Optimal Algorithms for Stochastic Contextual Preference Bandits

Aadirupa Saha[*]

Abstract

We consider the problem of preference bandits in the contextual setting. At each round, the learner is presented with a context set of K items, chosen randomly from a potentially infinite set of arms $\mathcal{D} \subseteq \mathbb{R}^d$. However, unlike classical contextual bandits, our framework only allows the learner to receive feedback in terms of item preferences: At each round, the learner is allowed to play a subset of size q (any $q \in \{2, \ldots, K\}$) upon which only a (noisy) winner of the subset is revealed. Yet, same as the classical setup, the goal is still to compete against the best context arm at each round. The problem is relevant in various online decision-making scenarios, including recommender systems, information retrieval, tournament ranking–typically any application where it's easier to elicit the items' relative strength instead of their absolute scores. To the best of our knowledge, this work is the first to consider preference-based stochastic contextual bandits for potentially infinite decision spaces. We start with presenting two algorithms for the special case of pairwise preferences ($q = 2$): The first algorithm is simple and easy to implement with an $\tilde{O}(d\sqrt{T})$ regret guarantee, while the second algorithm is shown to achieve the optimal $\tilde{O}(\sqrt{dT})$ regret, as follows from our $\Omega(\sqrt{dT})$ matching lower bound analysis. We then proceed to analyze the problem for any general q-subsetwise preferences ($q \geq 2$), where surprisingly, our lower bound proves the fundamental performance limit to be $\Omega(\sqrt{dT})$ yet again, independent of the subsetsize q. Following this, we propose a matching upper bound algorithm justifying the tightness of our results. This implies having access to subsetwise preferences does not help in faster information aggregation for our feedback model. All the results are corroborated empirically against existing baselines.

1 Introduction

Sequential decision-making problems with side information (in the form of features or attributes), have been popular in machine learning as contextual bandits [16, 13, 26]. A contextual bandit learner, at each round, observes a context before taking action based on it. The resulting payoff is typically assumed to depend on the context and the action taken according to an unknown map. The learner aims to play the best possible action for the current context at each time and thus minimize its regret with respect to an oracle that knows the payoff function.

In many learning settings, however, it is more common to be able to only *relatively* compare actions, in a decision step, instead of being able to gauge their absolute utilities. E.g., information retrieval, search engine optimization, recommender systems, crowdsourcing, drug testing, tournament ranking, social surveys etc. [18, 21]. A specific application example is: Consider the problem of recommending items from a catalog to users on a shopping website. Each time, the context is determined by the visiting user's features together with all items' features. When a subset of items is presented, the user clicks on one of them according to a relative preference model; only the items presented matter – this plausibly models comparative cognitive choices being made by humans. The aim is to converge to identify the overall best item in the catalog. Additionally, since our model is designed to leverage

[*]Microsoft Research, New York, US; `aasa@microsoft.com`.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

the structured preference feedback, we can handle very large decision spaces, unlike state of the art dueling bandits algorithms that only deals with pairwise preferences [47], [23] [43], [41] etc., or even some of the recent works on general subsetwise preference bandits (e.g. [30, 33, 38]). All of them need to maintain a K^2 matrix explicitly, and their regret scales as $O(K)$. This becomes impractical for large (or infinite) action spaces of size K.

In this work, we consider a natural *structured contextual preference bandit* setting, comprised of items with intrinsic (absolute) scores depending on their features in an unknown way, e.g., linear with unknown weights. In the most general setup, the learner plays (compares) a subset $2 \le q \le K$ items and gets to see a (noisy) ℓ-length rank-ordered feedback (where $1 \le \ell \le q$) of the top-ranked items of the selected subset with a probability distribution governed by the items' scores. We are primarily interested in developing adaptive subset-selection algorithms for which guarantees can be given for a suitably defined measure of regret.

To the best of our knowledge, we are the first to give *theoretical guarantees* for the above problem of *regret minimization in contextual preference bandits for potentially infinitely large decision spaces* and design provably optimal algorithms for the same. Some recent works [17, 38] have considered this problem in the subset selection setup but their algorithms do not guarantee any finite time regret bounds, neither validate their performance optimality theoretically. In [15], authors considered a version of adversarial contextual dueling bandit problem, which only takes into account the special case of pairwise feedback ($q = 2$). Though similar in names, their problem framework and consequently the analysis are very different from ours. In their setup, at each time the preference matrix is determined by a randomly chosen known context variable, and the algorithm is assumed to have access to a pool of finite set of policies. The goal of the learner is to compete with the 'optimal policy in the pool' for which they proposed an sparring-EXP4 like algorithm—their regret bound though scales as $O(\sqrt{T})$ in T, it depends linearly on the size of arm space (K in their notation), and thus become vacuous for infinitely large decision space. Moreover, their computational complexity also scales linearly with the size of the policy class, which tends to be prohibitively large in practice.

1. We propose the problem of structured contextual preference bandits, where at each iteration, the learner is presented with a set of K-arms $\mathcal{S}_t$ (each represented by d-dimensional feature vectors), and the task of the learner is to select a subset of at most q-items ($1 \le q \le K$), with the objective being to identify the 'best arm' of $\mathcal{S}_t$ in every round. The novelty of the framework lies in the relative preference based feedback model, which only allows the learner to see a noisy draw of the top-ranked items of the selected subset, unlike the absolute reward feedback used for standard bandits setup.

2. We first address the problem for the special case of dueling bandits, where $q = 2$, i.e. the learner only have access to pairwise preferences of the selected item pairs at each round. We propose two algorithms for the basic dueling bandit setting: Our first algorithm, Maximum-Informative-Pair (Alg 1), is based on the idea of selecting the most uncertain pair ('max-variance') from the set of 'promising candidates', and we prove an $O(d\sqrt{T})$ regret for the same (Thm. 3).

3. Our second algorithm Stagewise-Adaptive-Duel (Alg. 3), is developed on the idea of tracking, in a phased fashion, the best arm of the context set, which ensures a sharper concentration rate of the pairwise scores. This results in $\tilde{O}(\sqrt{dT})$ [2] regret guarantee (Thm. 5), improving upon the regret bound of our previous algorithm by a $\sqrt{d}$ factor.

4. We also show that fundamental regret lower bound of $\Omega(\sqrt{dT})$ for the contextual dueling bandit problem addressed here (Thm. 3). Thus theoretically our second algorithm (Alg. 3) is provably optimal, however our Alg. 1 often works better in practice as we show in the experiments (Sec. 6).

5. We then analyze the problem for more general subsetwise preference feedback, where at each round, the learner is allowed to play a subset of q-items ($q \ge 2$), upon which the winner feedback of the played subset is revealed [10, 30, 33]. We first prove an $\Omega(\sqrt{dT})$ regret lower bound (Thm. 11) for the problem, which establishes that, interestingly, having access to subsetwise preferences does not really help in faster information aggregation (for the specific preference model considered here). Subsequently, we also discuss an algorithm with near-optimal regret guarantee, whose regret bound is also independent of the subsetsize q up to logarithmic factors (Thm. 12).

[2]The notation $\tilde{O}(\cdot)$ hides logarithmic dependencies.

Finally, we corroborate our theoretical findings with empirical evaluations. Detailed *Related works*, *Experiments* and most of the technical *proofs* are moved to the appendix.

Related Works. Surprisingly, following the same spirit of extending standard multiarmed bandits (MAB) to continuous decision spaces (as in *linear* or *GP-bandits*), there has been really very little work on the continuous extension of the Dueling Bandit problem [24]. The works in [38, 17] did attempt a similar objective, however, without any satisfactory theoretical performance guarantees. [24] considers the problem of dueling bandits on continuous arm set but under rather restrictive sets of assumptions: Twice continuously differentiable, lipschitz, strongly convex and smooth score/reward function, which are often impractical for modeling any real-world preference feedback. Recently, [28] consider the problem of k-way assortment selection, where the problem is to minimize regret against the set of highest revenue. However, their objective is much different from ours. We focus on the regret against the single best item per iteration. Thus our pairwise action set must allow repeated item pulls, unlike their setup, due to which their algorithm does not lead to sublinear regret for our objective. In another recent work, [10] did address the problem of regret minimization in continuous *Dueling Bandits*, however without any finite time regret guarantee of their proposed algorithms, unlike ours. A more elaborated survey is given in Appendix A.

2 Preliminaries and Problem Formulation

Notations. For any positive integer $n \in \mathbb{N}_+$, we denote by $[n]$ the set $\{1, 2, ..., n\}$. $\mathbf{1}(\varphi)$ is generically used to denote an indicator variable that takes the value 1 if the predicate φ is true, and 0 otherwise. The decision space is denoted by $\mathcal{D} \subseteq \mathbb{R}^d$, where $d \in \mathbb{N}_+$. We use $\mathbf{1}_d$ to denote an d-dimensional vector of all 1's. For any matrix $M \in \mathbb{R}^{d\times d}$, we denote respectively by $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ the maximum a minimum eigenvalue of matrix M. For any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_M := \sqrt{x^\top M\mathbf{x}}$ denotes the weighted ℓ_2-norm associated with matrix M (assuming M is positive-definite).

2.1 Problem Setup

We consider the stochastic K-armed contextual dueling bandit problem for T rounds, where at each round $t \in [T]$, the learner is presented with a context set $\mathcal{S}_t = \{\mathbf{x}_1^t, \mathbf{x}_2^t, \ldots, \mathbf{x}_K^t\} \subseteq \mathcal{D} \subset \mathbb{R}^d$ of size K which is drawn IID from some d-dimensional decision space $\mathcal{D}$ (according to some unknown distribution on $\mathcal{D}$, say $\mathcal{P}_{\mathcal{D}}$). The learner is permitted to play a subset $\mathcal{X}_t \subseteq \mathcal{S}_t$ of size $q \geq 2$, given a fixed $q \leq K$ (see the formal setup in Sec. 2.1). Clearly for $q = 2$, the problem reduces to the sequential duel (pair of items) selection, say in this case we denote $\mathcal{X}_t = \{\mathbf{x}_t, \mathbf{y}_t\}$. Upon this, the environment provides a stochastic subsetwise preference feedback as follows:

Subsetwise-Preference Feedback Model. At any round t, upon selecting $\mathcal{X}_t$, the learner receives a *winner feedback* o_t such that: $Pr\big(o_t = \mathbf{x} \mid \mathcal{X}_t\big) = \frac{e^{g(\mathbf{x})}}{\sum_{\mathbf{y}\in\mathcal{X}_t} e^{g(\mathbf{y})}}$ for any $\mathbf{x} \in \mathcal{X}_t$, where $g : \mathcal{D} \mapsto [0, 1]$ is a utility score function on each point in the decisions space $\mathbf{x} \in \mathcal{D}$. Note our preference model essentially boils down to the well studied Plackett Luce (PL) choice model with the individual PL score of item-x being $e^{g(\mathbf{x})}$ [21, 35, 32, 30]. If $q = 2$, then $o_t = \mathbf{1}(\mathbf{x}_t$ preferred over $\mathbf{y}_t)$ simply indicates the preferred arm of the duel $(\mathbf{x}_t, \mathbf{y}_t)$, such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, the probability $\mathbf{x}$ is preferred over $\mathbf{y}$, denoted by $Pr(\mathbf{x} \succ \mathbf{y})$, is drawn according to $\sim Ber\big(\sigma\big(h(\mathbf{x}, \mathbf{y})\big)\big)$, here $\sigma(\cdot)$ being the sigmoid transformation $\big($i.e. $\sigma(x) = \frac{1}{1+e^{-x}}$ for any $x \in \mathbb{R}\big)$.

Analysis with linear scores. In this paper, we assume that $g(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\theta}^*$, $\forall \mathbf{x} \in \mathcal{D}$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is some unknown fixed vector in $\mathbb{R}^d$ such that $\|\boldsymbol{\theta}^*\| \leq 1$. We will henceforth denote this linear utility based 'subsetwise preference model' as SPM($\boldsymbol{\theta}^*, d, q$).

Objective: Regret Minimization. Suppose $\mathbf{x}_t^* := \arg\max_{\mathbf{x}\in\mathcal{S}_t} \mathbf{x}^\top\boldsymbol{\theta}^*$ is the best arm (with highest score) of round t. Then the goal of the learner is to minimize the T-round cumulative regret $R_T = \sum_{t=1}^T r_t$ with respect to the best arm $\mathbf{x}_t^*$ of each round t, where we measure the instantaneous regret r_t of playing a set $\mathcal{X}_t$ in terms of the average score of the played duel $\frac{\sum_{\mathbf{x}\in\mathcal{X}_t}\mathbf{x}^\top\boldsymbol{\theta}^*}{|\mathcal{X}_t|}$. Precisely,

$$R_T = \sum_{t=1}^T \left({\mathbf{x}_t^*}^\top\boldsymbol{\theta}^* - \frac{\sum_{\mathbf{x}\in\mathcal{X}_t}\mathbf{x}^\top\boldsymbol{\theta}^*}{|\mathcal{X}_t|} \right). \tag{1}$$

Above notion of learner's regret is motivated from the definition of classical K*-armed dueling bandit regret* introduced by [44] which is later adopted by the dueling bandit literature [46, 23, 3, 43, 45, 38, 31]. Here the context set at any round t is assumed to be a fixed set of K arms $\mathcal{S}_t = [K]$, and at each round the instantaneous regret incurred by the learner for playing an arm-pair $(i_t, j_t) \in [K \times K]$ is given by $r_t^{(\text{DB})} = \frac{\mathbf{P}(i_*, i_t) + \mathbf{P}(i_*, j_t) - 1}{2}$, $i_* \in [K]$ being the 'best-arm' in the hindsight (e.g. condorcet winner [46] or copeland winner [23, 41]) depending on the underlying preference matrix $\mathbf{P} \in [0,1]^{K \times K}$.

Remark 1 (Equivalence with Dueling Bandit Regret when $q = 2$)**.** *It is easy to note that assuming the context set $\mathcal{S}_t \subseteq \mathcal{D}$ to be fixed $\forall t \in [T]$ and denoting $\mathbf{x}_t^* = \mathbf{x}^*$, our regret definition (Eqn.* (1)*) is equivalent to dueling bandit regret (up to constant factors), as in our case the pairwise advantage against the best arm (i.e. $Pr(\mathbf{x}^*, \mathbf{x}_t) - \frac{1}{2}$) can be both upper and lower bounded as: (1) $Pr(\mathbf{x}^*, \mathbf{x}_t) - \frac{1}{2} = \frac{(e^{\mathbf{x}^{*\top}\boldsymbol{\theta}^*} - e^{\mathbf{x}_t^\top \boldsymbol{\theta}^*})}{2(e^{\mathbf{x}^{*\top}\boldsymbol{\theta}^*} + e^{\mathbf{x}_t^\top \boldsymbol{\theta}^*})} \leq \frac{(\mathbf{x}^* - \mathbf{x}_t)^\top \theta^*}{2}$ and (2) $Pr(\mathbf{x}^*, \mathbf{x}_t) - \frac{1}{2} \geq \frac{(\mathbf{x}^* - \mathbf{x}_t)^\top \theta^*}{4e}$. Combining above claims we get $\frac{R_T}{4e} \leq R_T^{(DB)} \leq \frac{R_T}{2}$ (analysis detail given in Appendix B.1).*

3 Dueling Feedback $(q = 2)$: Algorithm and Analysis

We first analyze the problem for pairwise preference feedback ($q = 2$ case). Before proceeding to the actual algorithms, it is crucial to note that, same as generalized linear Bandits (GLB) [16, 26], both our algorithms use standard MLE techniques for maintaining a 'tight estimate' of $\boldsymbol{\theta}^*$ (Line 5 of Alg. 1, Line 8 of Alg. 3). This is since our dueling preference feedback can be seen as a generalized linear reward over item pairs (details in Appendix C). However, the regret definition of dueling bandits (Eqn. (1)) being very different than GLB objective, a direct application of any GLB algorithm will simply lead to $O(T)$ regret in our setup: The objective of any GLB learner is to converge to the arm with highest reward, unlike ours. Thus any GLM routine would always converge to the best-worst arm pair as that would be perceived to be the duel with highest reward (pairwise preference in our case). On the contrary, to have any sublinear regret, we require the learner to eventually play only the best arm in the duel, which does not have the highest pairwise-preference (reward). This is the inherent complexity and primary difference of any dueling bandit problem w.r.t. GLB objective or any MAB framework per se. As a result, a sublinear MAB algorithm never works for dueling/preference bandit objectives. To circumvent the problem, we design fairly non-trivial arm-selection rules for our proposed algorithms, e.g., Alg. 1 first needs to construct a promising-set $\mathcal{C}_t$ and then pick the maximum informative pair amongst them (Line 6-7, Alg1). Alg. 2 needs to optimally maintain the set of 'good-items' $\mathcal{G}^s$ with a careful arm-selection rule which significantly differs from any GLB approach (Line 19, Alg 2). Consequently, we also need to resolve to new proof ideas towards analyzing their regret guarantees which remains one of the primary novelty of this work.

