Proceedings of Machine Learning Research vol 132:1–27, 2022

Efficient and Optimal Algorithms for Contextual Dueling Bandits under Realizability

Aadirupa Saha AADIRUPA.SAHA@MICROSOFT.COM
Microsoft Research, NYC

Akshay Krishnamurthy AKSHAYKR@MICROSOFT.COM
Microsoft Research, NYC

Abstract

We study the K-armed contextual dueling bandit problem, a sequential decision making setting in which the learner uses contextual information to make two decisions, but only observes *preference-based feedback* suggesting that one decision was better than the other. We focus on the regret minimization problem under realizability, where the feedback is generated by a pairwise preference matrix that is well-specified by a given function class $\mathcal{F}$. We provide a new algorithm that achieves the optimal regret rate for a new notion of best response regret, which is a strictly stronger performance measure than those considered in prior works. The algorithm is also computationally efficient, running in polynomial time assuming access to an online oracle for square loss regression over $\mathcal{F}$. This resolves an open problem of Dudík et al. (2015) on oracle efficient, regret-optimal algorithms for contextual dueling bandits.

Keywords: Contextual, Dueling Bandits, Preference-based learning, Realizability, Function approximation, Regret analysis, Best-Response, Policy regret, Regression oracles, Efficient, Optimal algorithms, Markov games, Linear realizability, Agnostic

1. Introduction

In many decision-making scenarios, a significant obstacle towards deploying reinforcement learning is the design of the reward function. For example, in personalization applications, reward engineering to align the performance of the reinforcement learning algorithm with application-specific objectives often requires months of iterating in a trial-and-error manner, requiring substantial effort from domain experts and resulting in sub-optimal system performance in the interim. Thus instead of engineering a reward function that may misalign with long-term objectives, it may be beneficial to re-design the system to collect more reliable signals that enable efficient optimization.

Preference/comparative feedback is a particular signal that is often available — or can be made available — in many applications, and is often more reliable than ordinal/absolute rewards. Preference-based feedback can be easily collected in applications including online retail chain optimization, prediction markets, tournament ranking, recommender systems, search engine optimization and information retrieval, robotics, multiplayer games, and elsewhere. As just one example, Hofmann et al. (2013) interleave results from two different search engine ranking algorithms and use click information as a preference signal, which they show has high fidelity and is significantly less expensive than collecting relevance judgements from experts.

Motivated by such scenarios, recent work from the machine learning community has studied online decision making from pairwise/preference feedback through the *dueling bandits* framework. This framework is a variant of the widely-studied multi-armed bandit (MAB) setting (Auer et al., 2002; Slivkins, 2019; Lattimore and Szepesvári, 2020), where rather than receive rewards, the

learner obtains preference-based feedback information. In particular, the learner repeatedly selects a pair of items to be compared to each other in a so-called duel and observes a stochastic binary outcome, indicating the winning item in this duel. Performance is often measured by some notion of regret, and while many definitions have been studied (Yue et al., 2012; Zoghi et al., 2014; Saha et al., 2021), they all intuitively ask that the learner identify the good actions, i.e., those that are typically favored amongst the others. Over the last two decades, several algorithms have been proposed for dueling bandit problems (Ailon et al., 2014; Zoghi et al., 2014; Komiyama et al., 2015; Wu and Liu, 2016) and generalizations to subset-wise preference feedback (Sui et al., 2017; Brost et al., 2016; Saha and Gopalan, 2019a; Ren et al., 2018; Saha and Gopalan, 2019b).

In practice, preferences over items can vary substantially with auxilliary/side information like user demographics, search query, etc.; however, the majority of dueling bandits literature does not leverage contextual information to learn higher-quality decision making policies. This shortcoming motivated Dudík et al. (2015) to formulate the *contextual dueling bandits* problem, in which the agent first receives a context, chooses a pair of actions, and then observes the outcome of the duel, with the goal of learning a *policy* that maps contexts to actions that typically win *in that context*. They formulated a new notion of regret and designed two types of algorithms: (1) a regret-optimal algorithm that is computationally intractable, and (2) a computationally tractable algorithm with suboptimal regret. Thus, their work left open the following question:

Is there a computationally efficient and statistically optimal algorithm for contextual dueling bandits?

In this paper, we resolve this question in the affirmative under a natural realizability assumption.

1.1. Our contributions

Our main contribution is a new efficient algorithm for contextual dueling bandits. To state the guarantee, let $\mathcal{X}$ be a context space, let $\mathcal{A} := [K]$ be an action space of size K, and let $\mathcal{P} := \{\mathbf{P} \in [-1,1]^{K\times K} : P[i,j] = -P[j,i], P[i,i] = 0\}$ denote the set of *preference matrices*, which are skew-symmetric matrices with bounded entries and 0 along the diagonal. We interpret a preference matrix $\mathbf{P}$ as encoding a zero-sum game, in which the row player's goal is to maximize their value and the column player's goal is to minimize.

