-Duel* we detailed in the previous section for the Pref-DB, and also takes input parameters η, δ, and γ_t, and a ν-self concordant barrier function ψ.

Similar to Alg. 1, in this case too the idea is to build an estimate of $\boldsymbol{\theta}^*_t$ from the pairwise observations. However, due to the batched feedback of size B, we can build an estimate with better variance leading to $\sqrt{B}$-factor improvement in the final learning rate of $O(\frac{d}{\sqrt{B}}\sqrt{T})$. However, one would need $B \leq d$ since it is impossible to obtain a regret rate better than $\Omega(\sqrt{T})$, which is the rate one obtains in the full information setting Zinkevich (2003).

More precisely, at any round t, assuming $\mathbf{w}_t$ is the running estimate of the optimizer over the decision set $\mathcal{D}_\delta$, our proposed algorithm *Mirror-BPref* first computes the eigendecomposition of the Hessian $\nabla^2\psi(\mathbf{w}_t) = \sum_{i=1}^d \lambda_{t,i}\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top$, and samples B indices $i_t^1, i_t^2, \ldots, i_t^B$, uniformly from $[d]$. Upon this it assigns $\mathbf{x}_t^\ell = \mathbf{w}_t + \gamma_t \frac{1}{2\sqrt{\lambda_{t,i_t^\ell}}}\mathbf{v}_{t,i_t^\ell}$ and $\mathbf{y}_t^\ell = \mathbf{w}_t - \gamma_t \frac{1}{2\sqrt{\lambda_{t,i_t^\ell}}}\mathbf{v}_{t,i_t^\ell}$. and plays the batch of B-pairs $\{(\mathbf{x}_t^1, \mathbf{y}_t^1), (\mathbf{x}_t^2, \mathbf{y}_t^2), \ldots, (\mathbf{x}_t^B, \mathbf{y}_t^B)\}$. Upon this it receives the corresponding B pairwise preferences $o_t^1, \ldots, o_t^B$ and computes a gradient $(\boldsymbol{\theta}^*_t)$ estimate $\mathbf{g}_t = \frac{1}{k}\sum_{\ell=1}^k \mathbf{g}_t^\ell$, where $\mathbf{g}_t^\ell = \frac{d}{\gamma_t}\big(o_t^\ell - \frac{1}{2}\big)\sqrt{\lambda_{t,i_t^\ell}}\mathbf{v}_{t,i_t^\ell}$. The process is then repeated for a T rounds, iteratively, refining the running estimate $\mathbf{w}_{t+1}$ by minimizing the sum of the ψ-regularized linearized loss over $\mathcal{D}_\delta$. Due to space limitations, the pseudocode of *Mirror-BPref* is given in Sec. D.1.

Thm. 4 analyzes the regret performance of Alg. 2 which is shown to yield an optimal $O\big(\frac{d}{\min\{d,B\}}\sqrt{T}\big)$ regret for the the problem.

Theorem 4 (Regret Analysis of Alg. 2)**.** *Consider the decision space $\mathcal{D}$, such that $\nabla^2\psi(\mathbf{w}) \geq H^2_{\mathcal{D},\psi}\mathbf{I}$, $\forall \mathbf{w} \in \mathcal{D}$. Then for the choice of $\eta = \frac{\sqrt{\nu \min\{B,d\}}H_{\mathcal{D},\psi}}{d\sqrt{T\log T}}$, $\delta = \frac{1}{T}$ and $\gamma_t \leq 0.7H_{\mathcal{D},\psi}$, the Mirror-BPref (Alg. 1) guarantees a regret bound: $\widehat{\mathrm{Reg}}_T^{BPref\text{-}DB} \leq O\bigg(\frac{d\sqrt{\nu T\log T}}{\sqrt{\min\{B,d\}}H_{\mathcal{D},\psi}}\bigg)$.*

The regret analysis of *Mirror-BPref* (Alg. 2) is given in Sec. D.2. Similar to Corollary 2, one can derive the sample complexity bounds for *Mirror-BPref* (Alg. 2) using Rem. 2:

Corollary 5 (Sample Complexity Bound of *Mirror-BPref* (Alg. 2))**.** *Under the parameter settings of Thm. 4, the ϵ-sample complexity of Mirror-BPref (Alg. 2) is roughly $O\bigg(\frac{d^2\nu}{\min\{d,B\}H^2_{\mathcal{D},\psi}\epsilon^2}\bigg)$.*

Theorem 6 (Regret Lower Bound for Batched Pref-DB Problem)**.** *Consider any fixed time step T and batched size B. Then for any algorithm $\mathcal{A}$ for the B-Batched Pref-DB problem, there exists a decision space $\mathcal{D} \subset \mathbb{R}^d$ and a sequence of unknown linear functionals $\boldsymbol{\theta}^*_1, \ldots \boldsymbol{\theta}^*_T \in \mathbb{R}^d$, such that the regret of algorithm $\mathcal{A}$ in T rounds $\widehat{\mathrm{Reg}}_T^{BPref\text{-}DB} \geq \frac{d\sqrt{T\log T}}{256\sqrt{\min\{B,d\}}}$.*

5 RANKING FEEDBACK: ALGORITHM FOR RANK-MNL PROBLEM

We now investigate the Rank-MNL problem (Sec. 2.3) for the general top-m ranking feedback. We first analyze the fundamental performance limit proving a lower bound for the problem. Following this we design an optimal algorithm matching the lower bound.

5.1 PROPOSED ALGORITHM: *Mirror-Rank*

Useful Notations. We will find it useful to define some notations before describing our main algorithm *Mirror-Rank*.

We denote by $V_n = \{(\pm 1)^n\}$, for any $n \in \mathbb{N}_+$. Clearly $|V_n| = 2^n$. Let $\mathcal{G}(V_n)$ be the graph with vertex set $V_n \subseteq \{\pm 1\}^n$ and there exists an (undirected) edge between two nodes $\mathbf{v}$ and $\tilde{\mathbf{v}}$ iff $\mathbf{v}$ and $\tilde{\mathbf{v}}$

differ only in sign in one of the n coordinates, i.e. $\exists k \in [n],\ v(k) = \tilde{v}(k)$ and $v(k') = \tilde{v}(k')$ for any $k' \neq k$. Clearly the number of neighboring nodes of any vertex $\mathbf{v} \in V_n$ in graph $\mathcal{G}$ is $|\mathcal{N}(\mathbf{v}, \mathcal{G})| = n$. In other words, the degree of any node in graph $\mathcal{G}$ is n. Also, let us define $\ell_k = \lfloor \log k \rfloor$.

Algorithm Description: ***Mirror-Rank*** As before (in Alg. 1 and Alg. 2), this algorithm too maintains a running estimate of the minimizer $\mathbf{w}_t$ (initialized to $\mathbf{w}_1 \in \mathcal{D}$), and find the eigen decomposition of the Hessian at $\mathbf{w}_t$, say $\nabla^2\psi(\mathbf{w}_t) = \sum_{i=1}^{d} \lambda_{i_t} \mathbf{v}_{t,i} \mathbf{v}_{t,i}^\top$.

(i). **Structured Query Sets S_t:** At each time t, it queries a set S_t of k points around $\mathbf{w}_t$ such that for every point $\mathbf{x} \in S_t$, there exists exactly ℓ_k neighboring points which are *symmetrically opposite to* $\mathbf{x}$ *in exactly one of the realization of* $\mathbf{v}_{t,i}$*s*: More precisely, at each time t, the algorithm first samples ℓ_k directions $i_t^j \sim [d]$ for $j \in [\ell_k]$ and let $U_t = \left[\frac{\mathbf{v}_{t,i_t^1}}{\sqrt{\ell_k}}, \ldots, \frac{\mathbf{v}_{t,i_t^{\ell_k}}}{\sqrt{\ell_k}}\right] \in \mathbb{R}^{d \times \ell_k}$. We then define $S_t = \{\mathbf{w}_t + \gamma_t U_t \mathbf{v} \mid \mathbf{v} \in V_{\ell_k}\}$.

