Dueling in the Dark: An Efficient and Optimal Mirror Descent Approach for Online Optimization with Adversarial Preferences

Aadirupa Saha, Apple AADIRUPA.SAHA@GMAIL.COM
Yonathan Efroni, Meta JONATHAN.EFRONI@GMAIL.COM
Barry-John Theobald, Apple BARRYJOHN_THEOBALD@APPLE.COM

Abstract

Recent developments in Large Language Models (LLMs) have sparked significant attention in Reinforcement Learning from Human Feedback (RLHF). A simple, widely used, and resource-efficient method for gathering human feedback is through relative queries based on human preferences, where the pairwise preference of two alternatives is often modeled as the sigmoid of their respective utility scores. Despite the popularity of these 'sigmoid-based RLHF' frameworks, their theoretical foundations remain underdeveloped as existing algorithms often lack the desired performance guarantees or are limited to small-scale problems due to computationally intractable steps. We address this challenge by developing the first efficient online gradient descent-based algorithm for the problem with provably optimal performance guarantees. In fact, our proposed methods work even for adversarially changing preferences, unlike the existing attempts which assume a fixed underlying stochastic preference model. Formally, we consider the adversarial online convex (linear) optimization (OLO) problem in d-dimension, but unlike the existing OLO framework, we only assume that the learner can only observe a (weaker) preference feedback upon choosing a few alternatives at each round. With the objective of identifying the 'best arm', we propose an efficient online mirror descent (OMD) based approach for the problem with regret and sample complexity guarantees. The main challenge lies in finding a suitable 'gradient approximation' of the underlying (adversarially changing) utility functions solely from the weak preference feedback, as opposed to the conventional gradient or value feedback used in OLO. We also extend our methods beyond pairwise preferences to multi-way preferences (B-sized batched pairwise) with improved performance guarantees. Additionally, our algorithms are optimal as we proved by matching lower bounds closing the potential of any better algorithms for the settings. Our contribution lays the groundwork for practical gradient descent-based algorithm in RLHF. Supported by robust theoretical guarantees, our approach holds promise in the current landscape of developing efficient algorithms for LLMs and addressing human-AI alignment challenges.

1. Introduction

The rapidly advancing field of AI has sparked 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 in humanity's best interests. This paper explores the theoretical aspects of RLHF with preference feedback, emphasizing its potential to enhance AI alignment.

Preference feedback, in particular, is a critical form of human feedback within the field of RLHF. Unlike the conventional feedback models used in the ML optimization literature for designing predictive AI models, which includes demonstration [11, 25, 26], gradient-based [2, 9, 29], value-based feedback [8, 10, 14], preference feedback is a much weaker form of feedback that only receives rel-

© A.S. Apple, Y.E. Meta & B.-J.T. Apple.

ative desirability (a.k.a. preference) of different outcomes/actions for a given task. However, on the positive end, preference feedback is a much more nuanced understanding of human values and priorities by explicitly capturing human judgments about the relative desirability of different outcomes. Studies in psychology and cognitive neuroscience also corroborate the fact that humans are often naturally more comfortable providing relative feedback compared to the other modes [12, 19], hence the training data tend to be less biased and cost-effective. Consequently, this form of feedback enables AI systems to learn more complex and subtle aspects of human preferences, which are often difficult to encode explicitly through demonstrations or reward feedback. For instance, in environments where trade-offs between multiple objectives are necessary, preference feedback can help an AI system align its actions more closely with 'human expected system behaviour' by understanding their tradeoff across different objectives from the relative choices, which otherwise would have been hard to gauge from. This helps the AI systems to become more interpretable, trustworthy, and 'human-aligned'!

Existing work on preference-based learning for AI alignment, whether empirical or theoretical, is limited by computationally inefficient algorithms. Many current approaches struggle to scale effectively with the complexity of real-world scenarios, often requiring extensive computational resources and time to process human feedback and update AI models accordingly. This inefficiency not only hampers the practical deployment of preference-based learning systems but also restricts their ability to quickly adapt to dynamic environments and evolving human preferences. Consequently, there is a pressing need for the development of more 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 gradient descent based algorithm that addresses the problem of regret minimization in RLHF, accessing only preference feedback, and providing an optimal guarantee with a bound of $O(\sqrt{T})$, where T is the time horizon of the game. Our main motivation in doing so is to deepen our understanding of the design of popular RLHF algorithms [5, 21**?**] which lack convergence guarantees. The gradient descent based approach we study in this work represents an advancement over existing attempts to provide provable guarantees for the RLHF or trajectory feedback settings [4, 6, 13, 22, 24, 28]. These works based their algorithms on the optimism in the face of uncertainty principle, and require optimizing over confidence sets. Such an optimization procedure is often non-convex and cannot be easily implemented by contemporary neural network architectures. Alternatively, several studies [7, 16, 27] considered a Thompson Sampling (TS) based algorithm, which requires sampling from a posterior distribution that adds additional complexity to the design of algorithm. Hence, previous efforts often fail to scale effectively and suffer from computational overhead, rendering them impractical for real-world applications. In contrast, our algorithm not only achieves optimal regret minimization but also does so in a computationally efficient way. This breakthrough is particularly important for the practical deployment of RLHF systems, enabling them to quickly and effectively incorporate preference feedback in a manner that is both scalable and resource-efficient. In this work we propose a solution that marries optimal theoretical performance with practical feasibility, setting a new benchmark for future research in the field.

Advantage of Gradient Descent Methods: Gradient-based methods have multiple advantages compared to confidence-based methods: (1) GD/OMD handle high-dimensional problems efficiently due to their reliance on gradient information: (2) They are suitable for both stochastic and adversarial environments, making the gradient-based methods robust to changing data distributions or the underlying loss/reward functions which is often more practical for modeling real-world problems, (3) These methods can optimize a wide range of objective functions, including non-linear, non-convex, and constrained problems, (4) Gradient descent algorithms are simple to implement, even seamlessly integrate with modern deep learning frameworks, making these methods computationally efficient, unlike many UCB and TS based methods which often do not have a closed form solution [6, 24] or sampling from the posteriors could be complicated [20], and (5) Gradient descent techniques are inherently robust to model misspecification and smoothly integrate with differential privacy techniques.

Contributions. Our contributions are twofold. At a high level, we address the problem of solving adversarial linear bandits with gradient descent and with preference feedback, more popularly studied as the RLHF with preference feedback setting. Our specific contributions are:

1. In Sec. 3, we design an online mirror-descent based algorithm to obtain an optimal $\tilde{O}(d\sqrt{T})$ algorithm for online optimization with adversarial preferences (Alg. 1,Thm. 1).