In a stochastic contextual dueling bandit instance, the learner interacts with a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{P}$ via the following protocol: at each round t (1) nature samples $(x_t, \mathbf{P}_t) \sim \mathcal{D}$ and reveals x_t to the learner, (2) learner chooses (potentially randomly) two actions $(a_t, b_t) \in [K]^2$, (3) learner observes $o_t \sim \text{Ber}(\frac{P_t[a_t,b_t]+1}{2})$. The goal of the learner is to choose actions (a_t, b_t) so as to minimize the *best-response regret* over T rounds:

$$\text{BR-Regret}_T := \sum_{t=1}^{T} \max_{\mathbf{q}_t\in\Delta_K} \frac{1}{2}\mathbb{E}_{a\sim\mathbf{q}_t}\mathbb{E}_{(a_t,b_t)\sim\mathbf{p}_t}\left[f^\star(x_t)[a,a_t] + f^\star(x_t)[a,b_t]\right]. \tag{1}$$

Here we use $(a_t, b_t) \sim \mathbf{p}_t$ to capture the learner's randomness and define $f^\star : x \mapsto \mathbb{E}_{\mathcal{D}}[\mathbf{P} \mid x]$ to denote the conditional mean of the preference matrix $\mathbf{P}$ given context x. Intuitively, achieving low regret requires that the learner's distribution $\mathbf{p}_t$ cannot be exploited by an adversary that knows the expected preference matrix $f^\star$, so that $\mathbf{p}_t$ is typically preferred over any other distribution.

We consider the function approximation setting, where we are given a function class $\mathcal{F} : \mathcal{X} \to \mathcal{P}$ that we may use to learn the conditional mean function $f^\star$. To enable this we make two somewhat

standard assumptions about $\mathcal{F}$, which have also appeared in prior work (Foster and Rakhlin, 2020; Foster et al., 2020; Simchi-Levi and Xu, 2020; Agarwal et al., 2012; Foster et al., 2021).

Assumption *[Function approximation assumptions, informal] We assume* realizability*, that is $f^\star \in \mathcal{F}$. We also assume access to an* online square loss oracle *for $\mathcal{F}$ whose T-step square loss regret w.r.t. $\mathcal{F}$ is bounded by a known function* $\mathrm{Reg}_{\mathrm{Sq}}(T)$.

See Section 2.1 for a detailed description of the online square loss oracle. In this setting, our main theorem is as follows.

Theorem 1 *Under the above function approximation assumptions, Algorithm* **`MinMaxDB`**(γ)*, with learning rate $\gamma = O\left(\sqrt{\frac{KT}{\mathrm{Reg}_{\mathrm{Sq}}(T)}}\right)$ ensures:*

$$\text{BR-Regret}_T \leq O\left(\sqrt{KT \cdot \mathrm{Reg}_{\mathrm{Sq}}(T)}\right). \tag{2}$$

for any $T > 4K\mathrm{Reg}_{\mathrm{Sq}}(T)$. Additionally, **`MinMaxDB`** *incurs at most a poly(K) factor of run-time overhead over the square loss oracle.*

In the sequel, we list several instantiations for the online square loss oracle, but briefly the algorithm achieves (1) the optimal $\sqrt{KT\log|\mathcal{F}|}$ regret for finite function classes, (2) $\sqrt{K^3T}$ regret for the non-contextual problem, which is a factor of K worse than the optimal rate, and (3) $\sqrt{KdT}$ regret when $\mathcal{F}$ is (low-dimensional) linear functions. This is the first oracle-efficient algorithm for contextual dueling bandits with regret scaling at the optimal $\sqrt{T}$-rate. See Corollary 4 for detailed discussions.

Other Contributions. In addition to our main result (Theorem 1), the paper contains the following contributions:

(1). The notion of best-response regret in Eq. (1) itself is new, and we provide connections to other regret definitions in the literature. In particular, we show that it upper bounds the policy regret definition from Dudík et al. (2015) and also subsumes some other notions studied in dueling bandits.

(2). We also make a connection between the dueling bandits literature and the literature on Markov games, which we believe was previously unexplored. In particular, the Markov games literature has used game-theoretic techniques to developed UCB-based algorithms that can be applied directly to dueling bandits problems. We elaborate on this connection and provide complete analyses for these algorithms in an effort to encourage more cross-pollination between these communities (Sec. 3).

(3). Finally, we provide some evidence suggesting that in the absence of realizability, significantly new techniques are required to develop oracle-efficient $\sqrt{T}$-regret algorithms for contextual dueling bandits. Our evidence does not rule out such a result altogether, but it shows that the existing techniques from the standard contextual bandits literature are insufficient. We leave developing such an algorithm as an interesting open problem (Sec. 5).

1.2. Related work

Non-contextual dueling bandits. Our work builds on a large body of literature on the non-contextual (stochastic) dueling bandits problem, which can be seen as a special case of our setup

where there is only a single context, $|\mathcal{X}| = 1$, and hence a single preference matrix $\mathbf{P}$. For the non-contextual problem, the dominant algorithmic strategy is based on optimism in the face of uncertainty, which is widely deployed across sequential decision making. In terms of results, various regret definitions, largely motivated by social choice theory, have been studied. The most frequently used benchmark is the *Condorcet winner*, which is an arm $a^\star$ that beats all others on average (Yue et al., 2012; Zoghi et al., 2014, 2015b; Komiyama et al., 2015; Yue and Joachims, 2011). Generalizing slightly, one can consider a *Fixed-Benchmark* regret:

$$\text{FB-Regret}_T := \mathbf{E}_{a^\star \sim \mathbf{q}^\star}\left[\sum_{t=1}^{T} \mathbf{E}_{(a_t,b_t)\sim \mathbf{p}_t} \frac{P[a^\star, a_t] + P[a^\star, b_t]}{2}\right], \tag{3}$$

where $q^\star \in \Delta_K$ is a fixed (possibly unknown) distribution over the actions. Regret against the Condorcet winner is a special case, although a Condorcet winner may not exist for a given preference matrix $\mathbf{P}$ (Jamieson et al., 2015). To connect this definition with our results, note that our definition of best-response regret, Eq. (1), upper bounds fixed-benchmark regret for any $q^\star$:

Fact 1 *For the non-contextual setting, we have* $\text{FB-Regret}_T \le \text{BR-Regret}_T$*, for any* $q^\star \in \Delta_K$.