Note that by construction indeed $S_t = 2^{\ell_k} \leq k$. Further, note for any point $\mathbf{x} = \mathbf{w}_t + \gamma U_t \mathbf{v} \in S_t$ there exists exactly ℓ_k symmetrically opposing points $\mathbf{x}_i' = \mathbf{w}_t + \gamma U_t \mathbf{v}_i' \in S_t$, for all $\mathbf{v}_i' \in \mathcal{N}(\mathbf{v}, \mathcal{G})$ such that $\frac{(\mathbf{x} - \mathbf{x}_i')}{2\gamma v_i} = \mathbf{v}_{t,i_t^j},\ j \in [\ell_k]$.Given any such point $\mathbf{x}_\mathbf{v} := \mathbf{w}_t + \gamma U_t \mathbf{v}$, let us denote by the set $\mathcal{N}(\mathbf{x}_\mathbf{v}) = \{\mathbf{w}_t + \gamma U_t \mathbf{v}_i' \mid \mathbf{v}_i' \in \mathcal{N}(\mathbf{v}, \mathcal{G})\}$ of all symmetrically opposing points of $\mathbf{x}$ in S_t around $\mathbf{w}_t$ which differs in exactly one of the realizations of $\mathbf{v}_{t,i}$s. This property will play a very crucial role in our analysis.

(ii). **Inferring Pairwise Preferences from Top-m Ranking $\sigma_{m,t} \in \Sigma_{s_t}^m$:** One of our critical observations is Lem. 14. Thanks to this result, we actually break the top-m ranking feedback $\sigma_{m,t}$ to m pairwise comparisons. In particular for any $i \in [m]$, let us denote by $\tilde{\mathbf{x}}_t^{(i)} := \mathbf{x}_t^{\sigma_{m,t}(i)}$, $S_{t \setminus i} := S_t \setminus \{\mathbf{x}_t^{(j)}\}_{j=1}^{i}$, and $\mathbf{x} \succ_\sigma \mathbf{y}$ denotes $\mathbf{x}$ is preferred over $\mathbf{y}$ in the ranking $\sigma \in \Sigma_S$, for all $\mathbf{x}, \mathbf{y} \in S \subseteq \mathcal{D}$. Now note that for any $i \in [m]$, we can always find at least one $\mathbf{z} \in \mathcal{N}(\mathbf{x}_t^{(i)}) \cap \mathcal{S}_{t \setminus i}$ such that $\mathbf{x}_t^{(i)} \succ \mathbf{z}$. For ease of notation, for any such $\mathbf{x}_t^{(i)}$, we denote the corresponding rank-broken pair $\mathbf{z}$ by $\tilde{\mathbf{x}}_t^{(i)}$. Thus by definition $\mathbf{x}_t^{(i)} \succ_{\sigma_{m,t}} \tilde{\mathbf{x}}_t^{(i)},\ \forall i \in [m]$.

(iii). **Extracting m `BPref-DB` (batched pairwise preference) Feedback to Obtain Aggregated $\boldsymbol{\theta}_t^*$ Estimate:** Following the notations from $\#(ii)$ above, we extract all pairwise comparisons $(\mathbf{x}_t^{(\ell)}, \tilde{\mathbf{x}}_t^{(\ell)})$, for all $\ell \in [m]$. Further since by definition $\tilde{\mathbf{x}}_t^{(\ell)} \in \mathcal{N}(\mathbf{x}_t^{(\ell)})$, let us denote $(\mathbf{x}_t^{(\ell)} - \tilde{\mathbf{x}}_t^{(\ell)}) = \gamma_t \mathbf{v}_t^{(\ell)}$, where note $\mathbf{v}_t^{(\ell)} = \mathbf{v}_{t,i}$ for some $i \in [d]$ and construct an aggregated gradient estimate: $\mathbf{g}_t = \frac{1}{B} \sum_{\ell=1}^{m} \mathbf{g}_t^\ell$, where $\mathbf{g}_t^\ell := \frac{d}{2\gamma_t} \sqrt{\lambda_{t,i_t^{(\ell)}}} \mathbf{v}_{t,i_t^{(\ell)}}$.