2. In Sec. 4, we generalized the above algorithm to multiwise (batched) preference feedback, where the learner can query a set of B pairwise preferences in one go. Our improved analysis of Alg. 2 shows that one can achieve a faster $\tilde{O}(\frac{d}{\sqrt{\min\{B,d\}}}\sqrt{T})$ regret learning rate for this case (Thm. 2).

We start our technical results by introducing the basic problem setup first.

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. 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. 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 $\Omega \subset \mathbb{R}^d$, int$(\Omega)$ denotes the interior of the set Ω. Ber(p) defines *Bernoulli* distribution with parameter $p \in [0, 1]$.

2.1. Problem: Adversarial Logistic Dueling Bandits (`Logit-DB`):

We consider a decision space $\mathcal{D} \subset \mathbb{R}^d$ and the finite T horizon adversarial linear optimization setting. At every round, the algorithm plays $\mathbf{x}_t, \mathbf{y}_t \in \mathcal{D}$ and observes a binary o_t s.t.

$$o_t \sim \text{Ber}\Big(\sigma\big({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}_t - \mathbf{y}_t)\big)\Big).$$

We denote the probability of arm $\mathbf{x}$ being preferred over arm $\mathbf{y}$ as:

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

Note we call the problem `Logit-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: Regret Minimization w.r.t. the Best Choice. The goal of the algorithm is to minimize the cumulative regret, defined as:

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

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 $x \in \mathcal{D}$,*

$$\frac{\boldsymbol{\theta}_t^{*\top}(\mathbf{x}^* - \mathbf{x})}{4} \leq P_t(\mathbf{x}^*, \mathbf{x}) - 1/2 \leq \boldsymbol{\theta}_t^{*\top}(\mathbf{x}^* - \mathbf{x})$$

when $\mathcal{D} \subseteq \mathcal{B}_d(1)$. We prove this in App. A. Consequently, in the rest of the paper, we will address the regret

$$\widehat{\text{Reg}}_T^{\textit{Logit-DB}} := \sum_{t=1}^{T} \left[\frac{\boldsymbol{\theta}_t^{*\top}(\mathbf{x}^* - \mathbf{x}_t) + \boldsymbol{\theta}_t^{*\top}(\mathbf{x}^* - \mathbf{y}_t)}{2} \right],$$

noting $\text{Reg}_T^{\textit{Logit-DB}} \leq \widehat{\text{Reg}}_T^{\textit{Logit-DB}}$ from Rem. 1, thus designing algorithm to bound $\widehat{\text{Reg}}_T^{\textit{Logit-DB}}$ would suffice to bound $\text{Reg}_T^{\textit{Logit-DB}}$.

3. Dueling Case: Algorithm for `Logit-DB` Problem

In this section, we investigate the `Logit-DB` problem (Sec. 2.1) for the pairwise preference (dueling) feedback.

Algorithm description: Our algorithm is motivated by the Scrible algorithm from [1, 10], which is a variant of the online mirror descent algorithm with a self-concordant barrier [2] 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. 3. 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 analyze the regret performance of Alg. 1 yielding an optimal $O(\sqrt{T})$ regret for the problem, as justified in Rem. 3. The detailed regret analysis of *Double-Scrible* (Alg. 1) is given in App. B.2.

[1] Interested readers may check [2, 10, 17] for the properties and examples of self-concordant barrier functions.

Algorithm 1 *Double-Scrible*

```
Input: Decision set $\mathcal{D}$ with $\nu$-self concordant barrier $\psi$, parameters $\eta, \delta, \gamma_t$.
for $t = 1$ to $T$ do
    Compute: $\mathbf{w}_t = \arg\min_{\mathbf{w}\in\mathcal{D}_\delta} \left\{\eta \sum_{\tau=1}^{t-1}(-\mathbf{g}_\tau)^\top \mathbf{w} + \psi(\mathbf{w})\right\}$.
    Compute eigendecomposition s.t. $\nabla^2\psi(\mathbf{w}_t) = \sum_{j=1}^{d} \lambda_{t,j}\mathbf{v}_{t,j}\mathbf{v}_{t,j}^\top$.
    Sample $i_t \in [d]$ uniformly at random
    Choose $\mathbf{x}_t = \mathbf{w}_t + \gamma_t \frac{1}{2\sqrt{\lambda_{t,i_t}}}\mathbf{v}_{t,i_t}$ and $\mathbf{y}_t = \mathbf{w}_t - \gamma_t \frac{1}{2\sqrt{\lambda_{t,i_t}}}\mathbf{v}_{t,i_t}$.
    Play $(\mathbf{x}_t, \mathbf{y}_t)$, observe $o_t \sim \text{Ber}(P_t(\mathbf{x}_t, \mathbf{y}_t))$.
    $\mathbf{g}_t = \frac{d}{\gamma_t}\left(o_t - \frac{1}{2}\right)\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}$.
end for
```

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 Double-Scrible (Alg. 1) guarantees a regret bound:*

$$\widehat{\text{Reg}}_T^{\textit{Logit-DB}} \leq O\left(\frac{d\sqrt{\nu T\log T}}{H_{\mathcal{D},\psi}}\right).$$

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. 5.

Remark 2 (Minimal Eigenvalue Assumption) *Thm. 1 holds assuming the minimal eigenvalue of $\nabla^2\psi(\mathbf{w})$ is larger than $H^2_{\mathcal{D},\psi}$. This assumption was not required in the analysis of Scrible [1], however, we could not circumvent it. The reason we are required to make this assumption lies in the fact the reward model we optimize is a non-linear model, whereas the reward model in [1] is linear in w. This assumption is equivalent for 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 `Logit-DB` problem ([23]).*

Remark 4 (Advantage of Our Approach over Existing Algorithms for `Logit-DB`) *(1) Prior works that considered online learning in the generalized linear bandit setting [15, 16] are required to assume a lower bound on the derivative of the sigmoid link function, which results in 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 the `Logit-DB` problem [6, 24]. Interestingly, we do not need to make this assumption, owing to the nice trick of exploiting the pairwise preference of symmetrically opposite points $\mathbf{x}_t$ and $\mathbf{y}_t$, as shown in Lem. 4. This is a clean advantage of our approach over the existing GLM-bandits based approach for `Logit-DB` which relies on UCB estimation based confidence bounding technique. (2) Further, since our approach relies on gradient based techniques, they are extremely computationally efficient–the runtime requirement of our method is just $O(dT)$, compared to the prior methods which are computationally infeasible and not implementable in practice [6, 13, 24].*

Due to page limitations, the complete proof is moved to App. B.