As such, Theorem 1 immediately yields guarantees for the non-contextual fixed-benchmark setting. In particular, we obtain $\tilde{O}(\sqrt{K^3T})$ worst-case fixed-benchmark regret, which is slightly worse than the minimax optimal $\tilde{O}(\sqrt{KT})$ rate (Dudík et al., 2015). On the other hand, our regret notion is strictly stronger, and, most importantly, our results generalize to the contextual setting which does not seem possible using techniques from this literature.

Beyond fixed-benchmark regret, notions involving Borda (Busa-Fekete and Hüllermeier, 2014; Jamieson et al., 2015; Falahatgar et al., 2017) and Copeland scores (Zoghi et al., 2015a; Komiyama et al., 2016; Wu and Liu, 2016) have also been considered. Our work does not directly yield results for these notions. However, we note that, as discussed by Dudík et al. (2015), these notions fail the *independence of clones criterion* (Schulze, 2011), which makes them somewhat undesirable in contextual settings. Finally we note that many other variations of the non-contextual dueling bandits problem have been considered, including adversarial preference matrices (Gajane et al., 2015; Saha et al., 2021), best-arm identification (Saha and Gopalan, 2019b, 2020; Yue and Joachims, 2011), full-ranking (Szörényi et al., 2015; Falahatgar et al., 2017; Saha and Gopalan, 2018), top-set detection (Busa-Fekete et al., 2013; Mohajer et al., 2017; Chen et al., 2018), etc. A very thorough literature survey on the recent developments in preference bandits can be found in Bengs et al. (2021); Sui et al. (2018).

Contextual dueling bandits. We are only aware of a few results for the contextual setting. The first is the work of Yue and Joachims (2009), which proposes a policy-gradient style algorithm, and establishes a $T^{3/4}$ style regret bound under convexity assumptions on the preference model. A follow-up work by Kumagai (2017) also uses gradient-based techniques to show an improved $T^{1/2}$ regret guarantee, but this result requires even stronger assumptions on the preference model.

Another line of work considers contextual dueling bandits under a special class of utility based preferences (González et al., 2017; Sui et al., 2017; Saha, 2021). Saha (2021) provides two $\tilde{O}(T^{1/2})$ regret algorithms assuming the preferences are a function of underlying utility scores of the individual arms, where utility scores of each arm are assumed to be a linear function of the arm-features. González et al. (2017) makes a similar assumption and provides empirical results, but they do not establish any theoretical guarantees. Sui et al. (2017) assumes the preference relations to be generated

from an unknown Gaussian process model, but also do not obtain regret bounds for their algorithms. In comparison, we do not require any assumptions (beyond realizability) on the preference model, and we still obtain a $T^{1/2}$ rate.

Closest to our work is the paper of Dudík et al. (2015), which studies the contextual dueling bandits problem with an abstract finite policy set $\Pi : \mathcal{X} \to \mathcal{A}$ and without realizability. They propose a minimax notion of regret given by:

$$\text{Policy-Regret}_T := \max_{\pi \in \Pi} \sum_{t=1}^{T} \frac{1}{2} \left[f^\star(x_t)[\pi(x_t), a_t] + f^\star(x_t)[\pi(x_t), b_t] \right].$$

This definition differs from our best-response notion in two ways: (1) the adversary chooses a policy for all T rounds, rather than a per-round distribution $\mathbf{q}_t$ and (2) we include the expectation over the learner's randomness in our definition. Based on these differences, we can show that our definition upper bounds the definition of Dudík et al. (2015):

Fact 2 *Let Π be given, with $|\Pi| < \infty$. Then for any learner that chooses $(a_t, b_t) \sim \mathbf{p}_t$ at round t, we have*

$$\text{Policy-Regret}_T \leq \text{BR-Regret}_T + \sqrt{T \log(|\Pi|/\delta)},$$

with probability $1 - \delta$.

Note that the minimax rate for policy regret is $\sqrt{KT \log |\Pi|}$, so the additive term is of lower order. Additionally, we are assuming realizability, without which it is impossible to attain sublinear best-response regret, while minimizing policy regret is always possible.

Apart from the difference in regret definition, Dudík et al. (2015) provides two algorithmic results: an inefficient sparring algorithm that achieves the optimal regret rate, and an ϵ-greedy algorithm that achieves a sub-optimal $T^{2/3}$-style regret assuming access to an *offline policy optimization oracle*. In comparison, we give an algorithm with optimal regret assuming realizability as well as access to an *online square loss minimization oracle*. We emphasize that the oracle models are quite different, so the results are not directly comparable. However, previous experimentation with standard contextual bandits suggests that algorithms based on square loss minimization may be more effective in practice (Foster et al., 2021).

Markov games. Finally, we highlight a growing body of work on preference based reinforcement learning in the Markov games framework (Littman, 1994). Broadly speaking, a Markov game models a multi-step decision making problem where several players compete to maximize their payoff over the course of an episode. While the specific formulations vary considerably, one formulation can be seen as a multi-step generalization of (contextual) dueling bandits, where regret is measured using our best response notion (which, as discussed, subsumes many other notion in the literature) (Bai and Jin, 2020; Xie et al., 2020; Bai et al., 2020). In particular, Xie et al. (2020) introduce the *coarse correlated equilibrium* strategy and show that it obtains $\sqrt{T}$ regret in the linear function approximation setting, which directly gives an efficient contextual dueling bandits algorithm under linear realizability. We highlight this technique in Sec. 3 in an attempt to better connect these two lines of work. On the other hand, our results for general function classes under realizability are novel, and we hope that they find applications in Markov games.