(iv). **FTRL update of $\mathbf{w}_t$:** Upon finding the gradient estimate $\mathbf{g}_t$, the rest of the algorithm proceeds exactly the same as Alg. 1 (or Alg. 2). More precisely, The algorithm iteratively updates the decision variable $\mathbf{w}_t$ by minimizing the sum of the ψ-regularized linearized loss within the δ-contracted decision set $\mathcal{D}_\delta$ such that $\mathbf{w}_t = \arg\min_{\mathbf{w} \in \mathcal{D}_\delta} \left\{\eta \sum_{\tau=1}^{t-1} \mathbf{g}_\tau^\top \mathbf{w} + \psi(\mathbf{w})\right\}$.

The complete pseudocode of *Mirror-Rank* is given in Sec. E.1.

Theorem 7 (Regret Analysis of *Mirror-Rank*(Alg. 3))**.** *Consider the decision space $\mathcal{D}$, such that $\nabla^2\psi(\mathbf{w}) \geq H_{\mathcal{D},\psi}^2 \mathbf{I},\ \forall \mathbf{w} \in \mathcal{D}$. Then for $\eta = \frac{\sqrt{\nu \min\{m,d\}} H_{\mathcal{D},\psi}}{d\sqrt{T \log T}}$, $\delta = \frac{1}{T}$ and $\gamma_t \leq \frac{0.7 H_{\mathcal{D},\psi}}{\ell_k}$, the Mirror-Rank (Alg. 1) guarantees a regret bound:* $\widehat{\mathrm{Reg}}_T^{\textit{Mirror-Rank}} \leq O\left(\frac{d\sqrt{\nu T \log T}}{\sqrt{\min\{m,d\}} H_{\mathcal{D},\psi}}\right)$.

The proof of Thm. 7 follows from the simple observation that upon receiving $\sigma_{m,t}$ at each round t, *Mirror-Rank* actually extracts m independent pairwise preference feedback from $\sigma_{m,t}$. This reduces the problem to our B-batched pairwise feedback setting with batchsize $B = m$. The result of Thm. 7 hence follows immediately from Thm. 4.

Following the argument of Rem. 2, the sample complexity bounds of *Mirror-Rank* (Alg. 3) is:

Corollary 8 (Sample Complexity Bound of *Mirror-Rank* (Alg. 3))**.** *Under the parameter settings of Thm. 7, the ϵ-sample complexity of Mirror-Rank (Alg. 3) is roughly* $O\left(\frac{d^2 \nu}{\min\{d,m\} H_{\mathcal{D},\psi}^2 \epsilon^2}\right)$.

Remark 5. *While querying larger subsets (k) intuitively provides more information per round, this advantage is exactly cancelled by the increased outcome variance — it becomes harder for the*

best item to prove its supremacy against $k-1$ competitors. Our result formally establishes this cancellation, yielding a k-independent guarantee, as justified by matching lower bound in Thm. 9.

Theorem 9 (Regret Lower Bound for Top-m Ranking `Rank-MNL` Problem)**.** *Consider any fixed time step T, subsetsize k and length of rank-ordered feedback $m \in \{1, \ldots, k\}$. Then for any algorithm $\mathcal{A}$ for the Top-m `Rank-MNL` problem, there exists a decision space $\mathcal{D} \subset \mathbb{R}^d$ and a sequence of unknown linear functionals $\boldsymbol{\theta}_1^*, \ldots \boldsymbol{\theta}_T^* \in \mathbb{R}^d$, such that the regret of algorithm $\mathcal{A}$ in T rounds $\widehat{\mathrm{Reg}}_T^{\texttt{Rank-MNL}} \geq \frac{d\sqrt{T \log T}}{256\sqrt{\min\{m,d\}}}$.*

6 Experiments

We run experiments to report the performance *Mirror-Duel* (Sec. 3) and *Mirror-Rank* (Sec. 5) for dueling and top-m ranking feedback. All results are averaged across 100 runs.