3.1 Algorithm-1: Maximum-Informative-Pair

Our first algorithm is a computationally efficient one with a $O(d\sqrt{T})$ regret guarantee (Thm. 3), which is only suboptimal by a factor of $O(\sqrt{d})$ (as reflects from our lower bound, Thm. 10, Sec. 4).

Main Idea: At any time t, the algorithm simply maintains an UCB estimate on the pairwise scores $\bar{h}(\mathbf{x}, \mathbf{y}) := \hat{\boldsymbol{\theta}}^\top (\mathbf{x} - \mathbf{y}) + \eta \|\mathbf{x} - \mathbf{y}\|_{V_t^{-1}}$ for any pair of arms $(\mathbf{x}, \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathcal{S}_t$, where $V_t = \sum_{\tau=1}^{t-1} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$, $\hat{\boldsymbol{\theta}}$ being the maximum likelihood estimator (MLE) of our preference model parameter. It then builds a set of promising arms $\mathcal{C}_t := \{\mathbf{x} \in \mathcal{S}_t \mid \bar{h}(\mathbf{x}, \mathbf{y}) > 0, \forall \mathbf{y} \in \mathbf{y} \in S_t \setminus \{\mathbf{x}\}\}$: Arms that beats the rest in terms of their UCB score $\bar{h}(\mathbf{x}, \mathbf{y})$. Finally it plays the most-uncertain (least sampled) pair $(\mathbf{x}_t, \mathbf{y}_t) := \arg\max_{\mathbf{x}, \mathbf{y} \in \mathcal{C}_t} \|\mathbf{x} - \mathbf{y}\|_{V_t^{-1}}$. Note (x_t, y_t) is the pair with highest pairwise score variance in $\mathcal{C}_t$, hence 'maximum informative'. (Detail in Alg. 1.)

Analysis. Regret guarantee of Alg.1 (Thm. 3) is based on the following main lemmas.

Lemma 1 (Self-Normalized Bound)**.** *Suppose $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_t, \mathbf{y}_t)\}$ be a sequence of arm-pair played such that all arms $\mathbf{x} \in \{\mathbf{x}_\tau, \mathbf{y}_\tau\}_{\tau=1}^t$ belong to the ball of unit radius. Also suppose the initial exploration length t_0 be such that $\lambda_{\min}\Big(\sum_{\tau=1}^{t_0} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top\Big) \geq 1$. Then $\forall\, t > t_0$, $\sum_{\tau=t_0+1}^t \|(\mathbf{x}_\tau - \mathbf{y}_\tau)\|_{V_{\tau+1}^{-1}} \leq \sqrt{2dt \log\left(\frac{4t_0 + t}{d}\right)}$, where recall $V_{\tau+1} := \sum_{j=1}^\tau (\mathbf{x}_j - \mathbf{y}_j)(\mathbf{x}_j - \mathbf{y}_j)^\top$.*

Algorithm 1 **Maximum-Informative-Pair (MaxInP)**

input: Learning rate $\eta > 0$, exploration length $t_0 > 0$
init: Select t_0 pairs $\{(\mathbf{x}_\tau, \mathbf{y}_\tau)\}_{\tau \in [t_0]}$, each drawn at random from $\mathcal{S}_\tau$, and observe the corresponding preference feedback $\{o_\tau\}_{\tau \in [t_0]}$
Set $V_{t_0+1} := \sum_{\tau=1}^{t_0} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$
for $t = t_0 + 1, t_0 + 2, \ldots T$ **do**
    Compute the MLE $\hat{\boldsymbol{\theta}}_t$ on $\{(\mathbf{x}_\tau, \mathbf{y}_\tau, o_\tau)\}_{\tau=1}^{t-1}$: $\sum_{\tau=1}^{t-1} \Big(o_\tau - \sigma\big((\mathbf{x}_\tau - \mathbf{y}_\tau)^\top \hat{\boldsymbol{\theta}}_t\big)\Big)(\mathbf{x}_\tau - \mathbf{y}_\tau) = \mathbf{0}$
    $\mathcal{C}_t := \{\mathbf{x}, \mathbf{y} \in \mathcal{S}_t \mid (\mathbf{x} - \mathbf{y})^\top \hat{\boldsymbol{\theta}}_t + \eta \|(\mathbf{x} - \mathbf{y})\|_{V_t^{-1}} > 0\}$
    Compute $(\mathbf{x}_t, \mathbf{y}_t) := \arg\max_{\mathbf{x}, \mathbf{y} \in \mathcal{C}_t} \|\mathbf{x} - \mathbf{y}\|_{V_t^{-1}}$
    Play the duel $(\mathbf{x}_t, \mathbf{y}_t)$. Receive $o_t = \mathbf{1}(\mathbf{x}_t \text{ beats } \mathbf{y}_t)$
    Update $V_{t+1} = V_t + (\mathbf{x}_t - \mathbf{y}_t)(\mathbf{x}_t - \mathbf{y}_t)^\top$
end for

Lemma 2 (Confidence Ellipsoid)**.** *Suppose the initial exploration length t_0 be such that $\lambda_{\min}\Big(\sum_{\tau=1}^{t_0} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top\Big) \geq 1$, and κ is as defined in Thm. 3. Then for any $\delta > 0$, with probability at least $(1-\delta)$, for all $t > t_0$, $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\|_{V_t} \leq \frac{1}{2\kappa}\sqrt{\frac{d}{2}\log\Big(1 + \frac{2t}{d}\Big) + \log\frac{1}{\delta}}$, where recall $V_{t+1} := \sum_{\tau=1}^{t} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$.*

Theorem 3 (Regret bound of Maximum-Informative-Pair (Alg. 1))**.** *Let $\eta = \frac{1}{2\kappa}\sqrt{\frac{d}{2}\log(1 + \frac{2T}{d}) + \log\frac{1}{\delta}}$, where $\kappa := \inf_{\|x-y\| \leq 2, \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\| \leq 1} \Big[\sigma'\big((\mathbf{x} - \mathbf{y})^\top \hat{\boldsymbol{\theta}}\big)\Big]$ is the minimum slope of the estimated sigmoid when $\hat{\boldsymbol{\theta}}$ is sufficiently close to $\boldsymbol{\theta}^*$ $\Big(\sigma'(\cdot)$ being the first order derivative of the sigmoid function $\sigma(\cdot)\Big)$. Then given any $\delta > 0$, with probability at least $(1 - 2\delta)$, the T round cumulative regret of Maximum-Informative-Pair satisfies:*

$$R_T \leq t_0 + \left(\frac{1}{\kappa}\sqrt{\frac{d}{2}\log\left(1 + \frac{2T}{d}\right) + \log\frac{1}{\delta}}\right)\sqrt{2dT\log\left(\frac{4t_0 + T}{d}\right)} = O\bigg(d\sqrt{T}\log\Big(\frac{T}{d\delta}\Big)\bigg),$$

where we choose $t_0 = 2\left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(1/\delta)}}{\lambda_{\min}(B)}\right)^2 + \frac{4}{\lambda_{\min}(B)}$, $B = \mathbf{E}_{\mathbf{x},\mathbf{y} \overset{iid}{\sim} \mathcal{P}_{\mathcal{D}}}[(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^\top]$ (for some universal problem independent constants $C_1, C_2 > 0$).

Proof. **(sketch)** Our choice of t_0 ensures that with probability at least $(1 - \delta)$, V_{t_0+1} is full rank, or more precisely $\lambda_{\min}(V_{t_0+1}) \geq 1$ [42] (see Lem. 13, Appendix D for the formal statement). We next apply the two key concentration lemmas (Lem 1 and 2), upon expressing the regret definition in terms of the above concentration results. Precisely, using Lem. 2 and our *'most informative pair'* based arm selection strategy, we can show at any round $t > t_0$, we can bound $r_t \leq 4\|\mathbf{x}_t - \mathbf{y}_t\|_{V_t^{-1}}$. The results now follows from the choice of η and Lem.1. The complete proof is given in Appendix D.1. □

3.2 Algorithm-2: Stagewise-Adaptive-Duel (Sta′D))

Our second algorithm runs with a provable optimal regret bound of $\tilde{O}(\sqrt{dT})$, except with an additional $\sqrt{logK}$ factor. When $K = O(1)$, the algorithm thus yields an optimal regret guarantee.

Main Idea. The algorithm proceeds in stages $s \in \lfloor \log T \rfloor$ with the aim of tracking a set of *'promising arms'* $\mathcal{G}^s$ per stage: At each such stage s, we maintain confidence interval on the pairwise scores of each index pair (i, j) $p_t^s(i, j)$. If at any stage s, the confidence-score of any arm-pair is not estimated to the *'sufficient accuracy'*, we play that pair and include it in the set of *'informative pairs'* of stage $\boldsymbol{\phi}^s$ to be further explored in following rounds. Otherwise, we sequentially eliminate the *'weakly-performing'* arms which gets defeated by some other arm in terms of its optimistic pairwise score, and proceed to the next stage $s + 1$ to examine the surviving arms with a stricter confidence interval. Now if the pairwise scores of every index pair in the set of *'promising-arms'* $\mathcal{G}^s$ is almost *'accurately estimated up to high confidence'*, we pick the first arm $\mathbf{x}_t$ as the one which has the maximum estimated score, followed by choosing its strongest challenger $\mathbf{y}_t$ which beats $\mathbf{x}_t$ with highest pairwise preference. The algorithm is given in Alg. 3 (Appendix 3).

Analysis. Thm. 5 proves an optimal $\tilde{O}(\sqrt{dT})$ regret bound of Alg. 3 (matching the regret lower bound, Thm. 10). It is worth pointing that the near optimal regret analysis of Stagewise-Adaptive-Duel crucially relies on the stronger concentration guarantees of the pairwise scores (compared to the weaker concentration of Lem. 2 used earlier for Alg. 1). Note, this is made possible by specifically maintaining the independent *'stage-wise informative samples'* $\boldsymbol{\phi}^s$ as also pioneered in few of the earlier works [8, 13] for multi-armed bandits (see our proof analysis of Lem. 4).

Lemma 4 (Sharper Concentration of Pairwise Scores)**.** *Consider any $\delta > 0$, and suppose we set the parameters of Stagewise-Adaptive-Duel (Alg. 3) as $\eta = \frac{3}{2\kappa}\sqrt{2\log\frac{3TK}{\delta}}$, where $\kappa := \inf_{\|\mathbf{x}-\mathbf{y}\|\le 2, \|\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}\|\le 1}\left[\sigma'\big((\mathbf{x}-\mathbf{y})^\top\hat{\boldsymbol{\theta}}\big)\right]$, and $t_0 = 2\left(\frac{C_1\sqrt{d}+C_2\sqrt{\log(2/\delta)}}{\lambda_{\min}(B)}\right)^2 + \frac{4\Lambda}{\lambda_{\min}(B)}$, where $\Lambda = \frac{8}{\kappa^4}\left(d^2+\log\frac{3}{\delta}\right)$ and $B = \mathbf{E}_{\mathbf{x},\mathbf{y}\overset{iid}{\sim}\mathcal{P}_\mathcal{D}}[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^\top]$ (for some universal problem independent constants $C_1, C_2 > 0$). Then with probability at least $(1-\delta)$, for all stages $s \in \lceil \log T\rceil$ at all rounds $t > t_0$ and for all index pairs $i, j \in \mathcal{G}^s$ of round t: $|(\mathbf{x}_i^t - \mathbf{x}_j^t)^\top(\theta^*-\theta_t^s) \le p_t^s(i,j)|$.*

Theorem 5 (Regret bound of Stagewise-Adaptive-Duel (Alg. 3))**.** *Consider we set t_0 and η as per Lem. 4. Then for any $\delta > 0$, with probability at least $(1-\delta)$, the regret of Alg. 3 can be bounded as:*

$$R_T \le t_0 + 4\eta\sqrt{2d\log\left(\frac{4t_0T}{d}\right)}\sqrt{T\log T} + 2\sqrt{T} = O\left(\frac{\sqrt{dT\log T}}{\kappa}\sqrt{\log\left(\frac{TK}{\delta}\right)\log\left(\frac{Td}{\kappa}\log\frac{1}{\delta}\right)}\right)$$

Proof. **(sketch)** Suppose we denote by $\boldsymbol{\phi}^c := \{t \in [T]\setminus[t_0] \mid t \notin \cup_{s=1}^{\lfloor \log T\rfloor}\boldsymbol{\phi}^s\}$ the set of all good time intervals where all the index pairs $p_t^s(i,j)$ are estimated within the confidence accuracy $\frac{1}{\sqrt{T}}$. The proof crucially relies on the concentration bound of Lem. 4, from which we first derive:

Lemma 6. *For any $t > t_0$, suppose the pair $(\mathbf{x}_t, \mathbf{y}_t)$ is chosen at stage s_t, and i_t^* denotes the best action of round t, i.e. $\mathbf{x}_{i_t^*}^t = \mathbf{x}_t^* = \arg\max_{\mathbf{x}\in\mathcal{S}_t}\mathbf{x}^\top\boldsymbol{\theta}^*$. Then for any $\delta \in (0,1)$, with probability at least $(1-\delta)$, for all $t > t_0$: $i_t^* \in \mathcal{G}^{s_t}$ and for both $\mathbf{x} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, $g(\mathbf{x}_t^*) - g(\mathbf{x}) \le \begin{cases}\frac{2}{\sqrt{T}} & \text{if } t \in \boldsymbol{\phi}^c\\ \frac{4}{2^{s_t}} & \text{otherwise}\end{cases}$.*

Owning to Lem. 1 and due to the construction of our *'stagewise-good item pairs'* we can also show:

Lemma 7. *At any stage $s \in \lfloor \log T\rfloor$, with probability at least $(1-\delta)$, $\sqrt{|\boldsymbol{\phi}^s|} \le \eta 2^s\sqrt{2d\log\left(\frac{4t_0T}{d}\right)}$.*

The final regret bound can be derived combining the results of Lem. 6 and 7. (see Appendix E.5). $\square$

4 Lower Bound for Dueling $(q = 2)$ Feedback

We now proceed to understand the fundamental performance limit of our contextual preference bandits problem for pairwise preference $(q = 2)$ case. Towards this we use a novel idea of reducing linear bandits problem to our setup which finally leads to the desired lower bound of $\Omega(\sqrt{dT})$.

Reducing Linear-Contextual Bandits to our framework. Let us instantiate any instance of our K-armed contextual dueling bandit problem by its problem parameter $\boldsymbol{\theta}^* \in R^d$ as $\mathcal{I}^{cdb}(\boldsymbol{\theta}^*, K, T)$. On the other hand define any instance of K-armed contextual linear bandit problem [13] with problem parameter $\boldsymbol{\theta}^* \in R^d$ as $\mathcal{I}^{clb}(\boldsymbol{\theta}^*, K, T)$: Recall in this setup, at each iteration the learner is provided with a context set $\mathcal{S}_t = \{\mathbf{x}_1^t, \mathbf{x}_2^t, \ldots, \mathbf{x}_K^t\} \subset \mathbb{R}^d$ of size K (as before $\|\mathbf{x}\|_2 \le 1,\ \forall \mathbf{x} \in \mathcal{S}_t$), upon which the learner choose an arm $\mathbf{x}_t \in \mathcal{S}_t$, and the environment feedbacks a reward $r(\mathbf{x}_t) = \mathbf{x}_t^\top\boldsymbol{\theta}^* + \varepsilon_t$, where ε_t is a zero mean random noise. Objective is to minimize the regret with respect to the best action, $\mathbf{x}_t^* := \arg\max_{\mathbf{x}\in\mathcal{S}_t}\mathbf{x}^\top\boldsymbol{\theta}^*$, of each round t, defined as: $R_T^{clb} := \sum_{t=1}^T\left(\mathbf{x}_t^{*\top}\boldsymbol{\theta}^* - \mathbf{x}_t^\top\boldsymbol{\theta}^*\right)$,

Main Idea. For proving a lower bound for $\mathcal{I}^{cdb}(\boldsymbol{\theta}^*, K, T)$, we first show under *Gumbel noise* [9, 34], any instance of contextual linear bandits $\mathcal{I}^{clb}$ can be reduced to an instance of $\mathcal{I}^{cdb}$.