2. Problem Setup

Notation. Let $[n] := \{1, 2, \dots n\}$, for any $n \in \mathbb{N}$. Given a set $\mathcal{A}$, for any two items $a, b \in \mathcal{A}$, we use $a \succ b$ to denote that a is preferred over b. We use lower case bold letters for vectors and upper case bold letters for matrices. $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. For any vector $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_2$ denotes the ℓ_2 norm of $\mathbf{x}$. $\Delta_K := \{\mathbf{p} \in [0,1]^K \mid \sum_{i=1}^K p_i = 1, p_i \geq 0, \forall i \in [K]\}$ denotes the K-simplex, $\Delta_{K^2} = \Delta_{K \times K}$. $\mathbf{e}_i$ denotes the i-th standard basis vector in $\mathbb{R}^K$. For this work, we consider the *zero-sum* representation of preference matrices:

$$\mathcal{P} := \{\mathbf{P} \in [-1,1]^{K \times K} \mid P[i,j] = -P[j,i], P[i,i] = 0, \ \ \forall i, j \in [K]\}^1.$$

Note any $\mathbf{P} \in \mathcal{P}$ can be viewed as a *zero-sum game*, where the two players, called row and column player resp., simultaneously choose two (possibly randomized) items from $[K]$, with their goal being to respectively maximize and minimize the value of the selected entry.

2.1. Online Regression Oracle

An online regression oracle (Cesa-Bianchi and Lugosi, 2006, Chapter 3), is an algorithm, which we denote by `SqrReg`, and which operates in the following online protocol: on each round t (1) it receives an abstract input $z_t \in \mathcal{Z}$, from some input space $\mathcal{Z}$, chosen adversarially by the environment, (2) it produces a real-valued prediction $\hat{y}_t \in \mathcal{Y} \subset \mathbb{R}$ where $\mathcal{Y}$ is some output space, and (3) it observes the true response $y_t \in \mathcal{Y}$ and incurs loss $\ell(\hat{y}_t, y_t) := (\hat{y}_t - y_t)^2$. The goal of the oracle is to predict the outcomes as well as the best function in a given function class $\mathcal{F} := \{f : \mathcal{Z} \mapsto \mathbb{R}\}$, such that for every sequence of outcomes, the square loss regret is bounded.[2]

Formally, adopting the notation of Foster and Rakhlin (2020), at time t and for input $z \in \mathcal{Z}$, `SqrReg` can be seen as a mapping $\hat{y}(z) :=$`SqrReg`$(z, \{z_\tau, y_\tau\}_{\tau=1}^{t-1})$. Note this corresponds to the prediction the algorithm would make at time t if we passed in the input z, although this input may not be what is ultimately selected by the environment. We will take $\mathcal{Z} = (\mathcal{X} \times [K]^2)$ to be the set of (context, action-pair) tuples such that $z_t = (x_t, a_t, b_t)$ with $x_t \in \mathcal{X}$ and $(a_t, b_t) \in [K]^2$. Our output space is simply $\mathcal{Y} = [-1, 1]$.

Assumption 1 *The online regression oracle* `SqrReg` *guarantees for every sequence* $\{z_t, y_t\}_{t \in [T]}$*, its regret is bounded as* $\sum_{t=1}^T \big(\hat{y}_t(z_t) - y_t\big)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^T \big(f(z_t) - y_t\big)^2 \leq \mathrm{Reg}_{\mathrm{Sq}}(T)$*, where* $\mathrm{Reg}_{\mathrm{Sq}}(T) = o(T)$ *is a known upper bound.*

If we further assume realizability, in the sense that there exists $f^\star \in \mathcal{F}$ such that $\forall t : f^\star(z_t) = \mathbb{E}[y_t \mid z_t]$, then it is well-known that Assumption 1 further implies

$$\sum_{t=1}^T \big(\hat{y}_t(z_t) - f^\star(z_t)\big)^2 \leq \mathrm{Reg}_{\mathrm{Sq}}(T). \tag{4}$$

As we will see, under our realizability assumption on the preference matrices, the underlying square loss regression problem is also realizable, allowing us to appeal to Eq. (4).

1. Standard dueling bandit literature generally represents preference matrices $\mathbf{Q} \in [0,1]^{K \times K}$, such that $Q[i,j]$ indicates the probability of item i being preferred over item j. Here $\mathbf{Q}$ satisfies $Q[i,j] = 1 - Q[j,i]$ and $Q[i,i] = 0.5$. Note both representations are equivalent as there exists a one to one mapping $\mathbf{P} = (2\mathbf{Q} - 1) \in \mathcal{P}$ (Dudík et al., 2015; Bengs et al., 2021)
2. The square loss itself does not play a crucial role and can be replaced by any loss function that is strongly convex with respect to the predictions (Foster and Rakhlin, 2020).