4. Batched Feedback: Algorithm for B-Batched `Logit-DB` Problem

In this section, we will analyze a variant of `Logit-DB` problem where the learner actively gets to query B-pairwise queries in a batched fashion. More precisely, at each round t, the learner gets to query 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 $\widehat{\text{Reg}}_T^{\texttt{Batched-LogitDB}}$, the objective of the learner, in this case, is to minimize the regret over T rounds, defined as:

$$\widehat{\text{Reg}}_T^{\text{Batched-LogitDB}} := \sum_{t=1}^{T} \frac{1}{B}\left[\sum_{i=1}^{B} \frac{{\boldsymbol{\theta}_t^*}^\top(\mathbf{x}_t^* - \mathbf{x}_t^i) + {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}_t^* - \mathbf{y}_t^i)}{2}\right].$$

4.1. Algorithm for `Batched-LogitDB`

Our proposed algorithm for this case is *Batched-DouBle-Scrible* (*BaBle-Scrible*) which is a variant of *Double-Scrible* we detailed in the previous section for the `Logit-DB`. Same as algorithm, it 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 B-factor improvement in the final learning rate of $O(\frac{d}{B}\sqrt{T})$. However, one would need $B \le d$ 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 [29].

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 *BaBle-Scrible* 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 $\mathcal{D}_\delta$. The algorithm pseudocode is given in Alg. 2.

Thm. 2 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. The regret analysis of *BaBle-Scrible* (Alg. 2) is given in App. C.1.

Theorem 2 (Regret Analysis of Alg. 2) *Consider the decision space $\mathcal{D}$, such that $\nabla^2\psi(\mathbf{w}) \ge H_{\mathcal{D},\psi}^2\mathbf{I}$, $\forall \mathbf{w} \in \mathcal{D}$. 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 $\gamma_t \le 0.7H_{\mathcal{D},\psi}$, the BaBle-Scrible (Alg. 1) guarantees a regret bound:*

$$\widehat{\text{Reg}}_T^{\textit{Logit-DB}} \le O\left(\frac{d\sqrt{\nu T\log T}}{\sqrt{\min\{B,d\}}H_{\mathcal{D},\psi}}\right).$$

Due to page limitations, the complete proof is moved to App. C.

Algorithm 2 *BaBle-Scrible*

```
Input: Decision set $\mathcal{D}$ with $\nu$-self concordant barrier $\psi$, parameters $\eta, \delta, \gamma_t$.
for $t = 1$ to $T$ do
    Compute: $\mathbf{w}_t = \arg\min_{\mathbf{w} \in \mathcal{D}_\delta} \left\{ \eta \sum_{\tau=1}^{t-1} (-\mathbf{g}_\tau)^\top \mathbf{w} + \psi(\mathbf{w}) \right\}$.
    Compute eigendecomposition s.t. $\nabla^2 \psi(\mathbf{w}_t) = \sum_{i=1}^{d} \lambda_{t,i} \mathbf{v}_{t,i} \mathbf{v}_{t,i}^\top$.
    for $\ell = 1, 2, \ldots, B$ do
        Sample $i_t^\ell \in [d]$ uniformly at random.
        Choose $\mathbf{x}_t^\ell = \mathbf{w}_t + \gamma_t \frac{1}{2\sqrt{\lambda_{t,i_t^\ell}}} \mathbf{v}_{t,i_t^\ell}$ and $\mathbf{y}_t^\ell = \mathbf{w}_t - \gamma_t \frac{1}{2\sqrt{\lambda_{t,i_t^\ell}}} \mathbf{v}_{t,i_t^\ell}$.
        Play $(\mathbf{x}_t^\ell, \mathbf{y}_t^\ell)$, observe $o_t^\ell \sim \text{Ber}(P_t(\mathbf{x}_t^\ell, \mathbf{y}_t^\ell))$.
        $\mathbf{g}_t^\ell = \frac{d}{\gamma_t} \left(o_t^\ell - \frac{1}{2}\right) \sqrt{\lambda_{t,i_t^\ell}} \mathbf{v}_{t,i_t^\ell}$.
    end for
    Update $\mathbf{g}_t = \frac{1}{B} \sum_{\ell=1}^{B} \mathbf{g}_t^\ell$
end for
```

5. Conclusion

In this paper, we introduced an efficient gradient descent-based approach for regret minimization for online linear optimization with adversarial preferences. Our results has critical implications in learning problems of RLHF which has wide applications in fields of AI-alignment, fine tuning language models, etc. Our proposed novel online mirror descent (OMD) algorithm achieves an optimal regret bound of $O(\sqrt{T})$ while only relying on binary preference feedback. This advancement improves upon existing methods by addressing key computational challenges, particularly in handling high-dimensional and adversarial environments while still respecting optimal performance guarantees. We also extended our algorithm to accommodate batched preference feedback, which yields provably faster performance guarantees for larger batch size. The computational efficiency of our algorithms makes them suitable for large-scale real-world applications.

Future Work. Building on this work, several promising avenues for future exploration emerge: One potential extension is to generalize the setting beyond linear scores which is certainly not straightforward even for value-feedback based convex optimization setting []. Extending to partially observable preferences or partial ranking feedback over a subset of alternatives is also an interesting open problem. Another direction is to explore hybrid approaches that combine gradient descent with other optimization techniques like Thompson sampling or Bayesian methods, to reduce variance in feedback-based learning. Finally, investigating how this algorithm can be adapted for different AI alignment challenges, such as incorporating fairness or ethical constraints in decision-making, presents an exciting opportunity for future research.

References

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

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

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

[4] 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.

[5] 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.

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

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

[8] 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*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

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

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

[11] 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.

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

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

[14] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

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

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

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

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

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

[20] 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*, pages 1029–1038. PMLR, 2020.

[21] 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.

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

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

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

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

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

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

[28] 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.