Lemma 8 (Reducing $\mathcal{I}^{clb}$ with *Gumbel noise* to $\mathcal{I}^{cdb}$)**.** *There exists a reduction from the $\mathcal{I}^{clb}$ problem (under Gumbel noise, i.e. $\varepsilon_t \overset{iid}{\sim}$ Gumbel$(0,1)$) to $\mathcal{I}^{cdb}$ which preserves the expected regret.*

Proof. **(sketch)** Suppose we have a blackbox algorithm for the instance of $\mathcal{I}^{cdb}$ problem, say $\mathcal{A}^{cdb}$. To prove the claim, our goal is to show that this can be used to solve the $\mathcal{I}^{clb}$ problem where the underlying stochastic noise, ϵ_t at round t, is generated from a Gumbel$(0,1)$ distribution [39, 9]. Precisely we can construct an algorithm for $\mathcal{I}^{clb}(\boldsymbol{\theta}^*, K, T)$ (say $\mathcal{A}^{clb}$) using $\mathcal{A}^{cdb}$ as follows:

Algorithm 2 $\mathcal{A}^{clb}$ for problem $\mathcal{I}^{clb}(\boldsymbol{\theta}^*, K, T)$

for $t = 1, 2, \ldots \lceil \frac{T}{2} \rceil$ **do**
  Receive: $(\mathbf{x}_t, \mathbf{y}_t) \leftarrow$ duel played by $\mathcal{A}^{cdb}$ at time t.
  Play $\mathbf{x}_t$ at round $(2t-1)$ of $\mathcal{I}^{clb}$. Receive $r(\mathbf{x}_t)$.
  Play $\mathbf{y}_t$ at round $2t$ of $\mathcal{I}^{clb}$. Receive $r(\mathbf{y}_t)$.
  Feedback: $o_t = \mathbf{1}(r(\mathbf{x}_t) > r(\mathbf{y}_t))$ to $\mathcal{A}^{cdb}$.
end for

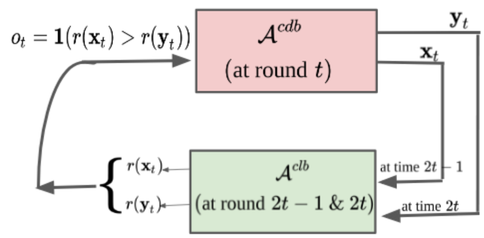

Figure 1: Demonstration of the reduction idea: $\mathcal{I}^{clb}$ to $\mathcal{I}^{cdb}$

Lemma 9. *If $\mathcal{A}^{clb}$ rums on a problem instance $\mathcal{I}^{clb}(\boldsymbol{\theta}^*, K, 2T)$ with Gumbel$(0,1)$ noise, then the internally the algorithm $\mathcal{A}^{cdb}$ runs on a problem instance of $\mathcal{I}^{cdb}(\boldsymbol{\theta}^*, K, T)$.*

The proof of the above lemma is given in Appendix F.2. Lem. 9 precisely shows a reduction of $\mathcal{I}^{clb}$ to $\mathcal{I}^{cdb}$. The claim of Lem. 9 now follows from the regret definitions of the $\mathcal{I}^{clb}$ and $\mathcal{I}^{cdb}$, precisely we can show for any fixed T, $2R_T^{cdb} = R_{2T}^{clb}$. Complete proof is deferred to Appendix F.1. □

Our lower bound result now immediately follows as a implication of Thm. 10 and from the existing lower bound result of K-armed d-dimensional contextual linear bandits problem [13].

Theorem 10 (SPM$(\boldsymbol{\theta}^*, d, 2)$: Regret Lower Bound)**.** *For any algorithm $\mathcal{A}^{cdb}$ for the problem of linear-score based stochastic K-armed contextual dueling bandit of dimensional-d, there exists a sequence of d-dimensional context sets $\{\mathbf{x}_1^t, \ldots \mathbf{x}_K^t\}_{t=1}^T$ and a constant $\gamma > 0$ such that the regret incurred by $\mathcal{A}^{cdb}$ on T rounds is at least $\Omega(\frac{\gamma}{2}\sqrt{2dT})$, i.e. $R_T(\mathcal{A}^{cdb}) \geq \frac{\gamma}{2}\sqrt{2dT}$, for any $T \geq d^2$.*

5 Analysis for General Subsetwise Preference Feedback (any $q \in [K]$)

We now extend our analysis to any general q-subsetwise feedback, where at round t the learner is permitted to play a subset $\mathcal{X}_t \subseteq \mathcal{S}_t$ of size $q \geq 2$, given a fixed $q \leq K$ (formal setup in Sec. 2.1). We first analyze the regret lower bound, which, somewhat surprisingly, turns out to be independent of q (Thm. 11). We also propose an algorithm following this. Proof details are given in Appendix G.

5.1 Regret Lower Bound

We first show that for any given $q \geq 2$, there exists a problem instance where no learner can achieve a better learning rate than $\Omega(\sqrt{dT})$. However the conclusions are much similar to the existing regret lower bounds for finite K arm preference bandits for Plackett-Luce (PL) model [30, 33]. As described in Sec. 2.1, since our preference model can also be seen a special case of PL model, our results show even in the contextual framework, the learner can not attain a faster rate by playing larger subsets.

Theorem 11 (Regret Lower Bound (Subsetwise Preferences))**.** *Given any $q \geq 2$ and $d > 1$, for any algorithm $\mathcal{A}$ for the problem of stochastic K-armed d-dimensional linear contextual bandit with SPM$(\boldsymbol{\theta}, d, q)$ feedback model, there exists a sequence of d-dimensional context sets $\{\mathbf{x}_1^t, \ldots \mathbf{x}_K^t\}_{t=1}^T$ and a choice of $\boldsymbol{\theta}$ such that the regret incurred by $\mathcal{A}$ on T rounds is at least $\Omega(\sqrt{dT})$.*

5.2 Algorithm and Regret Guarantee

Given the above lower bound, the first thing to note is our Alg. 3, itself yields an optimal $\tilde{O}(\sqrt{dT})$ algorithm (which only makes pairwise queries per round) in case the problem allows the learner to query preferences of any subsets of size $1, 2, \ldots, q$. However, we here propose a general version of Alg. 3 which is also based on the idea of *stagewise-elimination* but can exploit subsetwise preferences for any general $q \geq 2$; it works even if the learner is restricted to play only sets of size q. Moreover, even though for the worst case instances, a better regret guarantee is not possible (as shown in Thm. 11), it can exploit the problem structure when there is a sufficient 'quality gap' between items.

Main Ideas. The main idea is to exploit the subsetwise feedback using the idea of *rank-breaking* [21] for extracting pairwise estimates from subsetwise feedback. Given these pairwise estimates, now the algorithm may proceed the same as the original Alg. 3, however instead of selecting a pair of arms, we can now select a subset of q most-promising arms: First, by selecting a potential good arm and then recursively selecting the best challenger of the already selected items. The complete description is given in Alg. 4 (Appendix G.2). The challenging part, however, lies in its regret analysis which requires justifying the right concentration rates of $\boldsymbol{\theta}_t^s$, obtained from the above pairwise estimates. The detailed regret analysis is given in Appendix G.3, which finally lead to the following guarantee:

Theorem 12 (Regret bound of Sta′D++ (Alg. 4))**.** *Consider any $\delta > 0$, and suppose we set the parameters of Sta′D++ (Alg. 3) as $\eta = \frac{3}{2\kappa}\sqrt{2\log\frac{3qTK}{\delta}}$, and $t_0 = 2\left(\frac{C_1\sqrt{d}+C_2\sqrt{\log(2q/\delta)}}{\lambda_{\min}(B)}\right)^2 + \frac{4\Lambda}{\lambda_{\min}(B)}$, where Λ, κ, B is as defined in Lem. 4. Then with probability at least $(1-\delta)$, the T round cumulative regret of Sta′D++ is at most $O\left(\frac{\sqrt{dT\log(T)}}{\kappa}\sqrt{\log\left(\frac{qTK}{\delta}\right)\log\left(\frac{Td}{\kappa}\log\frac{q}{\delta}\right)}\right)$.*

6 Experiments

This section gives empirical performances of our algorithms (Alg. 1 and 3) and compare them with some existing preference learning algorithms. The details of the algorithms are given below:

Algorithms. 1. MaxInP: Algorithm Maximum-Informative-Pair (Alg. 1 as described in Sec. 3.1). 2. Sta′D: Our proposed algorithm Stagewise-Adaptive-Duel (Alg. 1 as described in Sec. 3.2). 3 *SS*: *Self-Sparring* (independent beta priors on each arm) algorithm for multi-dueling bandits [37] 4. *RUCB*: *The Relative Upper Confidence Bound* algorithm for regret minimization in standard dueling bandits [46]. 5. *DTS*: *Dueling-Thompson Sampling* algorithm for best arm identification problem in bayesian dueling bandits [17]. In every experiment, the performances of the algorithms are measured in terms of *cumulative regret* (sec. 1), averaged across 50 runs, reported with standard deviation.

Constructing Problem Instances. The difficulty of the instances depends on the difference of scores of the best and second best arms, which, in the hindsight, is actually governed by the 'worst case slope' of the sigmoid function κ (see the dependency of κ in Thm. 3 or Thm. 5), and also by the underlying problem parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$. So we used 3 different *linear score based* problem instances based on 3 different characterizations of $\boldsymbol{\theta}^* \in \mathbb{R}^d$ (with K arms and dimension d): **1. Easy** $h(d,K)$, **2. Extreme** $e(d,K)$, and **3. Intermediate** $m(d,K)$, by suitably adjusting the norm $\|\boldsymbol{\theta}^*\|_2$. Also in all settings, the d-dimensional feature vectors (of the arm set) are generated as random linear combination of each arm to be a random linear combination of the d-dimensional basis vectors (for scaling issues of the item scores, we limit each instance vector to be within ball of radius 1, i.e. ℓ_2-norm upper bounded by 1).

Regret vs Time. For this experiment we fix $d = 10$ and $K = 50$. Fig. 2 shows both our algorithms MaxInP and Sta′D always outperform the rest, and their performance gets comparatively better with increasing hardness of the instances. As expected, *RUCB* performs the worst as by construction it fails to exploit the structure of underlying linear score based item preferences, due to the same reason *SS* performs poorly as well (note we implement independent armed version of the *Self-Sparring* algorithm [37] for this case, and later the *Kernelized* version for the case of non-linear item scores). On the contrary, *DTS* performs reasonably well as it designed to exploit the underlying utility structures in the pairwise-preferences.

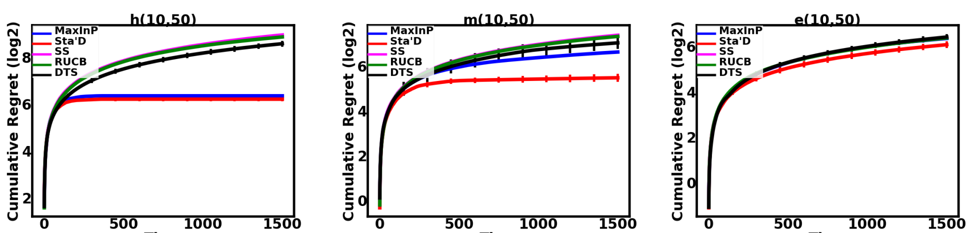

Figure 2: Average Cumulative Regret vs Time across algorithms on 3 problem instances (linear score based preferences, $d = 10, K = 50$)

Regret vs Context-size (K). We now compare the (averaged) final cumulative regret of each algorithm over varying context set size (K) over two different problem instances. For this experiment we fix $d = 10$ and $T = 1500$. From Fig. 3 note that again our algorithms superiorly outperforms the other baselines with *DTS* performing competitively. *SS* and *RUCB* performs very badly due to the same reason as explained for Fig 2. Interesting observation to make is that the performance of both our algorithms MaxInP and Sta′D is almost independent of K as also follows from their respective regret guarantees (see Thm. 3 and Thm. 5)–as long as d is fixed our algorithms clearly could identify the best item irrespectively of the size K of the context set, owning to their ability to exploit the underlying preference structures, unlike *SS* or *RUCB*.

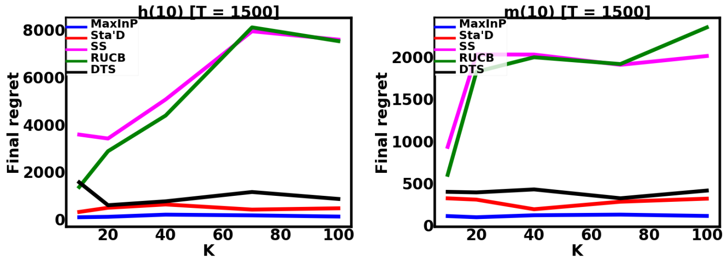

Figure 3: Final regret (averaged) vs context-set size (K) across different algorithms on two different problem instances $(d = 10)$

Regret vs Dimension (d). For this experiment we fix $K = 80$ and $T = 1500$. From Fig. 4 shows that in general the performance of every algorithm degrades over increasing d. However the effect is much most severe for the *DTS* baseline compared to ours. Since *RUCB* can not exploit the underlying preference structure, its performance is mostly independent of d and same goes for *SS* as well due to the same reason. The interesting observation to make is with increasing d, fixed T and K, our first algorithm MaxInP indeed performs worse than Sta′D, following their theoretical regret guarantees which shows the former has a multiplicative $O(\sqrt{d})$ worse regret than latter (see Thm. 3, 5).

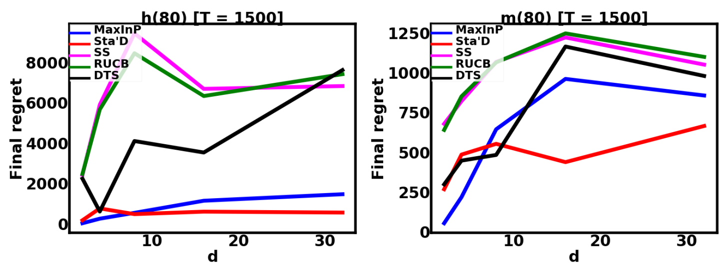

Figure 4: Final regret (averaged) vs featue dimension (d) across algorithms on two different problem instances $(K = 80)$

Non-Linear score based preferences We finally also run some experiments to compare our regret performances on non-linear score based preferences (i.e. the score function $g(\mathbf{x})$ is not linear in $\mathbf{x}$, see Sec. 2.1 for details). We use three different score functions for the non-linear setup.

Environments. We use thsese 3 functions as $g(\cdot)$: ***1.*** *Quadratic*, ***2.*** *Six-Hump Camel* and 3. *Gold Stein*. *Quadratic* is the reward function $f(\mathbf{x}) = \mathbf{x}^\top H\mathbf{x} + \mathbf{x}^\top \mathbf{w} + c$, where $H \in [-1, 1]^{d\times d}, \mathbf{w} \in [-1, 1]^d$ and $c \in [-1, 1]$ are randomly generated. The *Six-Hump Camel* and *Gold Stein* functions are as described in [17]. For all cases, we fix $d = 3$ and $K = 50$.

Algorithms. We use a slightly modified version of our two algorithms (MaxInP and Sta′D) for the non-linear scores, since the GLM based parameter estimation techniques would no longer work here. But unfortunately, without suitable assumptions, we do not have an efficient way to estimate the score functions for this general setup, so instead we fit a GP to the underlying unknown score function $g(\cdot)$ based on the Laplace approximation based technique suggested in [29] (see Chap 3). For *SS* also we now used the *kernelized self-sparring* version of the algorithm [37], and for *DTS* we now fit a GP model (instead of a linear model).

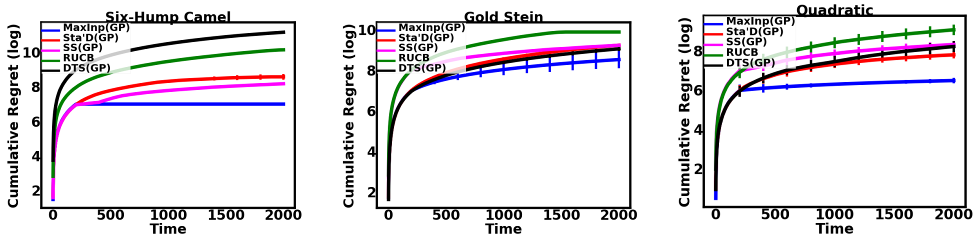

Figure 5: Avg. Cumulative Regret vs Time across algorithms on 3 problem instances (non-linear score based preferences, $d = 10, K = 50$)

Fig. 5 shows our algorithms still outperform the rest in almost all the instances. This actually implies the generality of our algorithmic ideas which applies beyond linear-scores (hence it is also worth understanding their theoretical guarantees for this general setup in future works). Moreover, unlike the previous scenarios *SS*, now starts to perform better since it could now exploit underlying preferences structures owing to the implementation of *kernelized self-sparring* [37].