Remark 1 (Some examples) *Online square loss regression is a well-studied problem, and efficient algorithms with provable regret guarantees are known for many specific function classes including finite classes where $|\mathcal{F}| < \infty$, finite and infinite dimensional linear classes, and others (Foster and Rakhlin, 2020; Foster et al., 2020). For completeness, we provide formal definition for some specific classes and instantiations of the regression oracles in Appendix B.1.*

2.2. Setup and Objective

We assume a context set $\mathcal{X} \subseteq \mathbb{R}^d$, action space of K actions denoted by $\mathcal{A} := [K]$, and a function class $\mathcal{F} = \{f : \mathcal{X} \mapsto [-1, 1]^{K \times K}\}$, all known to the learner ahead of the game. At each round, we assume a context-preference pair $(x_t, \mathbf{P}_t) \sim \mathcal{D}$ is drawn from a joint-distribution $\mathcal{D}$, such that $x_t \in \mathcal{X}$, and $\mathbf{P}_t \in [-1, 1]^{K \times K}$. The task of the learner is to select a pair of actions $(a_t, b_t) \in [K] \times [K]$, upon which an outcome $o_t \in \{\pm 1\}$ is revealed to the learner according to $\mathbf{P}_t$; specifically the probability that a_t is preferred over b_t, indicated by $o_t = +1$, is given by $\Pr(o_t = 1) := \Pr(a_t \succ b_t) = \frac{P_t[a_t, b_t]+1}{2}$, and hence $\Pr(o_t = -1) := \Pr(b_t \succ a_t) = \frac{1-P_t[a_t,b_t]}{2}$.

Assumption 2 (Realizability) *Define $f^\star : x \mapsto \mathbb{E}[\mathbf{P} \mid \mathbf{x}]$. We assume that $f^\star \in \mathcal{F}$. Thus, for any $a, b \in [K]$, we have $\mathbb{E}[\mathbf{P}[a, b] \mid x] = f^\star(x)[a, b]$.*

Objective: Best-Response Regret Assuming the learner selects the duel $(a_t, b_t) \sim \mathbf{p}_t \in \Delta_{K \times K}$ at each round t, we measure the learner's performance via a notion of best response regret, defined as:

$$\text{BR-Regret}_T := \sum_{t=1}^{T} \max_{q_t \in \Delta_K} \frac{1}{2} \mathbb{E}_{a \sim q_t} \mathbb{E}_{(a_t, b_t) \sim \mathbf{p}_t} \left[f^\star(x_t)[a, a_t] + f^\star(x_t)[a, b_t] \right].$$

Remark 2 (Learner's Obligation to Randomize) *As stated, we allow the learner to choose its actions randomly and our regret definition includes an expectation over this randomness. An alternative would be to measure the best response regret on the realized outcomes (a_t, b_t) chosen by the learner at each time:*

$$\sum_{t=1}^{T} \max_{\mathbf{q}_t \in \Delta_K} \frac{1}{2} \mathbb{E}_{a \sim q_t} \left[f^\star(x_t)[a, a_t] + f^\star(x_t)[a, b_t] \right]. \tag{5}$$

The two regret definitions (BR-Regret_T vs the one defined in (5)) are of course equivalent if the learner does not randomize, but they are very different in general. In fact, if we measure regret on the realized outcomes, as displayed above (irrespective of whether the learner is allowed to randomize or not), then sublinear regret is not possible in general, since the preference matrices $f^\star(\cdot)$ may not have pure-strategy Nash equilibria. See Appendix B.2 for a concrete example. On the other hand, if the learner randomizes and we incorporate this into the regret definition, then in principle the learner could set both marginals of $\mathbf{p}_t$ to be a Nash equilibrium for $f^\star(x_t)$ to guarantee 0 regret.

3. Warm up: Structured Function Classes

As a warm up, and to highlight an overlooked connection between dueling bandits and Markov games, we briefly sketch how UCB-based algorithms can achieve $\sqrt{T}$-regret in some structured

dueling bandits settings. These algorithms have appeared previously in the Markov games literature (Bai et al., 2020; Xie et al., 2020), so we summarize the key ideas here and defer additional details to the appendices. While these ideas are not technically novel, in light of the relationship between best-response regret and previously studied notions in the dueling bandit literature, we believe it is worthwhile to bring these techniques to the attention of the dueling bandit community.

We begin with the standard "non-contextual" dueling bandits setting where there is just a single unknown preference matrix $\mathbf{P} \in [-1,1]^{K\times K}$. Here, it is natural to deploy a confidence-based strategy that, at round t, maintains an estimate $\hat{\boldsymbol{P}}_t$ of the underlying parameter and a confidence set $C_t[i,j] := \tilde{O}\left(\sqrt{\frac{1}{N_t[i,j]}}\right)$ for each entry, where $N_t[i,j]$ is the number of times that entry (i,j) has been dueled prior to round t. The critical component of the algorithm, and the main departure from standard UCB approaches, is the action selection scheme. Here we find a *coarse correlated equilibrium* (CCE) of the "upper confidence" matrix $\boldsymbol{U}_t := \hat{\boldsymbol{P}}_t + \mathbf{C}_t$, defined as any joint distribution $\mathbf{p}_t \in \Delta_{K\times K}$ that satisfies:

$$\sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[a,b] \geq \max_{a^\star\in[K]} \sum_{b\in[K]\times[K]} p_t^r[b]U_t[a^\star,b],$$
$$\sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[b,a] \geq \max_{b^\star\in[K]} \sum_{a\in[K]\times[K]} p_t^\ell[a]U_t[b^\star,a],$$

where $p_t^\ell[\cdot] = \sum_b p_t[\cdot,b]$ is the "left" marginal and p_t^r is the analogously defined "right" marginal. In words, the CCE is a joint distribution over the actions of the two players, such that neither player is incentivized to deviate unilaterally from their marginal strategy (Dey, 2019). Since the matrix $\mathbf{U}_t$ is not zero-sum, a Nash equilibrium may not be efficiently computable (Dey, 2019; Daskalakis et al., 2009); however, a CCE is guaranteed to exist and is easily computed by solving the above linear feasiblity problem. Returning to the algorithm, we find a CCE solution $\mathbf{p}_t$ for the upper confidence matrix $\mathbf{U}_t$, sample $(a_t,b_t) \sim \mathbf{p}_t$, observe the outcome, and update our statistics for the next round. We refer this algorithm as `CCE-DB`. The full pseudocode is presented in Appendix C.1 (see Algorithm 2).