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

Supplementary: Dueling in the Dark: An Efficient and Optimal Mirror Descent Approach for Online Optimization with Adversarial Preferences

Appendix A. Appendix for Sec. 2

Remark 1 *For any $x \in \mathcal{D}$,*

$$\frac{{\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})}{4} \le P_t(\mathbf{x}^*, \mathbf{x}) - 1/2 \le {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})$$

when $\mathcal{D} \subseteq \mathcal{B}_d(1)$. We prove this in App. A. Consequently, in the rest of the paper, we will address the regret

$$\widehat{\text{Reg}}_T^{\textit{Logit-DB}} := \sum_{t=1}^{T} \bigg[\frac{{\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x}_t) + {\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{y}_t)}{2} \bigg],$$

noting $\text{Reg}_T^{\textit{Logit-DB}} \le \widehat{\text{Reg}}_T^{\textit{Logit-DB}}$ from Rem. 1, thus designing algorithm to bound $\widehat{\text{Reg}}_T^{\textit{Logit-DB}}$ would suffice to bound $\text{Reg}_T^{\textit{Logit-DB}}$.

Proof [Proof of Rem. 1] Let us fix a round t, and for simplicity denote $\mathbf{x}^* = \mathbf{x}_t^*$ (dropping the subscript). Note that due to the underlying preference structure for any $\mathbf{x} \in \mathcal{D}$,

$$\begin{aligned} P_t(\mathbf{x}^*, \mathbf{x}) - 1/2 &= \sigma({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) - 1/2 = \frac{\exp({\boldsymbol{\theta}_t^*}^\top\mathbf{x}^*)}{\exp({\boldsymbol{\theta}_t^*}^\top\mathbf{x}^*) + \exp({\boldsymbol{\theta}_t^*}^\top\mathbf{x})} - 1/2 \\ &= \frac{\big(\exp({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) - 1\big)}{2\big(\exp({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) + 1\big)} \overset{(a)}{\ge} \frac{\big(\exp({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) - 1\big)}{4} \\ &= \frac{1}{4}\bigg(1 + \sum_{i=1}^{\infty} \frac{({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x}))^i}{i!} - 1\bigg) > \frac{{\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})}{4} \end{aligned}$$

where (a) follows since ${\boldsymbol{\theta}_t^*}^\top\mathbf{x} \in [0, 1]$, $\forall \mathbf{x} \in \mathcal{D}$, assuming $\boldsymbol{\theta}_t^* \in \mathcal{B}_d(1)$ and $\mathcal{D} \subseteq \mathcal{B}_d(1)$. On the other hand,

$$\begin{aligned} P_t(\mathbf{x}^*, \mathbf{x}) - 1/2 &= \sigma({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) - 1/2 = \frac{\exp({\boldsymbol{\theta}_t^*}^\top\mathbf{x}^*)}{\exp({\boldsymbol{\theta}_t^*}^\top\mathbf{x}^*) + \exp({\boldsymbol{\theta}_t^*}^\top\mathbf{x})} - 1/2 \\ &= \frac{\big(\exp({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) - 1\big)}{2\big(\exp({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) + 1\big)} \overset{(b)}{\le} \frac{\big(\exp({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x})) - 1\big)}{2} \\ &= \frac{1}{2}\bigg(1 + \sum_{i=1}^{\infty} \frac{({\boldsymbol{\theta}_t^*}^\top(\mathbf{x}^* - \mathbf{x}))^i}{i!} - 1\bigg). \end{aligned}$$

Now let us denote $a = \boldsymbol{\theta}_t^{*\top}(\mathbf{x}^* - \mathbf{x})$ and note by assumption $0 < a \leq 2$.

$$P_t(\mathbf{x}^*, \mathbf{x}) - 1/2 = \frac{1}{2}\bigg(\sum_{i=1}^{\infty} \frac{a^i}{i!}\bigg) \leq \frac{a}{2} + \frac{a^2}{2}\bigg(1 + \frac{a}{2} + \frac{a^2}{2^2} + \dots\bigg)$$
$$= \frac{a}{2} + \frac{a^2}{2}\frac{2}{(2-a)} \leq \frac{a}{2} + \frac{a}{2} = a = \boldsymbol{\theta}_t^{*\top}(\mathbf{x}^* - \mathbf{x}),$$

where the second the last inequality holds since $a \in (0,2)$, assuming $\boldsymbol{\theta}_t^* \in \mathcal{B}_d(1)$ and $\mathcal{D} \subseteq \mathcal{B}_d(1)$.
■

Appendix B. Appendix for App. B.2

B.1. Key Lemmas for Thm. 1 (Regret Analysis of Alg. 1)

We define the useful notations which will be useful for stating the claims:
Notations: We denote the history $\mathcal{H}_t := \{(i_1, o_1), (i_2, o_2), \dots (i_{t-1}, o_{t-1})\}$ till time t. We define a norm associated with the Hessian of ψ at $\mathbf{w}$ as $\|\mathbf{x}\|_{\mathbf{w}} = \|\mathbf{x}\|_{\nabla^2\psi(\mathbf{w})} = \sqrt{\mathbf{x}^\top \nabla^2\psi(\mathbf{w})\mathbf{x}}$ for any $\mathbf{x} \in \mathbb{R}^d$. This is indeed a norm since a self-concordant barrier is strictly convex, such that $\nabla^2\psi(\mathbf{w})$ is positive definite for any $\mathbf{w} \in \text{int}(\mathcal{D})$.
Further considering the eigen-decomposition of $\nabla^2\psi(\mathbf{w}) = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, we further note that

$$\|\mathbf{x}\|_{\mathbf{w}} = \sqrt{\mathbf{x}^\top \nabla\psi(\mathbf{w})\mathbf{x}} = \sqrt{\sum_{i=1}^d \lambda_{t,i}\mathbf{x}^\top(\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top)\mathbf{x}} = \sqrt{\sum_{i=1}^d \lambda_i(\mathbf{x}^\top \mathbf{v}_{t,i})^2},\ \forall \mathbf{x} \in \mathbb{R}^d.$$

Further, one can define the dual norm of Hessian of ψ at $\mathbf{w}$ as:

$$\|\mathbf{x}\|_{\mathbf{w}}^* = \sqrt{\mathbf{x}^\top \nabla^{-2}\psi(\mathbf{w})\mathbf{x}} = \sqrt{\sum_{i=1}^d \frac{1}{\lambda_{t,i}}\mathbf{x}^\top(\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top)\mathbf{x}},\ \forall \mathbf{x} \in \mathbb{R}^d.$$

The Dikin ellipsoid centered at $\mathbf{w}$ with radius r is defined as the ellipsoid

$$\mathcal{E}_r(\mathbf{w}) = \Big\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{w}\|_{\mathbf{w}} \leq r\Big\}.$$

Property 1 ([2, 17]) *If ψ is a self-concordant barrier on $\mathcal{D}$, then $\mathcal{E}_1(\mathbf{w}) \subset \mathcal{D}$ for any $\mathbf{w} \in$ int$(\mathcal{D})$.*