7 Conclusion and Future Scopes

We consider the problem of regret minimization for contextual preference bandits for potentially infinite decision spaces. To the best of our knowledge, this is the first work to give optimal regret (up to logarithmic factors) $\tilde{O}(\sqrt{dT})$ algorithms along with a matching lower bound analysis. The problem of contextual preference bandits being a niche and highly practically relevant area, undoubtedly there are numerous interesting open threads to pursue along this direction: E.g. considering other link functions (probit, nested logit, etc.) based on the real-world system needs, analyzing the regret bound for adversarial preferences, or even extending preferences bandits setup to other related bandit frameworks like side information [27, 22], or feedback graphs [4, 5] etc. Analyzing instance dependent regret guarantees also remains to be an interesting future direction to see what improvements can be claimed for larger q under a sufficient 'quality gap' between the items.

Acknowledgment

Sincere thanks to Branislav Kveton and Ofer Meshi for some of the initial discussions when author was an intern at Google Research, Mountain View. Special thanks to Branislav Kveton for his insightful guidance on basics of linear-bandits and towards developing Algorithm 1.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[3] Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864, 2014.

[4] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *JMLR WORKSHOP AND CONFERENCE PROCEEDINGS*, volume 40. Microtome Publishing, 2015.

[5] Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

[6] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

[7] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[8] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[9] Hossein Azari, David Parks, and Lirong Xia. Random utility theory for social choice. In *Advances in Neural Information Processing Systems*, pages 126–134, 2012.

[10] Brian Brost, Yevgeny Seldin, Ingemar J. Cox, and Christina Lioma. Multi-dueling bandits and their application to online ranker evaluation. *CoRR*, abs/1608.06253, 2016.

[11] Xi Chen, Yuanzhi Li, and Jieming Mao. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2504–2522. SIAM, 2018.

[12] Yuxin Chen and Changho Suh. Spectral mle: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pages 371–380, 2015.

[13] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[14] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[15] Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587, 2015.

[16] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[17] Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1282–1291. JMLR. org, 2017.

[18] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2014.

[19] Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Optimal sample complexity of m-wise data for top-k ranking. In *Advances in Neural Information Processing Systems*, pages 1685–1695, 2017.

[20] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[21] Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016.

[22] Tomas Kocak, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, pages 613–621, 2014.

[23] Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on Learning Theory*, pages 1141–1154, 2015.

[24] Wataru Kumagai. Regret analysis for continuous dueling bandit. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

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

[26] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org, 2017.

[27] Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

[28] Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. In *Advances in Neural Information Processing Systems*, pages 3145–3155, 2019.

[29] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT press, 2006.

[30] Wenbo Ren, Jia Liu, and Ness B Shroff. Pac ranking from pairwise and listwise queries: Lower bounds and upper bound. *arXiv preprint arXiv:1806.02970*, 2018.

[31] Aadirupa Saha and Aditya Gopalan. Battle of bandits. In *Uncertainty in Artificial Intelligence*, 2018.

[32] Aadirupa Saha and Aditya Gopalan. Pac-battling bandits with plackett-luce: Tradeoff between sample complexity and subset size. *arXiv preprint arXiv:1808.04008*, 2018.

[33] Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, pages 983–993, 2019.

[34] Hossein Azari Soufiani, David C Parkes, and Lirong Xia. Preference elicitation for general random utility models. In *Uncertainty in Artificial Intelligence*, page 596. Citeseer, 2013.

[35] Hossein Azari Soufiani, David C Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *ICML*, pages 360–368, 2014.

[36] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[37] Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Multi-dueling bandits with dependent arms. In *Conference on Uncertainty in Artificial Intelligence*, UAI'17, 2017.

[38] Yanan Sui, Vincent Zhuang, Joel W Burdick, and Yisong Yue. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*, 2017.

[39] Jakub M Tomczak. On some properties of the low-dimensional gumbel perturbations in the perturb-and-map model. *Statistics & Probability Letters*, 115:8–15, 2016.

[40] Jakub M Tomczak. On some properties of the low-dimensional gumbel perturbations in the perturb-and-map model. *Statistics & Probability Letters*, 115:8–15, 2016.

[41] Tanguy Urvoy, Fabrice Clerot, Raphael Féraud, and Sami Naamane. Generic exploration and k-armed voting bandits. In *International Conference on Machine Learning*, pages 91–99, 2013.

[42] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[43] Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, 2016.

[44] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[45] Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pages 307–315, 2015.

[46] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten De Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. *arXiv preprint arXiv:1312.3393*, 2013.

[47] Masrour Zoghi, Shimon Whiteson, Remi Munos, Maarten de Rijke, et al. Relative upper confidence bound for the k-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pages 10–18. JMLR, 2014.

Supplementary: Optimal Algorithms for Stochastic Contextual Preference Bandits

A Related Works (detailed)

The problem of regret minimization for multiarmed bandits (MAB) is very well studied in the online learning literature [7, 2, 6, 20, 25], where the learner gets to see a noisy draw of absolute reward feedback of an arm upon playing a single arm per round. However, the classical multiarmed bandits only consider finitely many arms (i.e. finite decision set) [7, 2, 25], whereas in practice, it is much more realistic to consider large decision spaces with potentially infinitely many actions, which is the reason that continuum extensions of MABs are widely studied in the literature – this includes *linear bandits* [14, 13, 1] where the true mean rewards of the arms are some linear functions of the arm features, *GLM bandits* [16, 26], where instead of linear rewards, the expected rewards of the arms follow generalized linear models (GLMs), or more generally *GP Bandits* [36] where the arms' rewards are assumed to be non-linear functions of the arm features.

On the other hand, relative feedback variants of stochastic MAB problem have also been widely studied: The most popular one being the *Dueling Bandit*, where, instead of getting a noisy feedback of the reward of the chosen arm, the learner only gets to see a noisy feedback on the pairwise preference of two arms selected by the learner. The objective is to find a high-value arm in the stochastic model and algorithmic approaches based on both upper-confidence-bounds (UCBs) [46, 23] and Thompson sampling [43] are known. There are also very few recent developments on the subsetwise extension on *Dueling Bandit* problem [38, 10, 31, 32, 30, 11]. Some of the existing work also explicitly consider the Plackett-Luce parameter estimation problem with subset-wise feedback but for offline setup only [21, 12, 19].

Surprisingly though, following the same spirit of extending MAB to continuous decision spaces (as in *linear* or *GP-bandits*), there has been very little work on the continuous extension of Dueling Bandit problem [24], that too without any theoretical performance guarantees [38, 17]. Although [24] considers the problem of dueling bandits on continuous arm set, the underlying score/reward function of each arm needs to be twice continuously differentiable, lipschitz, strongly convex as well as smooth which are very restrictive assumptions to model the preference feedback. In a recent work, [28] considers the problem of k-way assortment selection, where the problem is to minimize regret with respect to the set of highest revenue—again, this objective is much different than ours, which focuses on regret with respect to the single best item per iteration and hence our pairwise action set allows repeated items, unlike their setup, due to which their algorithm does not lead to sublinear regret in our case. The recent works by [38, 10, 17] did address the problem of regret minimization in continuous *Dueling Bandits*, or even the subsetwise generalization of the setting termed as *Multi-dueling bandits* with *Sparring* based [3] thompson sampling algorithm. However, none of these works analyzed any finite horizon regret guarantee of their proposed algorithms, which remains the primary objective of our work.

B Appendix for Sec. 2

B.1 Derivations for Rem. 1

Claim: $\frac{R_T}{4e} \leq R_T^{(DB)} \leq \frac{R_T}{2}$.

Proof. Recall $r_T^{(\text{DB})} = \sum_{t=1}^{T} \frac{\mathbf{P}(i_*, i_t) + \mathbf{P}(i_*, j_t) - 1}{2}$.
Note that:

$$Pr(\mathbf{x}^*, \mathbf{x}_t) - \frac{1}{2} = \frac{(e^{\mathbf{x}^{*\top}\boldsymbol{\theta}^*} - e^{\mathbf{x}_t^\top \boldsymbol{\theta}^*})}{2(e^{\mathbf{x}^{*\top}\boldsymbol{\theta}^*} + e^{\mathbf{x}_t^\top \boldsymbol{\theta}^*})} = \frac{(e^{\boldsymbol{\theta}^{*\top}(\mathbf{x}^* - \mathbf{x}_t)} - 1)}{2(e^{\boldsymbol{\theta}^{*\top}(\mathbf{x}^* - \mathbf{x}_t)} + 1)} \leq \frac{(\mathbf{x}^* - \mathbf{x}_t)^\top \theta^*}{2}, \; \Big[\text{since } \boldsymbol{\theta}^{*\top}(\mathbf{x}^* - \mathbf{x}_t) \geq 0 \Big],$$

where the last inequality follows since $(e^{\boldsymbol{\theta}^{*\top}(\mathbf{x}^*-\mathbf{x}_t)} - 1) \le \boldsymbol{\theta}^{*\top}(\mathbf{x}^* - \mathbf{x}_t)\Big(\frac{1}{1-\frac{\boldsymbol{\theta}^{*\top}(\mathbf{x}^*-\mathbf{x}_t)}{2}}\Big) \le 2(\boldsymbol{\theta}^{*\top}(\mathbf{x}^* - \mathbf{x}_t))$ since $\|\mathbf{x}^*\| \le 1$ and $\|\boldsymbol{\theta}^*\| \le 1$, then applying Cauchy-Schwartz and by the definition of $\mathbf{x}^*$ we get $\boldsymbol{\theta}^{*\top}(\mathbf{x}^* - \mathbf{x}_t) \le \boldsymbol{\theta}^{*\top}\mathbf{x}^* \le 1$. Moreover since $e^x - 1 > x$ for any $x > 0$, we also have:

$$Pr(\mathbf{x}^*, \mathbf{x}_t) - \frac{1}{2} = \frac{(e^{\mathbf{x}^{*\top}\boldsymbol{\theta}^*} - e^{\mathbf{x}_t^\top\boldsymbol{\theta}^*})}{2(e^{\mathbf{x}^{*\top}\boldsymbol{\theta}^*} + e^{\mathbf{x}_t^\top\boldsymbol{\theta}^*})} \ge \frac{(\mathbf{x}^* - \mathbf{x}_t)^\top\theta^*}{4e}, \ \Big[\text{since } \boldsymbol{\theta}^{*\top}(\mathbf{x}^* - \mathbf{x}_t) \ge 0\Big].$$

where the last inequality follows since $\|\mathbf{x}^*\| \le 1$, $\|\mathbf{x}_t\| \le 1$, $\|\boldsymbol{\theta}^*\| \le 1$ and hence applying Cauchy-Schwartz both $\mathbf{x}^{*\top}\boldsymbol{\theta}^*$ and $\mathbf{x}_t^\top\boldsymbol{\theta}^* \le 1$. Note that the same inequalities can be applied for $Pr(\mathbf{x}^*, \mathbf{y}_t) - \frac{1}{2}$ as well. Finally combining above claims and summing over $t = 1, 2, \ldots T$ we get $\frac{R_T}{4e} \le R_T^{(DB)} \le \frac{R_T}{2}$. □

C Connection to GLM Bandit's Feedback Model

We start by observing the relation of our preference feedback model to that of generalized linear model (GLM) based bandits [16, 26]–precisely the feedback mechanism. The setup of *GLM bandits* generalizes the stochastic *linear bandits* problem [14, 1], where at each round t the learner is supposed to play a decision point $\mathbf{x}_t$ from a set fixed decision set $\mathcal{D} \subset \mathbb{R}^d$, upon which a noisy reward feedback f_t is revealed by the environment such that $f_t = \mu(\mathbf{x}_t^\top\boldsymbol{\theta}^*) + \varepsilon_t$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is some unknown fixed direction, $\mu : \mathbb{R} \mapsto \mathbb{R}$ is a fixed strictly increasing link function, and ε_t is a zero mean ν sub-Gaussian noise for some universal constant $\nu > 0$, i.e. $\mathbf{E}\big[e^{\lambda\varepsilon_t} \mid \mathcal{H}_t\big] \le e^{\frac{\lambda^2\nu^2}{2}}$ and $\mathbf{E}[\varepsilon_t \mid \mathcal{H}_t] = 0$ (here $\mathcal{H}_t$ denotes the sigma algebra generated by the history $\{(x_\tau, o_\tau)\}_{\tau=1}^t$ till time t).

The important connection now to make is that our structured dueling bandit feedback can be modeled as a GLM feedback model on the decision space of pairwise differences $\mathcal{D}' := \{(\mathbf{x} - \mathbf{y}) \mid \mathbf{x}, \mathbf{y} \in \mathcal{D}\}$, since in this case the feedback received by the learner upon playing a duel $(\mathbf{x}_t, \mathbf{y}_t)$ can be seen as: $o_t = \sigma\big((d_t)^\top\boldsymbol{\theta}^*\big) + \varepsilon_t'$ where ε_t' is a 0-mean $\mathcal{H}_t$-measurable random binary noise such that

$$\varepsilon_t' = \begin{cases} 1 - \sigma\big(d_t^\top\boldsymbol{\theta}^*\big), & \text{with probability } \sigma\big(d_t^\top\boldsymbol{\theta}^*\big), \\ -\sigma\big(d_t^\top\boldsymbol{\theta}^*\big), & \text{with probability } \big(1 - \sigma\big(d_t^\top\boldsymbol{\theta}^*\big)\big), \end{cases}$$

where we denote $d_t := (\mathbf{x}_t - \mathbf{y}_t) \in \mathcal{D}'$, and it is easy to verify that ε_t' is $\frac{1}{2}$ sub-Gaussian. Thus our dueling based preference feedback model can be seen as a special case of GLM bandit feedback on the decision space $\mathcal{D}'$ where the link function $\mu(\cdot)$ in our case is the sigmoid $\sigma(\cdot)$.

The above connection is crucially used in both of our proposed algorithms (Sec. 3.1 and 3.2) for estimating the unknown parameter $\boldsymbol{\theta}^*$, denoted by $\hat{\boldsymbol{\theta}}_t$, with high confidence using maximum likelihood estimation on the observed pairwise preferences $\{(\mathbf{x}_\tau, \mathbf{y}_\tau, o_\tau)\}_{\tau=1}^{t-1}$ upto time $(t-1)$, following the same technique suggested by [16, 26].

D Appendix for Sec. 3

D.1 Proof of Thm. 3

Theorem 3 (Regret bound of Maximum-Informative-Pair (Alg. 1)). *Let* $\eta = \frac{1}{2\kappa}\sqrt{\frac{d}{2}\log(1 + \frac{2T}{d}) + \log\frac{1}{\delta}}$, *where* $\kappa := \inf_{\|x-y\|\le 2, \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\| \le 1}\Big[\sigma'\big((\mathbf{x} - \mathbf{y})^\top\hat{\boldsymbol{\theta}}\big)\Big]$ *is the minimum slope of the estimated sigmoid when* $\hat{\boldsymbol{\theta}}$ *is sufficiently close to* $\boldsymbol{\theta}^*$ $\Big(\sigma'(\cdot)$ *being the first order derivative of the sigmoid function* $\sigma(\cdot)\Big)$. *Then given any* $\delta > 0$, *with probability at least* $(1 - 2\delta)$, *the* T *round cumulative regret of Maximum-Informative-Pair satisfies:*

$$R_T \le t_0 + \left(\frac{1}{\kappa}\sqrt{\frac{d}{2}\log\left(1 + \frac{2T}{d}\right) + \log\frac{1}{\delta}}\right)\sqrt{2dT\log\left(\frac{4t_0 + T}{d}\right)} = O\Big(d\sqrt{T}\log\Big(\frac{T}{d\delta}\Big)\Big),$$

where we choose $t_0 = 2\left(\frac{C_1\sqrt{d}+C_2\sqrt{\log(1/\delta)}}{\lambda_{\min}(B)}\right)^2 + \frac{4}{\lambda_{\min}(B)}$, $B = \mathbf{E}_{\mathbf{x},\mathbf{y}\overset{iid}{\sim}\mathcal{P}_{\mathcal{D}}}[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^\top]$ *(for some universal problem independent constants* $C_1, C_2 > 0$*).*

Proof. Our choice of t_0 ensures that with probability at least $(1-\delta)$, V_{t_0+1} is full rank. More precisely $\lambda_{\min}(V_{t_0+1}) \geq 1$ owing to the following standard results from random matrix theory:

Lemma 13. *Suppose* $\{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2,\ldots,(\mathbf{x}_n,\mathbf{y}_n)\}$ *be a sequence of* n *arm-pairs such that all* $\mathbf{x} \in \{\mathbf{x}_\tau,\mathbf{y}_\tau\}_{\tau=1}^n$ *are drawn iid from some fixed distribution* $\mathcal{P}$, $\|\mathbf{x}\|_2 \leq 1$. *Then for any positive constant* $C > 0$, *and any* $\delta \in (0,1)$, *there exist two positive constants* C_1 *and* C_2 *such that if we choose* $n > 2\left(\frac{C_1\sqrt{d}+C_2\sqrt{\log(1/\delta)}}{\lambda_{\min}(B)}\right)^2 + \frac{4C}{\lambda_{\min}(B)}$, *then* $Pr\left(\lambda_{\min}\left[\sum_\tau^n(\mathbf{x}_\tau-\mathbf{y}_\tau)(\mathbf{x}_\tau-\mathbf{y}_\tau)^\top\right] \geq C\right) \geq (1-\delta)$, *where* $B = \mathbf{E}_{\mathbf{x},\mathbf{y}\overset{iid}{\sim}\mathcal{P}}[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^\top]$.