We run our experiments on different environments (adversarial loss sequences):

Adversarial $(\boldsymbol{\theta}_t^*)$ Environments. We report our results on problem instances with varying dimension d generated as follows: (1) **Inst**-$1(d)$: For a given dimension d and round t, we choose the adversarial sequence of score functions $\boldsymbol{\theta}_t^* \in \mathbb{R}^d$ such that for any co-ordinate $i \in [d]$, we first select $\tilde{\theta}_t(i) = \begin{cases} \mathrm{Unif}(0,1) - 0.37 & \text{if } i \leq \lceil d/2 \rceil \\ \mathrm{Unif}(0,1) + 0.53 & \text{otherwise} \end{cases}$, and set $\boldsymbol{\theta}_t^* = \tilde{\boldsymbol{\theta}}_t / \|\tilde{\boldsymbol{\theta}}_t\|$. (2) **Inst**-$2(d)$: For this instance, again given a fixed d and round t, we first choose $\tilde{\boldsymbol{\theta}}_t$ as: $\tilde{\theta}_t(i) = \mathrm{Unif}(0,1) + 0.57X$ for all $i \in [d]$, where $X \sim 2\mathrm{Ber}(0.5) - 1$ represents a $\{\pm 1\}$-valued binary random variable, and further set $\boldsymbol{\theta}_t^* = \tilde{\boldsymbol{\theta}}_t / \|\tilde{\boldsymbol{\theta}}_t\|$.

The decision space is given by $\mathcal{D} = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, for some $\mathbf{A} \in \mathbb{R}^{c \times d}$ and $\mathbf{b} \in \mathbb{R}^c$, for some $c \in \mathbb{N}_+$. For our purpose we select $\mathbf{A}$ to be $\mathbf{A} = [\mathbf{I}_d; -\mathbf{I}_d; -(\mathbf{1}_d)^\top; (\mathbf{1}_d)^\top]$ and $\mathbf{b} = [0.95\mathbf{1}_d; \mathbf{1}_d; 0.45d; 0.45d]$.

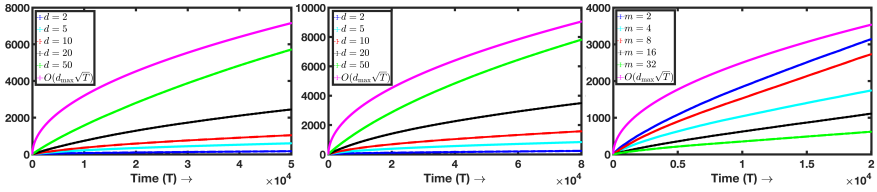


Figure 1: Avg-Regret of *Mirror-Duel* (Left, Center) & *Mirror-Rank* (Right) (More results in Sec. G)

Choice of the Self Concordant Barrier ψ: We use the following self concordant barrier $\psi(\mathbf{w}) = -\sum_{j=1}^{c} \ln(b_j - \mathbf{a}_j^\top \mathbf{w})$, for any $\mathbf{w} \in \mathbb{R}^d$, which is known to be an c-self-concordant barrier for $\mathcal{D}$, $\mathbf{a}_j$ being the j-th row of $\mathbf{A}$ and b_j denotes the j-th co-ordinate of $\mathbf{b}$ (Luo, 2017).

We now report a summary of our experiment results and a detailed description is deferred to Sec. G.

- Our first experiments are reported the regret performance of *Mirror-Duel* for the `Pref-DB` setting on Inst-1 and Inst-2. The results reported in Fig. 4 confirm its scalability and runtime efficiency, which are shown to scale as $O(d\sqrt{T})$ in Sec. G.1.
- We report the regret performance of *Mirror-Rank* on the `Rank-MNL` problem with varying rank-ordered feedback (on **Inst**-1) in Sec. G.2.
- Real-RL Environments: Finally, we compare *Mirror-Duel* against `APO` (Das et al., 2024), `DPO` (Rafailov et al., 2024), and `Nash-MD` (Munos et al., 2024) in a CartPole-v1 and Acrobot-v1 (Towers et al., 2024) preference-based policy search setting; results are deferred to Sec. G.3. Detailed descriptions of these baselines are given in Sec. G.4. [2]

7 Conclusion

A detailed conclusion, perspectives and discussion on future works is given in Sec. H.