3.1. Regret Analysis for CCE based Algorithms

The algorithm summarized above achieves the following regret guarantee. This theorem essentially appears in both Xie et al. (2020) and Bai et al. (2020) in more general forms; both study multi-step Markov games, Xie et al. (2020) considers linear function approximation, and Bai et al. (2020) allows the two players to have different action set sizes.

Theorem 2 (Regret of `CCE-DB` (Alg. 2), informal) *In the non-contextual standard K-armed dueling bandits setting, the above algorithm has regret*

$$\text{BR-Regret}_T \leq O(K\log(KT)\sqrt{T}).$$

An intuitive proof sketch of the algorithm is given below (while the complete analysis is presented in Appendix C.2). We mention a few remarks, before describing the key step in the analysis.

1. The above bound is optimal in the dependence on T, up to logarithmic factors, as the worst-case lower bound is known to be $\Omega(\sqrt{T})$ even when assuming existence of a Condorcet

Winner (Komiyama et al., 2015; Dudík et al., 2015). Recall from Fact 1, that the best-response regret upper bounds the regret to any fixed benchmark, including a Condorcet winner (if it exists).

2. While near-optimal in its dependence on T, the dependence on the number of arms is sub-optimal, as it is possible to achieve $\tilde{O}(\sqrt{KT})$ best-response regret in the non-contextual setting. Indeed, this optimal rate can be achieved here and in other related settings by sparring optimal adversarial bandit algorithms, such as Exp3 (Dudík et al., 2015; Ailon et al., 2014; Gajane et al., 2015; Sui et al., 2017). Unfortunately, all algorithms that achieve the optimal rate rely heavily on adversarial online learning techniques and do not seem to yield efficient algorithms in the more general contextual setting, and so we believe it is worthwhile to also study algorithms with a more statistical flavor.
3. The result above can be generalized to any setting where valid and shrinking confidence intervals can be constructed, including linear and generalized linear dueling bandit settings under realizability. These results are presented in Appendix C.4, where we give an $\tilde{O}(d\sqrt{T})$ regret algorithm (Alg. 3) for the linear case (see Theorem 8, Appendix C.4).

Proof Sketch: Regret Analysis of `CCE-DB` (Alg. 2). Turning to the analysis, the first part of the analysis involves verifying the validity of the confidence intervals, which is quite standard. The more interesting part involves relating the regret of $\mathbf{p}_t$ to the confidence bounds, where we must crucially use the fact that $\mathbf{p}_t$ is the CCE for the upper confidence matrix $\mathbf{U}_t$. The essential calculation is as follows: for any $\mathbf{q} \in \Delta_K$, we have

$$\mathbf{q}^\top \mathbf{P} \mathbf{p}_t^\ell \overset{(i)}{\leq} \mathbf{q}^\top \mathbf{U}_t \mathbf{p}_t^\ell \overset{(ii)}{\leq} \max_{b^\star} \sum_{a \in [K]} p_t^\ell(a) U_t[b^\star, a] \overset{(iii)}{\leq} \sum_{a,b \in [K]\times[K]} p_t[a,b] U_t[b,a],$$

where (i) follows from the upper confidence property of $\mathbf{U}_t$, (ii) follows since $\mathbf{q} \in \Delta_K$ and (iii) follows from the fact that $\mathbf{p}_t$ is a CCE for $\mathbf{U}_t$.

The exact same calculation applies for the right player p_t^r and yields the bound $\sum_{a,b} p_t(a,b) U_t[a,b]$. Finally using the fact that $U_t[a,b] = -U_t[b,a] + 2C_t[a,b]$ we obtain

$$\mathbf{q}^\top \mathbf{P}(\mathbf{p}_t^\ell + \mathbf{p}_t^r) \leq \sum_{a,b} p_t[a,b](U_t[a,b] + U_t[b,a]) = 2\sum_{a,b} p_t[a,b] C_t[a,b].$$

In other words, we can bound the per-round regret by the confidence width of the actions chosen by the learner. This enables us to use standard potential-function arguments for bounding the confidence sum, which yields the final regret bound.

4. Main result: General function classes

While the above CCE-based algorithm yields $\sqrt{T}$-type regret for function classes that admit pointwise confidence intervals, this is only possible for certain structured function classes. In this section, we turn to our main result: an efficient algorithm for more general classes $\mathcal{F}$.

Our algorithm is an adaptation of the contextual bandit algorithm `SquareCB`, due to Foster and Rakhlin (2020), which uses an online square loss oracle to make predictions and an inverse gap-weighting scheme that uses these predictions for action selection. Their key lemma establishes

a *per-round inequality* that relates the contextual bandit regret to the square loss of the online oracle. This inequality is established in a minimax sense, where no assumptions are made about the predictions of the oracle or the true reward function.