Proof. The result follows from the existing results of [26] (Proposition 1), which is adapted from [42] (Thm. 5.39), except we need to carefully construct the sample complexity bound considering that in our case all the iid vectors $(\mathbf{x}_t - \mathbf{y}_t) \in \mathbb{R}^d$ belong to a ball of radius 2. $\square$

We next derive the two key concentration lemmas, Lem. 1 and Lem. 2) that holds straightforwardly from the existing results of generalized linear bandits [16, 26], owing to the connection of our structured dueling bandits problem setup to that of GLM bandits.

The rest of the proof lies in expressing the regret bound in terms of the above concentration results which is possible owing to our *'most informative pair'* based arm selection strategy, as described below:

Now recall that the instantaneous regret at t: $r_t = \frac{(\mathbf{x}_t^*-\mathbf{x}_t)^\top\boldsymbol{\theta}^*+(\mathbf{x}_t^*-\mathbf{y}_t)^\top\boldsymbol{\theta}^*}{2}$. Then using above conditions and the by our arm selection strategy:

$$
\begin{aligned}
2r_t &= (\mathbf{x}_t^*-\mathbf{x}_t)^\top\boldsymbol{\theta}^* + (\mathbf{x}_t^*-\mathbf{y}_t)^\top\boldsymbol{\theta}^*\\
&= (\mathbf{x}_t^*-\mathbf{x}_t)^\top\hat{\boldsymbol{\theta}}_t + (\mathbf{x}_t^*-\mathbf{x}_t)^\top(\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}_t) + (\mathbf{x}_t^*-\mathbf{y}_t)^\top\hat{\boldsymbol{\theta}}_t + (\mathbf{x}_t^*-\mathbf{y}_t)^\top(\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}_t)\\
&\overset{(1)}{\leq} \eta\|\mathbf{x}_t^*-\mathbf{x}_t\|_{V_t^{-1}} + \|\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}_t\|_{V_t}\|\mathbf{x}_t^*-\mathbf{x}_t\|_{V_t^{-1}} + \eta\|\mathbf{x}_t^*-\mathbf{y}_t\|_{V_t^{-1}} + \|\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}_t\|_{V_t}\|\mathbf{x}_t^*-\mathbf{y}_t\|_{V_t^{-1}}\\
&\overset{(2)}{\leq} \eta\|\mathbf{x}_t^*-\mathbf{x}_t\|_{V_t^{-1}} + \eta\|\mathbf{x}_t^*-\mathbf{x}_t\|_{V_t^{-1}} + \eta\|\mathbf{x}_t^*-\mathbf{y}_t\|_{V_t^{-1}} + \eta\|\mathbf{x}_t^*-\mathbf{y}_t\|_{V_t^{-1}}\\
&\overset{(3)}{\leq} \eta\|\mathbf{x}_t-\mathbf{y}_t\|_{V_t^{-1}} + \eta\|\mathbf{x}_t-\mathbf{y}_t\|_{V_t^{-1}} + \eta\|\mathbf{x}_t-\mathbf{y}_t\|_{V_t^{-1}} + \eta\|\mathbf{x}_t-\mathbf{y}_t\|_{V_t^{-1}}\\
&= \left(\frac{2}{\kappa}\sqrt{\frac{d}{2}\log\left(1+\frac{2T}{d}\right)+\log\frac{1}{\delta}}\right)\|\mathbf{x}_t-\mathbf{y}_t\|_{V_t^{-1}},
\end{aligned}
$$

where inequality (1) holds since since both $\mathbf{x}_t, \mathbf{y}_t \in \mathcal{C}_t$, by definition of $\mathcal{C}_t$ this implies: $(\mathbf{x}_t^*-\mathbf{x}_t)^\top\hat{\boldsymbol{\theta}}_t < \eta\|\mathbf{x}_t^*-\mathbf{x}_t\|_{V_t^{-1}}$, and $(\mathbf{x}_t^*-\mathbf{y}_t)^\top\hat{\boldsymbol{\theta}}_t < \eta\|\mathbf{x}_t^*-\mathbf{y}_t\|_{V_t^{-1}}$. Inequality (2) follows from Lem. 2, and (3) follows from the arm selection strategy. The final inequality follows by simply replacing the value of η. We now proceed to bound the cumulative regret as follows:

$$
\begin{aligned}
R_t &= \sum_{t=1}^T r_t = \sum_{t=1}^{t_0} r_t + \sum_{t=t_0+1}^T r_t\\
&\overset{(1)}{\leq} t_0 + \sum_{t=t_0+1}^T r_t \leq t_0 + \frac{1}{2}\sum_{t=t_0}^T\left(\frac{2}{\kappa}\sqrt{\frac{d}{2}\log\left(1+\frac{2T}{d}\right)+\log\frac{1}{\delta}}\right)\|\mathbf{x}_t-\mathbf{y}_t\|_{V_t^{-1}}\\
&\overset{(2)}{\leq} t_0 + \left(\frac{1}{\kappa}\sqrt{\frac{d}{2}\log\left(1+\frac{2T}{d}\right)+\log\frac{1}{\delta}}\right)\sqrt{2dT\log\left(\frac{4t_0+T}{d}\right)}
\end{aligned}
$$

where the first inequality holds since $\|\mathbf{x}_t^*\| \le 1$ and $\|\boldsymbol{\theta}^*\| \le 1$. Thus by applying Cauchy-Schwartz and by the definition of $\mathbf{x}_t^*$ we get $\boldsymbol{\theta}^{*\top}(\mathbf{x}_t^* - \mathbf{x}) \le \boldsymbol{\theta}^{*\top}\mathbf{x}_t^* \le 1 \, \forall \mathbf{x} \in \mathcal{D}$–we consider the trivial bound $r_t = 1$ for the initial t_0 rounds of random exploration. Inequality (2) simply follows from Lem. 1, which concludes the proof. □

D.2 Proof of Lem. 1

Lemma 1 (Self-Normalized Bound). *Suppose $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_t, \mathbf{y}_t)\}$ be a sequence of arm-pair played such that all arms $\mathbf{x} \in \{\mathbf{x}_\tau, \mathbf{y}_\tau\}_{\tau=1}^t$ belong to the ball of unit radius. Also suppose the initial exploration length t_0 be such that $\lambda_{\min}\Big(\sum_{\tau=1}^{t_0}(\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top\Big) \ge 1$. Then $\forall\, t > t_0$, $\sum_{\tau=t_0+1}^{t}\|(\mathbf{x}_\tau - \mathbf{y}_\tau)\|_{V_{\tau+1}^{-1}} \le \sqrt{2dt\log\left(\frac{4t_0+t}{d}\right)}$, where recall $V_{\tau+1} := \sum_{j=1}^{\tau}(\mathbf{x}_j - \mathbf{y}_j)(\mathbf{x}_j - \mathbf{y}_j)^\top$.*

Proof. As explained in Sec. C, our problem setup being a special case of GLM bandits, Lem. 1 follows directly from Lem. 2 of [26], with the additional consideration that in our case: (1). the generalized linear model is sigmoid function, (2). the subgaussianity parameter of the noise model is $\frac{1}{2}$, and (3). any arm $(\mathbf{x}_t - \mathbf{y}_t) \in \mathcal{D}' \subset R^d$ played at round t belong to a ball of radius 2. □

D.3 Proof of Lem. 2

Lemma 2 (Confidence Ellipsoid). *Suppose the initial exploration length t_0 be such that $\lambda_{\min}\Big(\sum_{\tau=1}^{t_0}(\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top\Big) \ge 1$, and κ is as defined in Thm. 3. Then for any $\delta > 0$, with probability at least $(1-\delta)$, for all $t > t_0$, $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\|_{V_t} \le \frac{1}{2\kappa}\sqrt{\frac{d}{2}\log\left(1 + \frac{2t}{d}\right) + \log\frac{1}{\delta}}$, where recall $V_{t+1} := \sum_{\tau=1}^{t}(\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$.*

Proof. Following a similar argument as described in Lem.1, the result follows directly from Lem. 3 of [26]. □

E Appendix for Sec. 3.2

Lemma 14 (Stagewise Sample Independence). *At any time $t \in [T]$, at any stage $s \in \lfloor \log T \rfloor$, and given an fixed realization of the played arm-pairs $\{\mathbf{x}_\tau, \mathbf{y}_\tau\}_{\tau \in \boldsymbol{\phi}^s}$, the corresponding preference outcomes $\{o_\tau\}_{\tau \in \boldsymbol{\phi}^s}$ are independent random variables with $\mathbf{E}[o_\tau] = \sigma\Big((\mathbf{x}_\tau - \mathbf{y}_\tau)^\top \boldsymbol{\theta}^*\Big)$.*

Proof. Note that at any stage $s \in \lfloor \log T \rfloor$ of any trial $t > t_0$, the time index t is added to $\boldsymbol{\phi}^s$ only if $\exists p_t^s(a_t, b_t) > \frac{1}{2^s}$. But note the value of $p_t^s(a_t, b_t)$ only depends on the other existing instances of $\boldsymbol{\phi}^s$, i.e. $\{\mathbf{x}_\tau, \mathbf{y}_\tau\}_{\tau \in \boldsymbol{\phi}^s}$ and not in $\{o_\tau\}_{\tau \in \boldsymbol{\phi}^s}$.

Moreover, the fact that both the items $\mathbf{x} \in \{\mathbf{x}_t, \mathbf{y}_t\}$ has survived till stage s, means they must have passed all earlier stages $\tilde{s} < s$ which relies on their previously estimated scores $g_t^{\tilde{s}}(\mathbf{x})$ and pairwise confidence bounds $p_t^{\tilde{s}}(\mathbf{x}, j)$, $\forall j \in \mathcal{G}_t^{\tilde{s}}$—but this only depends on the observations $\cup_{\tilde{s}<s}\{\mathbf{x}_\tau, \mathbf{y}_\tau, o_\tau\}_{\tau \in \boldsymbol{\phi}^{\tilde{s}}}$. And by the modelling assumption of our preference feedback, given $(\mathbf{x}_\tau, \mathbf{y}_\tau)$, $\mathbf{E}[o_\tau] = \sigma\Big((\mathbf{x}_\tau - \mathbf{y}_\tau)^\top \theta^*\Big)$. Hence the claim follows. □

E.1 Pseudocode for Stagewise-Adaptive-Duel

Algorithm 3 Stagewise-Adaptive-Duel (Sta′D)

input: Learning rate $\eta > 0$, exploration length $t_0 > 0$
init: Select t_0 pairs $\{(\mathbf{x}_\tau, \mathbf{y}_\tau)\}_{\tau\in[t_0]}$, each drawn at random from $\mathcal{S}_\tau$, and observe the corresponding preference feedback $\{o_\tau\}_{\tau\in[t_0]}$
$S \leftarrow \lfloor \log T \rfloor$, $\boldsymbol{\phi}^s \leftarrow [t_0]$, $\forall s \in [\lfloor \log T \rfloor]$
Set $V_{t_0+1} := \sum_{\tau=1}^{t_0} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$
while $t \le T$ **do**
  $s \leftarrow 1, \mathcal{G}^1 \leftarrow [K]$
  repeat
    Compute the MLE estimate on $\boldsymbol{\phi}^s$, i.e. solve for $\hat{\boldsymbol{\theta}}_t^s$ s.t.:
$$\sum_{\tau\in\boldsymbol{\phi}^s} \Big(o_\tau - \sigma\big((\mathbf{x}_\tau - \mathbf{y}_\tau)^\top \hat{\boldsymbol{\theta}}_t^s\big)\Big)(\mathbf{x}_\tau - \mathbf{y}_\tau) = \mathbf{0}$$
    Set: $V_t^s = \sum_{\tau\in\boldsymbol{\phi}^s} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$
    Compute: $g_t^s(i) = \hat{\boldsymbol{\theta}}_t^{s\top} \mathbf{x}_i^t$, $\forall i \in \mathcal{G}^s$, and $p_t^s(i,j) = \eta \|\mathbf{x}_i^t - \mathbf{x}_j^t\|_{V_t^{s-1}}$, $\forall i,j \in \mathcal{G}^s$
    if $p_t^s(i,j) \le \frac{1}{\sqrt{T}}$, $\forall i,j \in \mathcal{G}^s$ **then**
      $a_t \leftarrow \arg\max_{a\in\mathcal{G}^s} g_t^s(a)$
      $b_t \leftarrow \arg\max_{b\in\mathcal{G}^s} \big(g_t^s(b) + p_t^s(b, a_t)\big)$
      Set $\mathbf{x}_t = \mathbf{x}_{a_t}^t$, $\mathbf{y}_t = \mathbf{x}_{b_t}^t$
    else if $p_t^s(i,j) \le \frac{1}{2^s}$, $\forall i,j \in \mathcal{G}^s$ **then**
      Find $\mathcal{B}_t^s := \{i \in \mathcal{G}^s \mid \exists j \in \mathcal{G}^s \text{ s.t. } g_t^s(i) + \frac{1}{2^s} < g_t^s(j)\}$
      Update $\mathcal{G}^{s+1} \leftarrow \mathcal{G}^s \setminus \mathcal{B}^s$, $s \leftarrow s+1$
    else
      Choose any pair $a_t, b_t \in \mathcal{G}^s$ s.t. $p_t^s(a_t, b_t) > \frac{1}{2^s}$.
      Set: $\boldsymbol{\phi}^s \leftarrow \boldsymbol{\phi}^s \cup \{t\}$, $\mathbf{x}_t = \mathbf{x}_{a_t}^t$, $\mathbf{y}_t = \mathbf{x}_{b_t}^t$
    end if
  until a pair $(\mathbf{x}_t, \mathbf{y}_t)$ is found
  Play $(\mathbf{x}_t, \mathbf{y}_t)$, and update $t \leftarrow t+1$
end while

Using Lem. 14, one can derive sharper concentration bounds on the pairwise-arm scores as proved below:

E.2 Proof of Lem. 4

Lemma 4 (Sharper Concentration of Pairwise Scores)**.** *Consider any $\delta > 0$, and suppose we set the parameters of Stagewise-Adaptive-Duel (Alg. 3) as $\eta = \frac{3}{2\kappa}\sqrt{2\log\frac{3TK}{\delta}}$, where $\kappa := \inf_{\|\mathbf{x}-\mathbf{y}\|\le 2, \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|\le 1} \Big[\sigma'\big((\mathbf{x}-\mathbf{y})^\top \hat{\boldsymbol{\theta}}\big)\Big]$, and $t_0 = 2\left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(2/\delta)}}{\lambda_{\min}(B)}\right)^2 + \frac{4\Lambda}{\lambda_{\min}(B)}$, where $\Lambda = \frac{8}{\kappa^4}\big(d^2 + \log\frac{3}{\delta}\big)$ and $B = \mathbf{E}_{\mathbf{x},\mathbf{y} \overset{iid}{\sim} \mathcal{P}_\mathcal{D}}[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^\top]$ (for some universal problem independent constants $C_1, C_2 > 0$). Then with probability at least $(1-\delta)$, for all stages $s \in \lceil \log T \rceil$ at all rounds $t > t_0$ and for all index pairs $i, j \in \mathcal{G}^s$ of round t: $|(\mathbf{x}_i^t - \mathbf{x}_j^t)^\top(\theta^* - \theta_t^s) \le p_t^s(i,j)|$.*

Proof. The first thing to note is that due to Lem. 13, our choice of the length of initial exploration phase t_0 ensures that with probability at least $(1 - \frac{\delta}{2})$, we have $\lambda_{\min}(V_{t_0+1}) \ge \frac{8}{\kappa^4}\Big(d^2 + \log\frac{3}{\delta}\Big)$.