[2]Note that other existing baselines discussed in Sec. 1 do not efficiently apply to continuous decision spaces. Existing approaches suggested by Di et al. (2024); Saha (2021b); Bengs et al. (2022) are computationally infeasible for continuous action spaces, and so experiments were conducted on only finite decision spaces. Kausik et al. (2024); Saha et al. (2023) do not report experiments. They were hence excluded from comparisons.

REFERENCES

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. *Conference on Learning Theory*, 2008.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. URL https://arxiv.org/abs/2310.12036.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pp. 1764–1786. PMLR, 2022.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006a.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006b.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34:3401–3412, 2021.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pp. 3773–3793. PMLR, 2022.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.03741.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably sample efficient rlhf via active preference optimization. *arXiv preprint arXiv:2402.10500*, 2024.

Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. *arXiv preprint arXiv:2404.10776*, 2024.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R Srikant. Exploration-driven policy optimization in rlhf: Theoretical insights on efficient data utilization. *arXiv preprint arXiv:2402.10342*, 2024.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7288–7295, 2021.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394. Society for Industrial and Applied Mathematics, 2005.

Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pp. 1704–1713, 2017.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Neural Information Processing Systems*, 2021.

Daniel Kahneman and Amos Tversky. The psychology of preferences. *Scientific American*, 246(1):160–173, 1982.

Sydney M. Katz, Amir Maleki, Erdem Bıyık, and Mykel J. Kochenderfer. Preference-based learning of reward function features. In *arXiv preprint arXiv:2103.02727*, 2021. URL https://arxiv.org/abs/2103.02727.

Chinmaya Kausik, Mirco Mutti, Aldo Pacchiano, and Ambuj Tewari. A framework for partially observed reward-states in rlhf. *arXiv preprint arXiv:2402.03282*, 2024.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. *arXiv preprint arXiv:1703.00048*, 2017.

Xuheng Li, Heyang Zhao, and Quanquan Gu. Feel-good thompson sampling for contextual dueling bandits. *arXiv preprint arXiv:2404.06013*, 2024.

Haipeng Luo. Lecture notes 17: Introduction to online learning. 2017.

Tengyu Ma. Lecture notes 15: Cs229t/stats231 statistical learning theory. 2018.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2312.00886.

Sam Musallam, BD Corneil, Bradley Greger, Hans Scherberger, and Richard A Andersen. Cognitive control signals for neural prosthetics. *Science*, 305(5681):258–262, 2004.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Dorsa Sadigh, Anca D. Dragan, Shankar Sastry, and Sanjit A. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems (RSS)*, 2017. doi: 10.15607/RSS.2017.XIII.053.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021a.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021b.

Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, 2019.

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 6263–6289. PMLR, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ohad Shamir. On the complexity of bandit linear optimization. In *Conference on Learning Theory*, pp. 1523–1551. PMLR, 2015.

Gokul Swamy, David Wu, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Inverse reinforcement learning without reinforcement learning. In *International Conference on Machine Learning*, pp. 33299–33318. PMLR, 2023.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Kronauer, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024. URL https://arxiv.org/abs/2407.17032.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is RLHF more difficult than standard RL? *arXiv preprint arXiv:2306.14111*, 2023.

Runzhe Wu and Wen Sun. Making rl with preference-based feedback efficient via randomization. *International Conference on Learning Representations*, 2024.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of Nash learning from human feedback under general KL-regularized preference. *arXiv preprint arXiv:2402.07314*, 2024.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Provable reward-agnostic preference-based reinforcement learning. *arXiv preprint arXiv:2305.18505*, 2023.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.