We follow their recipe and instantiate the square loss oracle with instance space $\mathcal{Z} := \mathcal{X} \times [K] \times [K]$. Then at each round $t \in [T]$, after observing the context x_t, we can query the oracle for predictions on (x_t, a, b) for each $(a, b) \in [K] \times [K]$ and collect the predictions into a zero-sum matrix $\widehat{\mathbf{Y}}_t \in [-1, 1]^{K \times K}$. Now, the goal is to use $\widehat{\mathbf{Y}}_t$ to construct a distribution $\mathbf{p}_t \in \Delta_{K \times K}$ such that

$$\max_{f^\star \in D_{\mathrm{ZS}}} \underbrace{\max_{\mathbf{q} \in \Delta_K} \frac{1}{2}\mathbb{E}_{a^\star \sim \mathbf{q}, (a_t, b_t) \sim \mathbf{p}_t}\left[f^\star[a^\star, a_t] + f^\star[a^\star, b_t]\right]}_{\text{best-response regret of } \mathbf{p}_t} - \frac{\gamma}{4} \underbrace{\mathbb{E}_{(a_t, b_t) \sim \mathbf{p}_t}\left[(\widehat{Y}_t[a_t, b_t] - f^\star[a_t, b_t])^2\right]}_{\text{square loss of } \widehat{\mathbf{Y}}_t} \leq \frac{\mathrm{poly}(K)}{\gamma}.$$

Here $D_{\mathrm{ZS}} := \{\mathbf{M} \in [-1, 1]^{K \times K} \mid M[i, j] = -M[i, j], M[i, i] = 0\}$ is the set of zero-sum matrices and $\gamma > 0$ is a exploration rate parameter. In words, we are asking that, no matter the true preference matrix $f^\star$, $\mathbf{p}_t$ has best response regret that is upper bounded by the square loss of $\widehat{\mathbf{Y}}_t$ on actions chosen from $\mathbf{p}_t$, up to an additive $\mathrm{poly}(K)/\gamma$ term. If we can find such a $\mathbf{p}_t$, then we can sample $(a_t, b_t) \sim \mathbf{p}_t$, observe the outcome o_t and pass the example $z_t = (x_t, a_t, b_t)$ along with outcome o_t to the square loss oracle. This ensures that the observed loss is an unbiased estimate for the second term above, so if we add up the per-round inequality for all $t \in [T]$, we obtain

$$\begin{aligned}
\text{BR-Regret}_T &\leq \frac{\gamma}{4} \cdot \sum_{t=1}^{T} \mathbb{E}_{(a_t, b_t) \sim \mathbf{p}_t}\left[(\widehat{Y}_t[a_t, b_t] - f^\star(x_t)[a_t, b_t])^2\right] + \frac{\mathrm{poly}(K) \cdot T}{\gamma} \\
&\leq \frac{\gamma}{4} \cdot \mathrm{Reg}_{\mathrm{Sq}}(T) + \frac{\mathrm{poly}(K) \cdot T}{\gamma} = O\left(\sqrt{\mathrm{poly}(K) \cdot T \cdot \mathrm{Reg}_{\mathrm{Sq}}(T)}\right).
\end{aligned}$$

Here, the second inequality follows from Assumption 1, in particular, Eqn. (4) and the fact that the example $z_t = (x_t, a_t, b_t)$ that we feed to the square loss oracle will have $(a_t, b_t) \sim \mathbf{p}_t$. Then the final bound is based on tuning $\gamma \asymp \sqrt{\frac{\mathrm{poly}(K)T}{\mathrm{Reg}_{\mathrm{Sq}}(T)}}$.

Thus, the main remaining step is to establish the per-round regret inequality. This is where our analysis departs from that of Foster and Rakhlin (2020), as we must account for the game-theoretic structure of the best response regret definition. To proceed with this analysis, let us define the per-round minimax value as:

$$V(\gamma) := \max_{\mathbf{Y} \in D_{\mathrm{ZS}}} \min_{\mathbf{p} \in \Delta_{K \times K}} \max_{\substack{f^\star \in D_{\mathrm{ZS}} \\ \mathbf{q} \in \Delta_K}} \left[\sum_{(a,b)} q[a] f^\star[a, b] \frac{p^\ell[b] + p^r[b]}{2} - \frac{\gamma}{4} \sum_{i \neq j} p[i, j](f^\star[i, j] - Y[i, j])^2\right], \tag{6}$$

where $p^\ell[\cdot] = \sum_b p[\cdot, b]$ is the "left" marginal and $p^r[\cdot] = \sum_a p[a, \cdot]$ is the "right" marginal of $\mathbf{p}$. The following lemma shows that the minimax value $V(\gamma)$ is bounded by $O(K/\gamma)$.