Now recall the finite samples classical asymptotic normality of MLE estimates of GLM distributions (see Thm. 1 of [26]). Then if $\hat{\boldsymbol{\theta}}_t$ is the MLE estimate of t independent random samples from any GLM model $\{Y_\tau\}_{\tau\in[t]}$ against the corresponding instance set $\{X_\tau\}_{\tau\in[t]}$, for any $\mathbf{x} \in \mathbb{R}^d$ and any $\delta > 0$, with probability at least $(1 - 3\delta)$,

$$|\mathbf{x}^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)| \le \frac{3\gamma}{\kappa}\Big(\sqrt{\log\frac{1}{\delta}}\|\mathbf{x}\|_{V_{t+1}^{-1}}\Big),$$

whenever t is such that $\lambda_{\min}(V_{t+1}) \geq \frac{512M^2\alpha^2}{\kappa^4}\bigg(d^2 + \log\frac{1}{\delta}\bigg)$, M being the upper bound of the second order derivative of the GLM link function and γ being the sub-Gaussianity parameter of the noise model, $V_{t+1} = \sum_{\tau=1}^{t} X_\tau X_\tau^\top$.

Now for specific case when the GLM link function turms out to be *the logistic / sigmoid function* $\sigma(\cdot)$ we have $M = \frac{1}{4}$, and for bernoulli noise the sub-Gaussian parameter $\gamma = \frac{1}{2}$. Thus for a GLM model with logistic link and Bernoulli noise, we now have that for any $\mathbf{x} \in \mathbb{R}^d$, with probability at least $1-\delta$,

$$|\mathbf{x}^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)| \leq \frac{3}{2\kappa}\Big(\sqrt{\log\frac{3}{\delta}}\|\mathbf{x}\|_{V_{t+1}^{-1}}\Big), \tag{2}$$

whenever $\lambda_{\min}(V_{t+1}) \geq \frac{8}{\kappa^4}\bigg(d^2 + \log\frac{3}{\delta}\bigg)$.

So the coming back to our setting of algorithm Stagewise-Adaptive-Duel (Alg. 3), first note that our choice of t_0 already ensures that with probability at least $1 - \frac{\delta}{2}$ we have:

$$\lambda_{\min}(V_{t_0+1}) \geq \frac{8}{\kappa^4}\Big(d^2 + \log\frac{3}{\delta}\Big). \tag{3}$$

Then combining the result from Eqn. (2) along with the independent samples guarantee derived from Lem. 14, and owing to the connection of our preference feedback model to GLM models (as explained in Sec. C), we further have that for at any stage $s \in \lceil \log T \rceil$, at any round $t > t_0$ for any index-pair $i, j \in G^s$, denoting $\mathbf{z}_s^t(ij) = \mathbf{x}_i^t - \mathbf{x}_j^t$, with probability at least $1 - \frac{\delta}{2TK(K-1)\lceil \log T \rceil}$,

$$|(\mathbf{z}_s^t(ij))^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)| \leq \frac{3}{2\kappa}\Big(\sqrt{\log\frac{6TK(K-1)\lceil \log T \rceil}{\delta}}\|\mathbf{z}_s^t(ij)\|_{V_{t+1}^{-1}}\Big), \tag{4}$$

as our choice of initial exploration length t_0 already ensures $\lambda_{\min}(V_t^s) \geq \frac{8}{\kappa^4}\bigg(d^2 + \log\frac{3}{\delta}\bigg)$. Now taking union bound over all round $t \in T \setminus [t_0]$, all stages $s \in \lceil \log T \rceil$ and pairs $i, j \in \mathcal{G}^s$, $i \neq j$ we get that:

$$Pr\Bigg(\forall i, j \in \mathcal{G}^s, s \in \lceil \log T \rceil \text{ of all round } t \in T \setminus [t_0],\ |(\mathbf{z}_s^t(ij))^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)| \leq \frac{3}{2\kappa}\Big(\sqrt{2\log\frac{3TK}{\delta}}\|\mathbf{z}_s^t(ij)\|_{V_{t+1}^{-1}}\Big)\Bigg) > 1 - \frac{\delta}{2}, \tag{5}$$

upon noting for any stage $s \in \lceil \log T \rceil$, $|\mathcal{G}^s| \leq K$ and $6TK(K-1)\lceil \log T \rceil \leq (3TK)^2$. The result finally follows by taking a union bound over the two events of Eqn. (3) and (5). $\square$

E.3 Proof of Lem. 6

Lemma 6. *For any $t > t_0$, suppose the pair $(\mathbf{x}_t, \mathbf{y}_t)$ is chosen at stage s_t, and i_t^* denotes the best action of round t, i.e. $\mathbf{x}_{i_t^*}^t = \mathbf{x}_t^* = \arg\max_{\mathbf{x} \in \mathcal{S}_t} \mathbf{x}^\top\boldsymbol{\theta}^*$. Then for any $\delta \in (0,1)$, with probability at least $(1-\delta)$, for all $t > t_0$: $i_t^* \in \mathcal{G}^{s_t}$ and for both $\mathbf{x} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, $g(\mathbf{x}_t^*) - g(\mathbf{x}) \leq \begin{cases} \frac{2}{\sqrt{T}} & \text{if } t \in \boldsymbol{\phi}^c \\ \frac{4}{2^{s_t}} & \text{otherwise} \end{cases}$.*

Proof. Let us consider the event $\mathcal{E} = \{\forall i, j \in \mathcal{G}^s, s \in \lceil \log T \rceil \text{ of all round } t \in T \setminus [t_0],\ |(\mathbf{x}_i^t - \mathbf{x}_j^t)^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)| \leq p_t^s(ij)\}$. The rest of the proof will assume $\mathcal{E}$ to be true which holds good with probability at least $(1-\delta)$ as proved in Lem. 4. We will now prove the lemma breaking it into three parts:

Part-1: We first prove that for all $t > t_0$: $i_t^* \in \mathcal{G}^{s_t}$ following a recursive argument.

For any $t > t_0$, first note that if $s_t = 1$ (first phase) then obviously $i_t^* \in \mathcal{G}^{s_t}$ as by initialization $\mathcal{G}^{s_t} = [K]$. Now for any $s_t > 1$, if $i_t^* \notin \mathcal{G}^{s_t}$ then it must have got eliminated by some phase say $s < s_t$. But that means, at phase s there was an item with index say $j \in \mathcal{G}^s \setminus \{i_t^*\}$ such that $g_t^s(j) > g_t^s(i_t^*) + \frac{1}{2^s}$.

With slight abuse of notation, let us denote for any index $i \in \mathcal{G}^s$ its true score as $g_t^*(i) := g(\mathbf{x}_i^t) = \mathbf{x}_i^{t\top}\theta^*$. Recall that $g_t^*(i) = g(\mathbf{x}_i^t) = \mathbf{x}_i^{t\top}\theta^*$, and $g_t^s(i) = \mathbf{x}_i^{t\top}\theta_t^s$. Let us also denote for any index pair $i, j \in \mathcal{G}^s$, their estimated score difference $d_t^s(i,j) := g_t^s(i) - g_t^s(j)$, and true pairwise score difference $d_t^*(i,j) := g_t^*(i) - g_t^*(j)$. So by definition $d_t^*(i_t^*, i) > 0$, $\forall i \in \mathcal{S}_t \setminus \{i_t^*\}$. So in particular $d_t^*(i_t^*, j) > 0$ as well.

But since both i_t^* and j have passed stage s and we assume the event $\mathcal{E}$ to be true, from Lem. 4 we have that $|d_t^s(i_t^*, j) - d_t^*(i_t^*, j)| \le p_t^s(i_t^*, j) \le \frac{1}{2^s}$. But this further implies

$$d_t^s(i_t^*, j) \ge d_t^*(i_t^*, j) - \frac{1}{2^s} > -\frac{1}{2^s} \implies g_t^s(i_t^*) \ge g_t^s(j) - \frac{1}{2^s},$$

which gives a contradiction as for i_t^* to get eliminated at stage s, we earlier assumed $g_t^s(j) > g_t^s(i_t^*) + \frac{1}{2^s}$. So i_t^* must be present at stage s_t.

Part-2: We now prove that for both $\mathbf{x} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, $g(\mathbf{x}_t^*) - g(\mathbf{x}) \le \frac{2}{\sqrt{T}}$, if $t \in \boldsymbol{\phi}^c$.

Recall $\mathbf{x}_t = \mathbf{x}_{a_t}^t$ and $\mathbf{y}_t = \mathbf{x}_{b_t}^t$. We would only consider the cases $a_t \ne i_t^*$ and $b_t \ne i_t^*$, as the claim is trivially true otherwise.

First let us analyse the case for $a_t \ne i_t^*$, by our arm selection strategy this means $d_t^s(a_t, i_t^*) > 0$ since both i_t^* and a_t are present at s_t. Also as $t \in \boldsymbol{\phi}^c$, and we assume the event $\mathcal{E}$ to be true, from Lem. 4 we have $|d_t^s(i_t^*, a_t) - d_t^*(i_t^*, a_t)| \le p_t^s(i_t^*, a_t) \le \frac{1}{\sqrt{T}}$. But this further implies

$$d_t^*(a_t, i_t^*) \ge d_t^s(a_t, i_t^*) - \frac{1}{\sqrt{T}} > -\frac{1}{\sqrt{T}} \implies g_t^s(a_t) \ge g_t^s(i_t^*) - \frac{1}{\sqrt{T}},$$

Now if $a_t = b_t$, then the claim follows from the earlier bound itself. Assuming $a_t \ne b_t$ and $b_t \ne i_t^*$, once again by our arm selection strategy this implies $g_t^s(b_t) + p_t^s(b_t, i_t^*) > g_t^s(i_t^*) \implies g_t^s(b_t) > g_t^s(i_t^*) - \frac{1}{\sqrt{T}}$. Also since $t \in \boldsymbol{\phi}^c$, and we assume the event $\mathcal{E}$ to be true, from Lem. 4 we have $|d_t^s(i_t^*, a_t) - d_t^*(i_t^*, b_t)| \le p_t^s(i_t^*, b_t) \le \frac{1}{\sqrt{T}}$, which further implies

$$d_t^*(b_t, i_t^*) \ge d_t^s(b_t, i_t^*) - \frac{1}{\sqrt{T}} > -\frac{2}{\sqrt{T}} \implies g_t^s(b_t) \ge g_t^s(i_t^*) - \frac{2}{\sqrt{T}}.$$

This validates the claim of this part.

Part-3: Finally in this part we show that for both $\mathbf{x} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, $g(\mathbf{x}_t^*) - g(\mathbf{x}) \le \frac{8}{2^{s_t}}$, if $t \in [T] \setminus \boldsymbol{\phi}^c$

Assuming any stage $s_t \in \lfloor \log T \rfloor$, if $\mathbf{x}_t$ and $\mathbf{y}_t$ has survived till s_t this means they were not eliminated by $\mathbf{x}_t^*$ at any stage $s < s_t$, as by the claim of **Part-1** $\mathbf{x}_t^*$ survives till stage s_t as well.

Let us first prove the claim for $\mathbf{x}_t = \mathbf{x}_{a_t}^t$. Since its corresponding index a_t did not get eliminated at stage $s_t - 1$ this implies $d_t^s(a_t, i_t^*) > \frac{1}{2^{s_t-1}}$. Moreover as we assume the event $\mathcal{E}$ to be true, from Lem. 4 we have $|d_t^s(i_t^*, a_t) - d_t^*(i_t^*, b_t)| \le \frac{2}{2^{s_t}}$, which further implies

$$d_t^*(a_t, i_t^*) \ge d_t^s(a_t, i_t^*) - \frac{2}{2^{s_t}} \implies g_t^*(a_t) \ge g_t^*(i_t^*) - \frac{4}{2^{s_t}},$$

which proves the claim for $\mathbf{x}_t$. The same claim for $\mathbf{y}_t$ can be proved as well by simply following the exact same chain of arguments shown for $\mathbf{x}_t$. □

E.4 Proof of Lem. 7

Lemma 7. *At any stage $s \in \lfloor \log T \rfloor$, with probability at least $(1-\delta)$,* $\sqrt{|\boldsymbol{\phi}^s|} \le \eta 2^s \sqrt{2d \log\left(\frac{4t_0 T}{d}\right)}$.

Proof. Firstly note that due to Lem. 1, at any state $s \in \lfloor \log T \rfloor$ of round T

$$\sum_{\tau \in \boldsymbol{\phi}^s} \|(\mathbf{x}_\tau - \mathbf{y}_\tau)\|_{(V_T^s)^{-1}} \le \sqrt{2d|\boldsymbol{\phi}^s| \log\left(\frac{4t_0 + |\boldsymbol{\phi}^s|}{d}\right)} \le \sqrt{2d|\boldsymbol{\phi}^s| \log\left(\frac{4Tt_0}{d}\right)},$$

since our choice of t_0 already ensures $\lambda_{\min}(V_T^s) \ge 1$ (noting by definition $\kappa < 1$), the second inequality follows from the fact that by definition $t_0, |\boldsymbol{\phi}^s| \ge 1$ (claim holds trivially if $\boldsymbol{\phi}^s = \emptyset$) and also $|\boldsymbol{\phi}^s| \le T$.

Recalling that a_t, b_t respectively denotes the index of the played pair $\mathbf{x}_t, \mathbf{y}_t$ at any round $t > t_0$, above further implies

$$\sum_{\tau \in \boldsymbol{\phi}^s} p_\tau^s(a_\tau, b_\tau) \le \eta \sqrt{2d|\boldsymbol{\phi}^s| \log\left(\frac{4Tt_0}{d}\right)}. \tag{6}$$

But on the other hand, by the construction of sets $\boldsymbol{\phi}^s$, we have: $\sum_{\tau \in \boldsymbol{\phi}^s} p_\tau^s(a_\tau, b_\tau) \ge \frac{|\boldsymbol{\phi}^s|}{2^s}$.

Then combining above with Eqn. (6) we have: $\frac{|\boldsymbol{\phi}^s|}{2^s} \le \sum_{\tau \in \boldsymbol{\phi}^s} p_\tau^s(a_\tau, b_\tau) \le \eta \sqrt{2d|\boldsymbol{\phi}^s| \log\left(\frac{4Tt_0}{d}\right)}$,

which finally implies $\sqrt{\boldsymbol{\phi}^s} \le \eta 2^s \sqrt{2d \log\left(\frac{4Tt_0}{d}\right)}$, and the claim follows. □

E.5 Proof of Thm. 5

Theorem 5 (Regret bound of Stagewise-Adaptive-Duel (Alg. 3))**.** *Consider we set t_0 and η as per Lem. 4. Then for any $\delta > 0$, with probability at least $(1-\delta)$, the regret of Alg. 3 can be bounded as:*

$$R_T \le t_0 + 4\eta\sqrt{2d\log\left(\frac{4t_0T}{d}\right)}\sqrt{T\log T} + 2\sqrt{T} = O\Bigg(\frac{\sqrt{dT\log T}}{\kappa}\sqrt{\log\left(\frac{TK}{\delta}\right)\log\left(\frac{Td}{\kappa}\log\frac{1}{\delta}\right)}\Bigg)$$

Proof. Suppose we denote by $\boldsymbol{\phi}^c := \{t \in [T] \setminus [t_0] \mid t \notin \cup_{s=1}^{\lfloor \log T \rfloor} \boldsymbol{\phi}^s\}$ the set of all good time intervals where all the index pairs $p_t^s(i,j)$ are estimated within the confidence accuracy $\frac{1}{\sqrt{T}}$. The proof crucially relies on the concentration bound of Lem. 4, from which we first derive Lem. 6. Further owing to Lem. 1 and due to the construction of our *'stagewise-good item pairs'* we also derive another main result: Lem. 7.