Property 2 ([17]) *Let $\mathbf{x} \in$ int$(\mathcal{D})$ be such that $\|\nabla\Phi(\mathbf{x})\|_{\mathbf{x}}^* \leq \frac{1}{4}$, and let $\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathcal{D}} \Phi(\mathbf{x})$. Then for any $\Phi : \mathcal{D} \mapsto \mathbb{R}$,*

$$\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{x}} \leq 2\|\nabla\Phi(\mathbf{x})\|_{\mathbf{x}}^*.$$

Property 3 ([10, 17]) *Let ψ be a ν-self concordant function over $\mathcal{D}$, then for all $\mathbf{x}, \mathbf{y} \in$ int$(\mathcal{D})$:*

$$\psi(\mathbf{y}) - \psi(\mathbf{x}) \leq \nu \log \frac{1}{1 - \pi_{\mathbf{x}}(\mathbf{y})},$$

where $\pi_{\mathbf{x}}(\mathbf{y}) = \inf\{t \geq 0 : \mathbf{x} + t^{-1}(\mathbf{y} - \mathbf{x}) \in \mathcal{D}\}$.

The proof sketch of Thm. 1 depends on some key lemmas. First we claim that g_t given an 'almost' unbiased estimate of $\boldsymbol{\theta}_t^*$ up to some constants.

Lemma 3 (Ensuring Decision Boundaries) *At any round t, $\mathbf{x}_t$ and $\mathbf{y}_t \in \mathcal{D}$ in Alg. 1.*

Proof We will prove the result for $\mathbf{x}_t$. A similar analysis will apply to $\mathbf{y}_t$ as well. Note since $\mathbf{w}_t \in \text{int}(\mathcal{D})$, and $\|\mathbf{x}_t - \mathbf{w}_t\|_{\mathbf{x}} \leq \gamma_t \leq 1$. Note Rem. 5 ensures $\gamma_t \leq 1$ and thus the results follows using Property 1. ■

Lemma 4 (Gradient Estimation) *It can be shown that for any round t,*

$$\mathbf{E}[\mathbf{g}_t \mid \mathcal{H}_t] = C\boldsymbol{\theta}_t^*,$$

for some $C \in [0.22, 0.25]$, whenever $\gamma_t \leq 0.7\sqrt{\lambda_{\min}(\nabla^2\psi(\mathbf{w}_t))}$.

Remark 5 (Ensuring appropriate choice of γ_t) *Noting that, given the decision space $\mathcal{D}$, since $\nabla^2\psi(\mathbf{w}_t) \geq H_{\mathcal{D},\psi}^2\mathbf{I}_d$, one can easily satisfy $\gamma_t \leq 0.7\sqrt{\lambda_{\min}(\nabla^2\psi(\mathbf{w}_t))}$ by choosing $\gamma_t = \min\{1, 0.7H_{\mathcal{D},\psi}\}$. We have given some specific examples in Rem. 2.*

Proof Consider any fixed round $t \in [T]$. We note that:

$$\begin{aligned}\mathbf{E}_{o_t}[(o_t - 1/2) \mid i_t, \mathcal{H}_t] &= \mathbf{E}_{o_t}[o_t \mid i_t, \mathcal{H}_t] - 1/2 = \sigma\big(\boldsymbol{\theta}_t^{*\top}(\mathbf{x}_t - \mathbf{y}_t)\big) - 1/2\\ &= \sigma\big(({}^{2\gamma_t}/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}\big) - 1/2\\ &= \sigma'(\varepsilon_t)({}^{\gamma_t}/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t} \quad \text{[using MVT, where } |\varepsilon_t| \in [0, |({}^{\gamma_t}/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}|]. \end{aligned} \tag{1}$$

Let us denote $c_t = |({}^{\gamma_t}/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}|$ and note that we can bound $c_t \leq \frac{\gamma_t}{\sqrt{\lambda_{t,i_t}}}\|\boldsymbol{\theta}_t^*\|\|\mathbf{v}_{t,i_t}\| \leq \frac{\gamma_t}{\sqrt{\lambda_{\min}(\nabla^2\psi(\mathbf{w}_t))}}$, where the first inequality follows from the Cauchy-Schwarz inequality.
Then by choosing any $\gamma_t \leq 0.7\sqrt{\lambda_{\min}(\nabla^2\psi(\mathbf{w}_t))}$ we get $c_t \leq 0.7$. This along with the results of Lem. 8 (App. D) implies that $\sigma'(\epsilon_t) \leq [0.222, 0.25]$ for the appropriate choice of γ_t. Note Rem. 5 explains the suitable choice of γ_t For simplicity, we will use $L = 0.222$, $U = 0.25$ for the rest of this proof and let $\sigma'(\varepsilon_t) \in [L, U]$. The interesting thing now is to note that, given the history $\mathcal{H}_t$ till time t, $\mathbf{g}_t$ in Alg. 1 satisfies:

$$\begin{aligned}&\mathbf{E}_{o_t,i_t}[g_t \mid \mathcal{H}_t] = \mathbf{E}_{i_t,\omega_t}\left[\frac{d}{\gamma_t}\mathbf{E}_{o_t}\left[\big(o_t - \frac{1}{2}\big) \mid i_t, \mathcal{H}_t\right]\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}\right]\\ &= \mathbf{E}_{i_t}\left[\frac{d}{\gamma_t}\left(\sigma'(\varepsilon_t)({}^{\gamma_t}/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}\right)\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}\right] \quad \text{using Eq. (1)}\\ &\in \left[L\mathbf{E}_{i_t}\left[\frac{d}{\gamma_t}\left(({}^{\gamma_t}/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}\right)\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}\right], U\mathbf{E}_{i_t}\left[\frac{d}{\gamma_t}\left(({}^{\gamma_t}/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}\right)\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}\right]\right]\\ &\in [L\boldsymbol{\theta}_t^*, U\boldsymbol{\theta}_t^*],\end{aligned}$$

where the last inequality follows noting:

$$\begin{aligned}\mathbf{E}_{i_t}\left[\frac{d}{\gamma_t}\Big((\gamma_t/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}\Big)\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}\right] &= \mathbf{E}_{i_t}\left[d\Big((1/\sqrt{\lambda_{t,i_t}})\boldsymbol{\theta}_t^{*\top}\mathbf{v}_{t,i_t}\Big)\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}\right]\\ &= d\bigg(\sum_{i=1}^{d}\frac{1}{d\sqrt{\lambda_{t,i}}}\sqrt{\lambda_{t,i}}\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top\bigg)\boldsymbol{\theta}_t^* = \boldsymbol{\theta}_t^*,\end{aligned}$$

since $\sum_i \mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top = \mathbf{I}_d$ by the fact that $\{\mathbf{v}_{t,i}\}_{i\in[d]}$ are orthonormal vectors that span $\mathbb{R}^d$. ■

Equipped with the previous results, we are now ready to proof our main theorem, Thm. 1, as shown below.