The final regret bound now follows clubbing the results of Lem. 6 and 7 as given below:

$$\begin{aligned}
R_t = \sum_{t=1}^{T} r_t &= \sum_{t=1}^{t_0} r_t + \sum_{s=1}^{\lfloor \log T \rfloor} \sum_{t \in \boldsymbol{\phi}^s} r_t + \sum_{t \in \boldsymbol{\phi}^c} r_t \\
&\overset{(a)}{\le} t_0 + \sum_{s=1}^{\lfloor \log T \rfloor} |\boldsymbol{\phi}^s| \frac{4}{2^s} + |\boldsymbol{\phi}^c| \frac{2}{\sqrt{T}} \\
&\overset{(b)}{\le} t_0 + 4 \sum_{s=1}^{\lfloor \log T \rfloor} \frac{2^s \eta \sqrt{2d|\boldsymbol{\phi}^s|}}{2^s} \sqrt{\log\left(\frac{4t_0T}{d}\right)} + 2\sqrt{T} \\
&\le t_0 + 4\eta\sqrt{2d\log\left(\frac{4t_0T}{d}\right)} \sum_{s=1}^{\lfloor \log T \rfloor} \sqrt{|\boldsymbol{\phi}^s|} + 2\sqrt{T} \\
&\overset{(c)}{\le} t_0 + 4\eta\sqrt{2d\log\left(\frac{4t_0T}{d}\right)}\sqrt{T\log T} + 2\sqrt{T}
\end{aligned}$$

$$= O\left(\frac{\sqrt{dT\log T}}{\kappa}\sqrt{\log\left(\frac{TK}{\delta}\right)\log\left(\frac{Td}{\kappa}\log\frac{1}{\delta}\right)}\right)$$

where recall that $\boldsymbol{\phi}^c := \{t \in [T] \setminus [t_0] \mid t \notin \cup_{s=1}^{\lfloor \log T \rfloor} \boldsymbol{\phi}^s\}$. We consider the trivial bound of $r_t = 1$ for the initial t_0 rounds. Note that here the inequality (a) follows from Lem. 6, (b) from Lem. 7 and since $\boldsymbol{\phi}^c \leq T$. Inequality (c) uses Cauchy-Schwartz along with the fact that $\cup_{s=1}^{\lfloor \log T \rfloor} \boldsymbol{\phi}^s \leq T$. Finally the order of the regret bound follows by considering our particular choice of η, t_0 and rearranging the terms. $\square$

F Appendix for Sec. 4

F.1 Proof of Lem. 8

Lemma 8 (Reducing $\mathcal{I}^{clb}$ with *Gumbel noise* to $\mathcal{I}^{cdb}$)**.** *There exists a reduction from the $\mathcal{I}^{clb}$ problem (under Gumbel noise, i.e. $\varepsilon_t \overset{iid}{\sim}$ Gumbel$(0,1)$) to $\mathcal{I}^{cdb}$ which preserves the expected regret.*

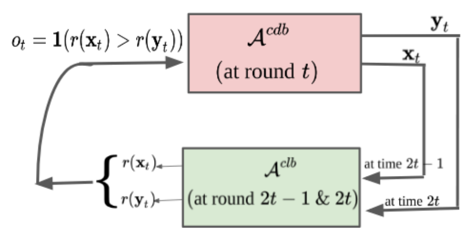

Figure 6: Pictorial demonstration of the reduction: Reducing $\mathcal{I}^{clb}$ to $\mathcal{I}^{cdb}$

Proof. Suppose we have a blackbox algorithm for the instance of $\mathcal{I}^{cdb}$ problem, say $\mathcal{A}^{cdb}$. To prove the claim, our goal is to show that this can be used to solve the $\mathcal{I}^{clb}$ problem where the underlying stochastic noise, ϵ_t at round t, is generated from a Gumbel$(0,1)$ distribution [39, 9]: Precisely we can construct an algorithm for $\mathcal{I}^{clb}(\boldsymbol{\theta}^*, K, T)$ (say $\mathcal{A}^{clb}$) using $\mathcal{A}^{cdb}$ as shown in Alg. 2.

The reduction now follows from Lem. 9 which establish the first half of the claim as it precisely shows a reduction of $\mathcal{I}^{clb}$ to $\mathcal{I}^{cdb}$. The second half of the claim is easy to follow from the corresponding regret definitions of the $\mathcal{I}^{clb}$ and $\mathcal{I}^{cdb}$ problem, Eqn. (4) and (1) respectively: Precisely owing to the reduction on Lem. 9, for any fixed T, $2R_T^{cdb} = R_{2T}^{clb}$. $\square$

F.2 Proof of Lem. 9

Lemma 9. *If $\mathcal{A}^{clb}$ rums on a problem instance $\mathcal{I}^{clb}(\boldsymbol{\theta}^*, K, 2T)$ with Gumbel$(0,1)$ noise, then the internally the algorithm $\mathcal{A}^{cdb}$ runs on a problem instance of $\mathcal{I}^{cdb}(\boldsymbol{\theta}^*, K, T)$.*

Proof. Firstly it is easy to note from the construction of $\mathcal{A}^{clb}$ that one round of $\mathcal{A}^{cdb}$, say round $t \in \lceil \frac{T}{2} \rceil$, goes in two consecutive rounds of $\mathcal{A}^{clb}$, round $2t-1$ and $2t$ of $\mathcal{A}^{cdb}$.

We now show the main claim that by construction of $\mathcal{A}^{clb}$, the internal world of $\mathcal{A}^{cdb}$ indeed receives feedback from an instance of $\mathcal{I}^{cdb}(\boldsymbol{\theta}^*, K, T)$: Precisely, recalling our feedback model for any problem instance of $\mathcal{I}^{cdb}(\boldsymbol{\theta}^*, K, T)$ from Sec. 2.1, we want to establish the following claim:

Claim: At any round $t \in \lfloor \frac{T}{2} \rfloor$ of $\mathcal{A}^{cdb}$, $o_t = \mathbf{1}(\mathbf{x}_t$ preferred over $\mathbf{y}_t) \sim Ber\big(\sigma\big(\mathbf{x}_t - \mathbf{y}_t\big)^\top \boldsymbol{\theta}^*\big)\big)$.

Towards this note that by construction of $\mathcal{A}^{cdb}$ we have $o_t = \mathbf{1}(r(\mathbf{x}_t) > r(\mathbf{y}_t))$. Now by the setting of any problem instance $\mathcal{I}^{cdb}(\boldsymbol{\theta}^*, K, T)$ with iid Gumbel$(0,1)$ noise, note that given any $\mathbf{x} \in \mathbb{R}^d$, $r(\mathbf{x}) \sim$ Gumbel$(\mathbf{x}^\top \boldsymbol{\theta}^*, 1)$ [40]. But then given arm pair $\mathbf{x}_t$ and $\mathbf{y}_t$ and by defining $Z_t = \max(r(\mathbf{x}_t), r(\mathbf{y}_t))$, by the property of max of two independent Gumbel distributions [9, 34]:

$$Pr(Z_t = \mathbf{x}_t \mid \{\mathbf{x}_t, \mathbf{y}_t\}) = \frac{e^{\mathbf{x}_t^\top \boldsymbol{\theta}^*}}{e^{\mathbf{x}_t^\top \boldsymbol{\theta}^*} + e^{\mathbf{y}_t^\top \boldsymbol{\theta}^*}} = \frac{1}{1 + e^{(\mathbf{x}_t - \mathbf{y}_t)^\top \boldsymbol{\theta}^*}} = \sigma((\mathbf{x}_t - \mathbf{y}_t)^\top \boldsymbol{\theta}^*).$$

The result now follows noting $o_t = 1$, if $Z_t = \mathbf{x}_t$ and $o_t = 0$, if $Z_t = \mathbf{y}_t$, implying $o_t \sim \text{Ber}\big(\sigma\big(\mathbf{x}_t - \mathbf{y}_t\big)^\top \boldsymbol{\theta}^*\big)$. $\square$

F.3 Proof of Thm. 10

Theorem 10 (SPM($\boldsymbol{\theta}^*, d, 2$): Regret Lower Bound)**.** *For any algorithm $\mathcal{A}^{cdb}$ for the problem of linear-score based stochastic K-armed contextual dueling bandit of dimensional-d, there exists a sequence of d-dimensional context sets $\{\mathbf{x}_1^t, \ldots \mathbf{x}_K^t\}_{t=1}^T$ and a constant $\gamma > 0$ such that the regret incurred by $\mathcal{A}^{cdb}$ on T rounds is at least $\Omega(\frac{\gamma}{2}\sqrt{2dT})$, i.e. $R_T(\mathcal{A}^{cdb}) \geq \frac{\gamma}{2}\sqrt{2dT}$, for any $T \geq d^2$.*

Proof. The proof immediately follows from the known regret lower bound for of K-armed d-dimensional contextual linear bandits problem (see Thm. 2 of [13]), and from the fact that for any T, $2R_T^{cdb} = R_{2T}^{clb}$ as we proved in Lem. 8: This is because any smaller regret for $\mathcal{A}^{cdb}$ would violate the best achievable regret bound of $\mathcal{A}^{clb}$ which is a logical contradiction as this would imply $R_{2T}^{clb} = 2R_T^{cdb} < \gamma\sqrt{2dT}$. So it must be the case that $R_T^{cdb} \geq \frac{\gamma}{2}\sqrt{2dT}$. $\square$

G Appendix to Sec. 5

G.1 Proof of Thm. 11

Theorem 11 (Regret Lower Bound (Subsetwise Preferences))**.** *Given any $q \geq 2$ and $d > 1$, for any algorithm $\mathcal{A}$ for the problem of stochastic K-armed d-dimensional linear contextual bandit with SPM($\boldsymbol{\theta}, d, q$) feedback model, there exists a sequence of d-dimensional context sets $\{\mathbf{x}_1^t, \ldots \mathbf{x}_K^t\}_{t=1}^T$ and a choice of $\boldsymbol{\theta}$ such that the regret incurred by $\mathcal{A}$ on T rounds is at least $\Omega(\sqrt{dT})$.*

Remark 2. *The interesting part here to note is that when $q = 2$ (the dueling bandit/pairwise preference case), we already know the lower bound is $\Omega(\sqrt{dT})$ from (Thm. 10). So whatever is possible to achieve for the dueling feedback is also achievable with the q-subsetwise feedback simply using $q = 2$. The question really to ask here is if its possible to achieve an faster learning rate (smaller regret) with general q-subsetwise queries. However, our lower bound result clearly shows the impossibility of any such hope.*

Proof. The proof relies on constructing a 'hard enough' problem instance for the learning framework and showing no algorithm can achieve a smaller rate of regret on that instance than the claimed lower bounds. We assume the decision space $\mathcal{D}$ to be the set of standard basis vectors: $\mathcal{D} = \{\mathbf{e}_1, \ldots \mathbf{e}_d\}$, and at each round t, the learner receives the set of context vectors $\mathcal{S}_t = \mathcal{D}$. Now let us construct $d+1$ problem instances $\mathcal{I}^1, \mathcal{I}^2, \ldots, \mathcal{I}^d$ and $\mathcal{I}^0$, where each instance is uniquely identified by its underlying score vector $\boldsymbol{\theta}$ as defined below:

Base Problem Instance ($\mathcal{I}^0$): $\boldsymbol{\theta}^0(i) = 0.5,\ \forall i \in [d]$. Now let us consider d alternative problem instances $\mathcal{I}^m\ \ \forall m \in [d]$:

Problem Instance-m ($\mathcal{I}^m$): For all $t \in [T]$, $\boldsymbol{\theta}^m(i) = 0.5,\ \forall i \in [d] \setminus \{m\},\ \boldsymbol{\theta}^m(m) = 0.5 + \epsilon$, for some $\epsilon \in (0, 0.2]$.

Clearly, for instance $\mathcal{I}^m$, the unique *'best'* item is $i_*^m := m$, and rest of all the $d-1$ items are *'bad'* playing which at any round $t \in [T]$ yields a regret of $\epsilon/|\mathcal{X}_t|$.

Let $N_T(S) := \mathbf{E}[\sum_{t=1}^T \mathbf{1}(S = \mathcal{X}_t)]$, denotes the expected number of times $\mathcal{A}$ pulls the subset $S \subseteq \mathcal{D}$ in T rounds. One key remark before proceeding to the main analysis is: We consider *only the class of all deterministic algorithms*, i.e. where $\mathcal{X}_t$ is a deterministic function of the past history $\mathcal{H}_{t-1} := \{\mathcal{S}_1, \mathcal{X}_1, i_1, \ldots \mathcal{S}_{t-1}, \mathcal{X}_{t-1}, i_{t-1}, \mathcal{S}_t\}$. Note this is without loss of generality, since any randomized strategy can be seen as a randomization over deterministic querying strategies. Thus, a lower bound which holds uniformly for any deterministic class of algorithms, would also hold over a randomized class of algorithms.

Note since for any instance $\mathcal{I}^m$, $m \in [d]$, the regret in T round can be lower bounded as:

$$\mathbf{E}_{\mathcal{I}^m}[R_T(\mathcal{A})] \geq \sum_{t=1}^T \left(\mathbf{E}_{\mathcal{I}^m}\left[\mathbf{1}(\mathcal{X}_t \cap [d] \setminus \{m\} \neq \emptyset)\right] \frac{\epsilon(|\mathcal{X}_t| - 1)}{|\mathcal{X}_t|} \right)$$

$$\geq \sum_{t=1}^{T} \frac{\epsilon}{2}\Big(T - \mathbf{E}_{\mathcal{I}^m}\big[\mathbf{1}(\mathcal{X}_t = \{i_*^m\})\big]\Big) = \frac{\epsilon}{2}\big(T - \mathbf{E}_{\mathcal{I}^m}[N_T(i_*^m)]\big).$$

Then taking average over $\mathcal{I}^m$s for all $m \in [d]$:

$$\mathbf{E}[R_T(\mathcal{A})] = \sum_{m\in[d]} \frac{\mathbf{E}_{\mathcal{I}^m}[R_T(\mathcal{A})]}{d} \geq \frac{\epsilon}{2}\bigg(T - \frac{\sum_{m\in[d]} \mathbf{E}_{\mathcal{I}^m}[N_T(m)]}{d}\bigg) \tag{7}$$

since $i_*^m = m$. Now note that:

$$\begin{aligned}
&\mathbf{E}_{\mathcal{I}^m}[N_T(m)] - \mathbf{E}_{\mathcal{I}^0}[N_T(m)] \\
&= \sum_{t=1}^{T} \big(Pr_{\mathcal{I}^m}(\mathcal{X}_t = \{m\}) - Pr_{\mathcal{I}^0}(\mathcal{X}_t = \{m\})\big) \leq T.D_{TV}(\mathcal{I}^0, \mathcal{I}^m),
\end{aligned} \tag{8}$$

where $D_{TV}(\mathcal{I}^0, \mathcal{I}^m)$ denotes the total variation distance between the probability distribution of $\mathcal{I}^0$ and $\mathcal{I}^m$ with respect to history $\mathcal{H}_T$, i.e. $D_{TV}(\mathcal{I}^0, \mathcal{I}^m) := \sup_{\mathcal{E}\in\mathcal{H}_T} |Pr_{\mathcal{I}^0}(\mathcal{E}) - Pr_{\mathcal{I}^m}(\mathcal{E})|$.

Further using Pinksker's inequality we have

$$D_{TV}(\mathcal{I}^0, \mathcal{I}^m) \leq \sqrt{\frac{1}{2} D_{KL}(\mathcal{I}^0, \mathcal{I}^m)}, \tag{9}$$

where $D_{KL}(\mathcal{I}^0, \mathcal{I}^m)$ denotes the KL-divergence between the probability distribution induced on the observed history $\mathcal{H}_T$ by the problem instance $\mathcal{I}^0$ and $\mathcal{I}^m$. Thus averaging over $\mathcal{I}^m$s for all $m \in [d]$:

$$\begin{aligned}
&\frac{\sum_{m\in[d]} \mathbf{E}_{\mathcal{I}^m}[N_T(m)]}{d} \leq \frac{\sum_{m\in[d]} \big(\mathbf{E}_{\mathcal{I}^0}[N_T(m)] + T.D_{TV}(\mathcal{I}^0, \mathcal{I}^m)\big)}{d} \\
&= \frac{\sum_{m\in[d]} \mathbf{E}_{\mathcal{I}^0}[N_T(m)]}{d} + T \sum_{m\in[d]} \frac{1}{d}\Big(\sqrt{\frac{1}{2} D_{KL}(\mathcal{I}^0, \mathcal{I}^m)}\Big) \\
&= \frac{\sum_{m\in[d]} \mathbf{E}_{\mathcal{I}^0}[N_T(m)]}{d} + T\sqrt{\bigg(\frac{1}{2d}\sum_{m\in[d]} D_{KL}(\mathcal{I}^0, \mathcal{I}^m)\bigg)}
\end{aligned} \tag{10}$$

With slight abuse of notation, by denoting $\mathcal{I}_t^0 := Pr_{\mathcal{I}^0}\big(\mathbf{P}_t(\mathcal{X}_t) \mid \mathcal{H}_{t-1}\big)$ and $\mathcal{I}_t^m := Pr_{\mathcal{I}^m}\big(\mathbf{P}_t(\mathcal{X}_t) \mid \mathcal{H}_{t-1}\big)$, we note that:

$$D_{KL}(\mathcal{I}_t^0, \mathcal{I}_t^m) \sim \begin{cases} KL\big(\text{Mult}(\boldsymbol{\theta}^0, \mathcal{X}_t), \text{Mult}(\boldsymbol{\theta}^m, \mathcal{X}_t)\big), \text{ if } m \in \mathcal{X}_t \text{ and } |\mathcal{X}_t| > 1 \\ 0, \text{ otherwise} \end{cases},$$

where $\text{Mult}(\boldsymbol{\theta}^0, \mathcal{X}_t)$ and $\text{Mult}(\boldsymbol{\theta}^m, \mathcal{X}_t)$ respectively denote the multinomial distributions ($Pr(\cdot \mid \mathcal{X}_t, \boldsymbol{\theta})$) induced by the instances $\mathcal{I}^0$ and $\mathcal{I}^m$ on the elements of $\mathcal{X}_t$. Let us also denote by $S^{(m)} := \{S \subseteq \mathcal{D} \mid m \in S, |S| > 1\}$.

So using chain rule of KL-divergence we get:

$$\begin{aligned}
D_{KL}(\mathcal{I}^0, \mathcal{I}^m) &= \sum_{t=1}^{T} D_{KL}(\mathcal{I}_t^0, \mathcal{I}_t^m) = \sum_{t=1}^{T}\sum_{S\in S^{(m)}} Pr_{\mathcal{I}^0}(\mathcal{X}_t = S) KL\big(\text{Mult}(\boldsymbol{\theta}^0, \mathcal{X}_t), \text{Mult}(\boldsymbol{\theta}^m, \mathcal{X}_t)\big) \\
&\leq \sum_{t=1}^{T}\sum_{S\in S^{(m)}} Pr_{\mathcal{I}^0}(\mathcal{X}_t = S)\frac{180\epsilon^2}{|\mathcal{X}_t|} \leq 180\epsilon^2 \sum_{t=1}^{T}\sum_{S\in S^{(m)}} \frac{\mathbf{E}_{\mathcal{I}^0}[N_T(S)]}{|S|},
\end{aligned}$$

where the last inequality follows by noting for any set $S \subseteq \mathcal{D}$, $KL\big(\text{Mult}(\boldsymbol{\theta}^0, S), \text{Mult}(\boldsymbol{\theta}^m, S)\big) \leq 180\frac{\epsilon^2}{|S|}$ for any $\epsilon \in (0, 0.2]$. Further averaging over $\mathcal{I}^m$s for all $m \in [d]$:

$$\frac{\sum_{m\in[d]} D_{KL}(\mathcal{I}^0,\mathcal{I}^m)}{d} \le \frac{180\epsilon^2 \sum_{m\in[d]} \sum_{S\in S^{(m)}}^{K} \mathbf{E}_{\mathcal{I}^0}[N_T(S)]}{d} \le \frac{180\epsilon^2 T}{d}.$$

Now combining above with Eqn. (7) and 10 we get:

$$\mathbf{E}[R_T(\mathcal{A})] \ge \frac{\epsilon}{2}\Bigg(T - \bigg(\frac{\sum_{m\in[d]} \mathbf{E}_{\mathcal{I}^0}[N_T(m)]}{d} + T\sqrt{\Big(\frac{1}{2d}\sum_{m\in[d]} D_{KL}(\mathcal{I}^0,\mathcal{I}^m)\Big)}\bigg)\Bigg)$$

$$\ge \frac{\epsilon}{2}\Bigg(T - \bigg(\frac{T}{d} + \epsilon T\sqrt{\Big(\frac{90T}{d}\Big)}\bigg)\Bigg) \overset{(a)}{\ge} \frac{\epsilon}{2}\Bigg(T - \bigg(\frac{T}{2} + \frac{T}{4}\bigg)\Bigg) = \frac{\epsilon}{2}\Big(T - \frac{3T}{4}\Big) = \frac{\sqrt{dT}}{32\sqrt{90}},$$

where (a) holds by setting $\epsilon = \frac{1}{4}\sqrt{\frac{d}{90T}}$ and since $d \ge 2$, and the last equality follows for the above choice of ϵ, concluding the proof. $\square$

G.2 Pseudocode for Sta′D++ (Alg. 4)

Given the above lower bound, we see that the learner having a provision to play q-subsetwise queries does not help in faster learning (small regret bound), the question we started asking in Rem. 2. It is hence easy to see that our Alg. 3, itself yields an optimal $\tilde{O}(\sqrt{dT})$ algorithm using $q = 2$ at every rounds (i.e. only making pairwise queries per round). However, we here propose a general version of Alg. 3 which is also based on the idea of *stagewise-elimination* but can exploit subsetwise preferences for any general $q \ge 2$. So even though for the worst case instances a better regret guarantee is not possible (as shown in Lem. 11), the hope is it would be able to exploit the problem structure when there is sufficient 'quality-gap' (in terms of $\boldsymbol{\theta}$) between best vs the rest of the items. Analyzing such instance dependent guarantees for the problem setup would definitely be interesting but beyond the scope of this work. Following (Alg. 4) describes our algorithm for general q-subsetwise preferences and the corresponding worst case (i.e. instance-independent) regret analysis (Thm. 12).

Main Ideas. The main idea is to exploit the subsetwise feedback using the idea of *rank-breaking* [21, 35] for extracting pairwise estimates from multiwise (subsetwise) preference information. Formally, at any round t if we play the set $\mathcal{X}_t$ and observe the winner i_t, *rank-breaking* suggests as if to treat item i_t has beaten all the rest of the items in $\mathcal{X}_t$ in a pairwise duel, resulting in $|\mathcal{X}_t| - 1$ pairwise preferences of the form $(i_t \succ j),\ \forall j \in \mathcal{X}_t \setminus \{i_t\}$. Upon extracting these pairwise duels, we can similarly get estimate the $\hat{\boldsymbol{\theta}}_{t+1}$ by running a MLE on these extracted pairwise preferences up to round t. Given the $\hat{\boldsymbol{\theta}}$ estimate, the algorithm then proceed same as the original Alg. 3. However instead of selecting a pair of arms, we can now select a subset of q 'most-promising arms', by first selecting a potential good arm and then recursively selecting the best challenger of the already selected items using a *recursive max-min subset selection rule* as used in [33]. Details is given in Alg. 4. The challenging part however lies in its regret analysis which requires to justify the right concentration rates of $\boldsymbol{\theta}_t^s$, obtained from the above (rank-broken) pairwise estimates (see Lem. 15 in the proof of Thm. 12).

G.3 Proofs for Thm. 12

Theorem 12 (Regret bound of Sta′D++ (Alg. 4))**.** *Consider any $\delta > 0$, and suppose we set the parameters of Sta′D++ (Alg. 3) as $\eta = \frac{3}{2\kappa}\sqrt{2\log\frac{3qTK}{\delta}}$, and $t_0 = 2\bigg(\frac{C_1\sqrt{d}+C_2\sqrt{\log(2q/\delta)}}{\lambda_{\min}(B)}\bigg)^2 + \frac{4\Lambda}{\lambda_{\min}(B)}$, where Λ, κ, B is as defined in Lem. 4. Then with probability at least $(1-\delta)$, the T round cumulative regret of Sta′D++ is at most $O\bigg(\frac{\sqrt{dT\log(T)}}{\kappa}\sqrt{\log\Big(\frac{qTK}{\delta}\Big)\log\Big(\frac{Td}{\kappa}\log\frac{q}{\delta}\Big)}\bigg)$.*

Proof. The main challenge in the regret analysis lies in justifying the right concentration rates of $\boldsymbol{\theta}_t^s$ computed from the MLE estimates of preferences in $\boldsymbol{\phi}^s$. Towards this we use the result from [32]

Algorithm 4 Sta$'$D++ (for general q-subsetwise preferences)

1: **input:** Learning rate $\eta > 0$, exploration length $t_0 > 0$
2: **init:** Select t_0 pairs $\{(\mathbf{x}_\tau, \mathbf{y}_\tau)\}_{\tau \in [t_0]}$, each drawn at random from $\mathcal{S}_\tau$, and observe the corresponding preference feedback $\{o_\tau\}_{\tau \in [t_0]}$
3: $S \leftarrow \lfloor \log T \rfloor$, $\boldsymbol{\phi}^s \leftarrow \{(\mathbf{x}_\tau, \mathbf{y}_\tau, o_\tau)\}_{\tau \in [t_0]}$, $\forall s \in [\lfloor \log T \rfloor]$
4: Set $V_{t_0+1} := \sum_{\tau=1}^{t_0} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$
5: **while** $t \le T$ **do**
6: $\quad s \leftarrow 1, \mathcal{G}^1 \leftarrow [K]$ $\quad\triangleright$ Init stagewise pruning
7: $\quad$ **repeat**
8: $\quad\quad$ Compute the MLE estimate on $\boldsymbol{\phi}^s$, i.e. solve for $\hat{\boldsymbol{\theta}}_t^s$ s.t.:
$$\sum_{(\mathbf{x}_\tau, \mathbf{y}_\tau, o_\tau) \in \boldsymbol{\phi}^s} \Big(o_\tau - \sigma\big((\mathbf{x}_\tau - \mathbf{y}_\tau)^\top \hat{\boldsymbol{\theta}}_t^s\big)\Big)(\mathbf{x}_\tau - \mathbf{y}_\tau) = \mathbf{0}$$
9: $\quad\quad$ Set: $V_t^s = \sum_{(\mathbf{x}_\tau, \mathbf{y}_\tau) \in \boldsymbol{\phi}^s} (\mathbf{x}_\tau - \mathbf{y}_\tau)(\mathbf{x}_\tau - \mathbf{y}_\tau)^\top$
10: $\quad\quad$ Compute: $g_t^s(i) = \hat{\boldsymbol{\theta}}_t^{s\top} \mathbf{x}_i^t$, $\forall i \in \mathcal{G}^s$, and $p_t^s(i,j) = \eta \|\mathbf{x}_i^t - \mathbf{x}_j^t\|_{V_t^{s-1}}$, $\forall i, j \in \mathcal{G}^s$
11: $\quad\quad$ **if** $p_t^s(i,j) \le \frac{1}{\sqrt{T}}$, $\forall i, j \in \mathcal{G}^s$ **then**
12: $\quad\quad\quad$ $a_t \leftarrow \arg\max_{a \in \mathcal{G}^s} g_t^s(a)$. $\mathcal{X}_t \leftarrow a_t$
13: $\quad\quad\quad$ **for** $q' = 1, 2, \ldots (q-1)$ **do**
14: $\quad\quad\quad\quad$ $b_t \leftarrow \arg\max_{b \in \mathcal{G}^s \setminus \mathcal{X}_t} \Big[\min_{a \in \mathcal{X}_t} \big(g_t^s(b) + p_t^s(b,a)\big)\Big]$ $\quad\triangleright$ Max-Min subset selection
15: $\quad\quad\quad\quad$ $\mathcal{X}_t \leftarrow \mathcal{X}_t \cup \{b_t\}$
16: $\quad\quad\quad\quad$ **if** $\mathcal{G}^s \setminus \mathcal{X}_t = \emptyset$: Break the For Loop (Goto Line 18)
17: $\quad\quad\quad$ **end for**
18: $\quad\quad\quad$ Play $\mathcal{X}_t$. Receive $i_t \in \mathcal{X}_t$ $\quad\triangleright$ Winner feedback of subset $\mathcal{X}_t$
19: $\quad\quad$ **else if** $p_t^s(i,j) \le \frac{1}{2^s}$, $\forall i, j \in \mathcal{G}^s$ **then**
20: $\quad\quad\quad$ Find $\mathcal{B}_t^s := \{i \in \mathcal{G}^s \mid \exists j \in \mathcal{G}^s \text{ s.t. } g_t^s(i) + \frac{1}{2^s} < g_t^s(j)\}$
21: $\quad\quad\quad$ Update $\mathcal{G}^{s+1} \leftarrow \mathcal{G}^s \setminus \mathcal{B}^s$, $s \leftarrow s+1$
22: $\quad\quad$ **else**
23: $\quad\quad\quad$ Select a subset $\mathcal{X}_t \subseteq \mathcal{G}^s$ (up to size q) s.t. any pair $(a,b) \subseteq \mathcal{X}_t$ satisfies $p_t^s(a,b) > \frac{1}{2^s}$
24: $\quad\quad\quad$ Play $\mathcal{X}_t$. Receive $i_t \in \mathcal{X}_t$
25: $\quad\quad\quad$ Update $\boldsymbol{\phi}^s \leftarrow \boldsymbol{\phi}^s \cup \{(i_t, j, 1)\}_{j \in \mathcal{X}_t \setminus \{i_t\}}$ $\quad\triangleright$ Converting q-subsetwise winner preferences to pairwise preferences: Rank-Breaking update
26: $\quad\quad$ **end if**
27: $\quad$ **until** a set $\mathcal{X}_t$ is found
28: $\quad$ Update: $t \leftarrow t+1$
29: **end while**

which shows that for Plackett-Luce model, the rank broken pairwise estimates are unbiased for true pairwise preferences (see Lem. 1, [32]).

Now recall that our subsetwise preference feedback model also corresponds to a Plackett-Luce model with utility parameters of item $\mathbf{x} \in \mathcal{D}$ being $e^{\mathbf{x}^\top \boldsymbol{\theta}^*}$ (as discussed in Sec. 2.1). Hence above result applies to our setup which implies for any triplet $(\mathbf{x}_\tau, \mathbf{y}_\tau, o_\tau)$ in $\boldsymbol{\phi}^s$, we have $o_\tau \sim \text{Ber}\Big(\sigma\big((\mathbf{x}_\tau - \mathbf{y}_\tau)^\top \boldsymbol{\theta}^*\big)\Big)$. Consequently, we can hence again apply the same *finite-samples-asymptotic-normality for MLE estimates of GLM models* (Thm. 1,[26]) to derive a similar *sharper concentration of pairwise scores* as used in Lem. 4 for the original Stagewise-Adaptive-Duel algorithm (Alg. 3) for pairwise preferences. Precisely, in this case we can derive the following concentration bound:

Lemma 15. *Consider any $\delta > 0$, and suppose we set the parameters of Sta$'$D++ (Alg. 3) as $\eta = \frac{3}{2\kappa}\sqrt{2\log\left(\frac{3qTK}{\delta}\right)}$, where $\kappa := \inf_{\|x-y\| \le 2, \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\| \le 1}\Big[\sigma'\big((\mathbf{x}-\mathbf{y})^\top \hat{\boldsymbol{\theta}}\big)\Big]$, and $t_0 = 2\left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(2q/\delta)}}{\lambda_{\min}(B)}\right)^2 + \frac{4\Lambda}{\lambda_{\min}(B)}$, where $\Lambda = \frac{8}{\kappa^4}\big(d^2 + \log(3q/\delta)\big)$ and $B = \mathbf{E}_{\mathbf{x},\mathbf{y} \overset{iid}{\sim} \mathcal{P}_\mathcal{D}}[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^\top]$ (for some universal problem independent constants $C_1, C_2 > 0$). Then with probability* *at least* $(1-\delta)$*, for all stages* $s \in \lceil \log T \rceil$ *at all rounds* $t > t_0$ *and for all index pairs* $i, j \in \mathcal{G}^s$ *of round* t*:* $|(\mathbf{x}_i^t - \mathbf{x}_j^t)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t^s) \le p_t^s(i,j)|$.

As shown in the proof of Thm. 5, this concentration guarantee is really the key result used towards proving the regret bound of Alg. 3. It is easy to check that, given above concentration, the other two supporting lemmas (Lem. 6 and Lem. 7) can easily be shown to hold good in this case as well due to the similar 'pairwise-preference' based arm-selection strategy as that of Alg. 3 (with some additional care needed to incorporate the *max-min subset selection* strategy used for this case, same as the technique suggested in [33]). Moreover Lem. 14 simply follows for this algorithm as well since we maintain independent pairwise samples in $\boldsymbol{\phi}^s$ across different stages $s \in [\lfloor \log T \rfloor]$ (same as Alg. 3). The result of Thm. 12 can now be derived by combining the above results, similar to argument shown in the proof of Thm. 5. $\square$