B.2. Regret Analysis: Proof of Thm. 1

Suppose *Be The Leader* (BTL) algorithm [3, 18] is run on the loss vector sequence $-\mathbf{g}_1, -\mathbf{g}_2, \ldots, -\mathbf{g}_T), \ \mathbf{g}_i \in \mathbb{R}^d$. We know that for any $\mathbf{u} \in \mathcal{D}_\delta$:

$$\sum_{t=1}^{T}\big(\mathbf{w}_t - \mathbf{u}\big)^\top(-\mathbf{g}_t) \le \sum_{t=1}^{T}\big(\mathbf{w}_t - \mathbf{w}_{t+1}\big)^\top(-\mathbf{g}_t) + \frac{(\psi(\mathbf{u}) - \psi(\mathbf{w}_1))}{\eta}.$$

Further applying Holder's inequality, we get:

$$\sum_{t=1}^{T}\big(\mathbf{u} - \mathbf{w}_t\big)^\top\mathbf{g}_t \le \sum_{t=1}^{T}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{\mathbf{w}_t}\|{-\mathbf{g}_t}\|_{\mathbf{w}_t}^* + \frac{(\psi(\mathbf{u}) - \psi(\mathbf{w}_1))}{\eta}. \tag{2}$$

Note we defined: $\mathbf{g}_t = \frac{d}{\gamma_t}\big(o_t - \frac{1}{2}\big)\sqrt{\lambda_{t,i_t}}\mathbf{v}_{t,i_t}$ and by Lem. 4, we have

$$0.22\boldsymbol{\theta}_t^* \le \mathbf{E}[g_t \mid \mathcal{H}_t] \le 0.25\boldsymbol{\theta}_t^*,$$

which implies :

$$\boldsymbol{\theta}_t^{*\top}\big(\mathbf{u} - \mathbf{w}_t\big) \le \frac{1}{0.22}\mathbf{E}[g_t^\top \mid \mathcal{H}_t]\big(\mathbf{u} - \mathbf{w}_t\big), \tag{3}$$

combining this with Eq. (2), we get:

$$0.22\boldsymbol{\theta}_t^{*\top}\big(\mathbf{u} - \mathbf{w}_t\big) \le \sum_{t=1}^{T}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{\mathbf{w}_t}\|\mathbf{g}_t\|_{\mathbf{w}_t}^* + \frac{(\psi(\mathbf{u}) - \psi(\mathbf{w}_1))}{\eta}. \tag{4}$$

On the other hand, by definition of $\|\cdot\|_{\mathbf{w}_t}^*$, we have that for any realization of $\mathbf{g}_t$:

$$\|\mathbf{g}_t\|_{\mathbf{w}_t}^* = \sqrt{\sum_{i=1}^{d}\frac{1}{\lambda_{i,t}}\mathbf{g}_t^\top(\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top)\mathbf{g}_t} = \frac{d}{2\gamma_t}. \tag{5}$$

Additionally, let us denote by $\Phi_t(\mathbf{w}) = \eta\sum_{\tau=1}^{t}(-\mathbf{g}_\tau)^\top\mathbf{w} + \psi(\mathbf{w})$, then note Alg. 1 have $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{D}_\delta}\Phi_t(x)$. Thus, applying Property 2, we get:

$$\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{\mathbf{w}_t} \le 2\|\nabla\Phi_t(\mathbf{w}_t)\|^*_{\mathbf{w}_t} = 2\|\nabla\Phi_{t-1}(\mathbf{w}_t) + \eta\mathbf{g}_t\|^*_{\mathbf{w}_t} = 2\eta\|\mathbf{g}_t\|^*_{\mathbf{w}_t},$$

where note by definition $\nabla\Phi_{t-1}(\mathbf{w}_t) = 0$ by definition of $\mathbf{w}_t$ for all t. But note for Property 2 to be applied we need $\|\nabla\Phi_t(\mathbf{w}_t)\|^*_t \le \frac{1}{4}$, but this is indeed true since since Eq. (7) implies

$$\eta\|\mathbf{g}_t\|^*_{\mathbf{w}_t} \le \frac{\eta d}{2\gamma_t},$$

and thus choosing any $\eta \le \frac{\gamma_t}{2d}$, we have $\|\nabla\Phi_t(\mathbf{w}_t)\|^*_t \le \frac{1}{4}$, as desired. We will see shortly how to choose η to ensure $\eta \le \frac{\gamma_t}{2d}$.
Further using Property 2, we have $\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{\mathbf{w}_t} \le 2\eta\|\mathbf{g}_t\|^*_{\mathbf{w}_t}$, which along with Eq. (4) we get:

$$\begin{aligned}0.22\sum_{t=1}^{T}\big(\mathbf{u}-\mathbf{w}_t\big)^\top\boldsymbol{\theta}^*_t &\le \sum_{t=1}^{T}2\eta\|\mathbf{g}_t\|^{*2}_{\mathbf{w}_t} + \frac{(\psi(\mathbf{u})-\psi(\mathbf{w}_1))}{\eta}\\ &= 2\eta\sum_{t=1}^{T}\frac{d^2}{4\gamma_t^2} + \frac{\nu\log\frac{1}{1-\pi_{\mathbf{w}_1}(\mathbf{u})}}{\eta}.\end{aligned}$$

However noting $\mathbf{u}$ and $\mathbf{w}_1 \in \mathcal{D}_\delta$, by definition of $\pi_{\mathbf{w}_1}(u) = (1-\delta)$ in Property 3, implying:

$$0.22\sum_{t=1}^{T}\big(\mathbf{u}-\mathbf{w}_t\big)^\top\boldsymbol{\theta}^*_t \le \eta\sum_{t=1}^{T}\frac{d^2}{2\gamma_t^2} + \frac{\nu\log\frac{1}{\delta}}{\eta}. \tag{6}$$

Further if we choose $\mathbf{u} := \arg\max_{\mathbf{x}\in\mathcal{D}_\delta}\sum_{t=1}^{T}{\boldsymbol{\theta}^*_t}^\top\mathbf{x}$, and recalling that we defined $\mathbf{x}^* := \arg\max_{\mathbf{x}\in\mathcal{D}}\sum_{t=1}^{T}{\boldsymbol{\theta}^*_t}^\top\mathbf{x}$, note that:

$$\begin{aligned}\sum_{t=1}^{T}\big(\mathbf{x}^*-\mathbf{w}_t\big)^\top\boldsymbol{\theta}^*_t &\le \sum_{t=1}^{T}\big(u-\mathbf{w}_t\big)^\top\boldsymbol{\theta}^*_t + \delta TLD\\ &= \frac{1}{0.22}\left[\eta\sum_{t=1}^{T}\frac{d^2}{2\gamma_t^2} + \frac{\nu\log\frac{1}{\delta}}{\eta}\right] + \delta TLD, && \text{from Eq. (6)}\\ &\le \frac{1}{0.22}\left[\frac{\eta d^2 T}{\min\{1, H^2_{\mathcal{D},\psi}\}} + \frac{\nu\log\frac{1}{\delta}}{\eta}\right] + \delta TLD, && \text{since we chose } \gamma \le \min\{1, 0.7H_{\mathcal{D},\psi}\}\\ &= \frac{d\sqrt{\nu T\log T}}{0.22H_{\mathcal{D},\psi}} + LD,\end{aligned}$$

choosing $\eta = \frac{\sqrt{\nu}H_{\mathcal{D},\psi}}{d\sqrt{T\log T}}$ and $\delta = \frac{1}{T}$, concludes the prove noting the diameter of the decision set $\mathcal{D} \subseteq \mathcal{B}_d(1)$ is bounded by 1, and the lipschitz constant $L \le \max_{t\in[T]}\|\boldsymbol{\theta}^*_t\| \le 1$.

Appendix C. Appendix for App. C.1

C.1. Regret Analysis of Alg. 2

We will need to prove some key lemmas before proceeding to the proof of the main theorem $Thm.$ 2.

Lemma 5 (Gradient $(\boldsymbol{\theta}_t^*)$ Estimation) *It can be shown that for any round t,*

$$\mathbf{E}[g_t \mid \mathcal{H}_t] = C\boldsymbol{\theta}_t^*,$$

for some $C \in [0.22, 0.25]$, whenever $\gamma_t \le 0.7\sqrt{\lambda_{\min}(\nabla^2\psi(\mathbf{w}_t))}$.

Proof [Proof of Lem. 5] Let us fix any $t \in [T]$. Recall that we defined $\mathbf{g}_t^\ell = \frac{d}{\gamma_t}\big(o_t^\ell - \frac{1}{2}\big)\sqrt{\lambda_{t,i_t^\ell}}\mathbf{v}_{t,i_t^\ell}$ and for any $\ell = 1, 2, \ldots, B$, Now noting since $i_t^\ell \sim \text{Unif}([d])$, following the notations and exact same proof of Lem. 4, we get that: for any $\ell \in [B]$, $\mathbf{E}[g_t^\ell \mid \mathcal{H}_t] = C\boldsymbol{\theta}_t^*$, for some $C \in [0.22, 0.25]$. The proof now follows noting $\mathbf{g}_t := \frac{1}{B}\sum_{\ell=1}^B \mathbf{g}_t^\ell$. ■

We next prove the most important claim of this analysis that shows that indeed the batched feedback helped to obtain a more accurate (reduced variance) estimate of the gradient $\boldsymbol{\theta}_t^*$ at each time step t. The proof involves a smart exploitation of the second moment of Binomial distribution, we will see in the proof of Lem. 6.

Lemma 6 (Improved Variance of $\mathbf{g}_t$ (Norm bound)) *At any time t, one can show that*

$$\mathbf{E}_{i_t^\ell, o_t^\ell}[\|\mathbf{g}_t\|_{\mathbf{w}_t}^*] \le \frac{d}{\gamma_t\sqrt{\{B, d\}}}. \tag{7}$$

Proof [Proof of Lem. 6] We start by recalling that we defined the dual norm of Hessian of ψ at $\mathbf{w}$ as

$$\|\mathbf{x}\|_{\mathbf{w}}^* = \sqrt{\mathbf{x}^\top\nabla^{-2}\psi(\mathbf{w})\mathbf{x}} = \sqrt{\sum_{i=1}^d \frac{1}{\lambda_{t,i}}\mathbf{x}^\top(\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top)\mathbf{x}}, \;\; \forall \mathbf{x} \in \mathbb{R}^d.$$

At any round t, let us now denote by $N_{t,i}$ the number of times the i-th eigen basis, $\mathbf{v}_{t,i}$, was drawn at round t, $i \in [d]$. Clearly $\sum_{i=1}^d N_{t,i} = B$. With this view we note that:

$$\mathbf{g}_t = \frac{1}{B}\sum_{\ell=1}^B \mathbf{g}_t^\ell = \frac{d}{B\gamma_t}\sum_{i=1}^d N_{t,i}(o_t^i - \frac{1}{2})\sqrt{\lambda_{t,i}}\mathbf{v}_{t,i},$$

and noting that since $\mathbf{v}_i$s are orthogonal to each other:

$$\begin{aligned}\mathbf{E}_{i_t^1,o_t^1,\ldots i_t^d,o_t^d}\left[\|\mathbf{g}_t\|_{\mathbf{w}_t}^*\right] &\le \frac{d}{2B\gamma_t}\mathbf{E}_{i_t^1,\ldots,i_t^d}\left[\sqrt{\sum_{i=1}^d N_{t,i}^2\mathbf{v}_{t,i}^\top(\mathbf{v}_{t,i}\mathbf{v}_{t,i}^\top)\mathbf{v}_{t,i}}\right] \\ &= \frac{d}{2B\gamma_t}\sqrt{\sum_{i=1}^d \mathbf{E}_{i_t}\left[N_{t,i}^2\right]}.\end{aligned}$$

We now note that $N_i \sim \text{Bin}(B, 1/d)$ and if $X \sim \text{Bin}(n,p)$, then $\mathbf{E}[X^2] = V(X) + \mathbf{E}[X]^2 = np(1-p) + n^2p^2$. Using this and denoting $B_d = \min\{B, d\} \leq d$, we get:

$$\mathbf{E}_{i_t^1, o_t^1, \ldots i_t^d, o_t^d}\left[\|\mathbf{g}_t\|_{\mathbf{w}_t}^*\right] \leq \frac{d}{2B_d\gamma_t}\sqrt{\sum_{i=1}^{d} \frac{3B_d}{d}} = \frac{d}{\gamma_t\sqrt{B_d}}.$$

■

Finally we are now ready to proof the bound of our main theorem, Thm. 2:
Proof [Proof of Thm. 2] The proof follows almost the same steps that of proof of Thm. 1. In particular, same as the proof of Thm. 1, one can bound:

$$\begin{aligned} 0.22\sum_{t=1}^{T}\left(\mathbf{u} - \mathbf{w}_t\right)^\top \boldsymbol{\theta}_t^* &\leq \sum_{t=1}^{T} 2\eta\|\mathbf{g}_t\|_{\mathbf{w}_t}^{*2} + \frac{\nu \log \frac{1}{\delta}}{\eta} \\ &\leq 2\eta\sum_{t=1}^{T}\frac{d^2}{\gamma_t^2 B_d} + \frac{\nu\log\frac{1}{\delta}}{\eta}, \end{aligned}$$

where the last inequality follows from Lem. 6. Same as before, choosing $\mathbf{u} := \arg\max_{\mathbf{x}\in\mathcal{D}_\delta}\sum_{t=1}^{T}\boldsymbol{\theta}_t^{*\top}\mathbf{x}$, and recalling that $\mathbf{x}^* := \arg\max_{\mathbf{x}\in\mathcal{D}}\sum_{t=1}^{T}\boldsymbol{\theta}_t^{*\top}\mathbf{x}$, we get:

$$\begin{aligned} \sum_{t=1}^{T}\left(\mathbf{x}^* - \mathbf{w}_t\right)^\top\boldsymbol{\theta}_t^* &\leq \sum_{t=1}^{T}\left(u - \mathbf{w}_t\right)^\top\boldsymbol{\theta}_t^* + \delta TLD \\ &= \frac{1}{0.22}\left[\eta\sum_{t=1}^{T}\frac{2d^2}{\gamma_t^2 B_d} + \frac{\nu\log\frac{1}{\delta}}{\eta}\right] + \delta TLD, && \text{from Eq. (6)} \\ &\leq \frac{1}{0.22}\left[\frac{\eta d^2 T}{\min\{1, H_{\mathcal{D},\psi}^2\}B_d} + \frac{\nu\log\frac{1}{\delta}}{\eta}\right] + \delta TLD, && \text{since we chose } \gamma \leq \min\{1, 0.7H_{\mathcal{D},\psi}\} \\ &= \frac{d\sqrt{\nu T\log T}}{0.22H_{\mathcal{D},\psi}\sqrt{B_d}} + LD, \end{aligned}$$

choosing $\eta = \frac{\sqrt{\nu B_d}H_{\mathcal{D},\psi}}{d\sqrt{T\log T}}$ and $\delta = \frac{1}{T}$, concludes the proof noting the diameter of the decision set $\mathcal{D} \subseteq \mathcal{B}_d(1)$ is bounded by 1, and the lipschitz constant $L \leq \max_{t\in[T]}\|\boldsymbol{\theta}_t^*\| \leq 1$. ■

C.2. Regret Analysis: Proof of Thm. 1

We start our analysis with noting a key property of Plackett Luce model, which will be crucial for our proof analysis. Let us denote a MNL$(\theta_1, \ldots, \theta_n)$ model on n items such that, probability of any ranking $\sigma \in \Sigma$ The property will rely on some notations, that I will describe below:
Ranking respecting pairwise comparison. Let $\sigma \in \Sigma_n$ any ranking such that $\sigma(i) \succ \sigma(j)$.
Recall then the probability of any permutation $\sigma \in \Sigma_[n]^m$ is given by:

$$\mathbf{P}_t(\sigma_m|S) = \prod_{i=1}^{m} \frac{\boldsymbol{\theta}_t^{*\top}\mathbf{x}_t^{\sigma(i)}}{\sum_{j=i}^{m} \boldsymbol{\theta}_t^{*\top}\mathbf{x}_t^{\sigma(j)} + \sum_{j \in S \setminus S_m} \boldsymbol{\theta}_t^{*\top}\mathbf{x}_t^{\sigma(j)}},$$

Then let us denote by $\Sigma_{i,j} = \{\sigma \in \Sigma \mid \sigma(i) < \sigma(j), \text{ i.e. i is preferred over j}\}$. Then it can be shown that

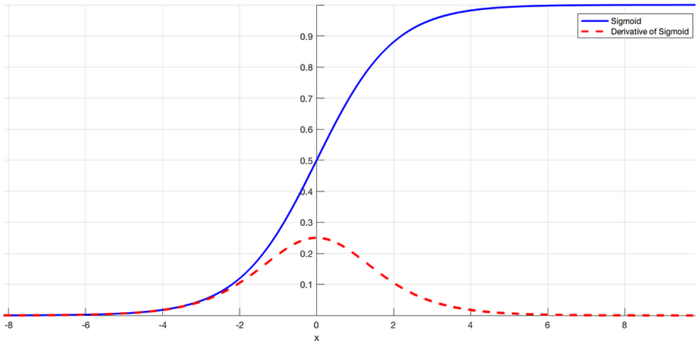
Lemma 7 (Pairwise Properties of MNL Model)

$$P(i > j) = \sum_{\sigma \in \Sigma_{i,j}} P(\sigma)$$

Appendix D. Some Useful Results

Lemma 8 *For any $x \in [-0.7, 0.7]$, $\sigma'(x) \in [0.222, 0.25]$.*

Proof Let us first consider the positive interval $x \in [0, 0.7]$. Note by definition, since $\sigma(x) = \frac{1}{1+e^{-x}}$, $\forall x \in \mathbb{R}$, first derivative and the second derivative of sigmoid is respectively given by:



$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})(1+e^{-x})}$$
$$= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2},$$

and

$$\sigma''(x) = \left[\frac{e^{-x}}{(1+e^{-x})^2} - \frac{2e^{-x}}{(1+e^{-x})^3}\right].$$

As shown in the right figure, this brings us to the observation that $\sigma''(x) < 0$ for any $x > 0$, and thus $\sigma'(\cdot)$ is a decreasing function in the interval $[0, \infty)$. Thus the function $\sigma'(\cdot)$ attains maximum at $x = 0$ and minimum at $x = 0.7$, yielding $\sigma'(x) \in [0.222, 0.25]$ in the range $x \in [0, 0.7]$.
The result follows from the symmetry of $\sigma'(\cdot)$ function around the Y-axis. ■