Lemma 3 *For any $\gamma \geq 2K$, $V(\gamma) \leq \frac{5K}{\gamma}$.*

Proof By a change of variables, we can write

$$V(\gamma) = \max_{\mathbf{Y} \in D_{\mathrm{ZS}}} \min_{\mathbf{p} \in \Delta_{K \times K}} \max_{\substack{\delta \in D_{\mathrm{ZS}} \\ \mathbf{q} \in \Delta_K}} \left[\sum_{(a,b)} q[a](Y[a, b] + \delta[a, b]) \frac{p^\ell[b] + p^r[b]}{2} - \frac{\gamma}{4} \sum_{i \neq j} p[i, j] \delta[i, j]^2\right].$$

Next, we relax all constraints on $\delta_{a,b}$, fix $\mathbf{p}$ and $\mathbf{q}$, and optimize over $\delta_{a,b}$. Maximizing the quadratic for each $\delta_{a,b}$, we find that setting $\delta_{a,b} = \frac{q[a](p^\ell[b]+p^r[b])}{\gamma p[a,b]}$ yields an upper bound on $V(\gamma)$:

$$\begin{aligned}
V(\gamma) &\leq \max_{\mathbf{Y}\in D_{\mathrm{ZS}}} \min_{\mathbf{p}\in\Delta_{K\times K}} \max_{\mathbf{q}\in\Delta_K} \left[\sum_{a,b} q[a]Y[a,b]\frac{p^\ell[b]+p^r[b]}{2} + \frac{1}{2\gamma}\sum_{a,b}\frac{q[a]^2(p^\ell[b]+p^r[b])^2}{p[a,b]}\right] \\
&\leq \max_{\mathbf{Y}\in D_{\mathrm{ZS}}} \min_{\mathbf{p}\in\Delta_K} \max_{\mathbf{q}\in\Delta_K} \left[\sum_{a,b} q[a]Y[a,b]p[b] + \frac{2}{\gamma}\sum_{a,b}\frac{q[a]^2p[b]^2}{p[a]p[b]}\right] \\
&\leq \max_{\mathbf{Y}\in D_{\mathrm{ZS}}} \min_{\mathbf{p}\in\Delta_K} \max_{\mathbf{q}\in\Delta_K} \left[\sum_{a,b} q[a]Y[a,b]p[b] + \frac{2}{\gamma}\sum_{a}\frac{q[a]}{p[a]}\right]. \qquad (7)
\end{aligned}$$

Above, the first inequality uses the maximizing value of $\delta_{a,b}$, while in the second inequality, we restrict the minimizing player $\mathbf{p}$ to sample (a,b) iid from a marginal distribution (which we overload and also call $\mathbf{p}$). In the last inequality we use the fact that $\mathbf{p},\mathbf{q}\in\Delta_K$ so that, e.g., $q[a]^2 \leq q[a]$.

We bound the final term by fixing $\mathbf{Y}$ and applying the minimax theorem. To do so, observe that the objective is linear (and hence concave) in $\mathbf{q}$ and convex in $\mathbf{p}$. To ensure that the objective is defined everywhere, we further shrink the domain for $\mathbf{p}$ by *smoothing*: Fixing $\epsilon > 0$, for any $\mathbf{p}\in\Delta_K$ we define $\mathbf{p}^{(\epsilon)} := (1-\epsilon)\mathbf{p} + \epsilon\mathbf{1}/K$. As $\mathbf{p}^{(\epsilon)}$ itself is a distribution, this upper bounds our objective while ensuring that the conditions for applying the minimax swap are satisfied. As such, we obtain

$$\begin{aligned}
V(\gamma) &\leq \max_{\mathbf{Y}\in D_{\mathrm{ZS}}} \min_{\mathbf{p}\in\Delta_K} \max_{\mathbf{q}\in\Delta_K} \left[\sum_{a,b} q[a]Y[a,b]p^{(\epsilon)}[b] + \frac{2}{\gamma}\sum_{a}\frac{q[a]}{p^{(\epsilon)}[a]}\right] \\
&= \max_{\mathbf{Y}\in D_{\mathrm{ZS}}} \max_{\mathbf{q}\in\Delta_K} \min_{\mathbf{p}\in\Delta_K} \left[\sum_{a,b} q[a]Y[a,b]p^{(\epsilon)}[b] + \frac{2}{\gamma}\sum_{a}\frac{q[a]}{p^{(\epsilon)}[a]}\right] \\
&\leq \max_{\mathbf{Y}\in D_{\mathrm{ZS}}} \max_{\mathbf{q}\in\Delta_K} \left[\sum_{a,b} q[a]Y[a,b]q^{(\epsilon)}[b] + \frac{2}{\gamma}\sum_{a}\frac{q[a]}{q^{(\epsilon)}[a]}\right] \\
&= \max_{\mathbf{Y}\in D_{\mathrm{ZS}}} \max_{\mathbf{q}\in\Delta_K} \left[(1-\epsilon)\sum_{a,b} q[a]Y[a,b]q[b] + \frac{\epsilon}{K}\sum_{a,b} q[a]Y[a,b] + \frac{2}{\gamma}\sum_{a}\frac{q[a]}{(1-\epsilon)q[a]+\epsilon/K}\right].
\end{aligned}$$

Here, the first inequality restricts the domain for $\mathbf{p}$ using the smoothing operator, the first equality is the minimax swap, and the second inequality follows by choosing $\mathbf{p} = \mathbf{q}$. The remaining three terms are bounded as follows: (i) the first term is zero since $\mathbf{Y}\in D_{\mathrm{ZS}}$, (ii) the second term is trivially upper bounded by ϵ, (iii) the third term is at most $4K/\gamma$ as long as $\epsilon \leq 1/2$. Setting $\epsilon = K/\gamma$, we obtain the result. ■

Lemma 3, combined with the above discussion, immediately certifies the existence of a strategy that achieves $O(\sqrt{KT\cdot\mathrm{Reg}_{\mathrm{Sq}}(T)})$ regret for contextual dueling bandits with realizability. For constructing an algorithm, the missing piece is the action selection scheme $\mathbf{p}_t$, whose existence is

guaranteed by Lemma 3. For this, an examination of Eq. (7) reveals that we can compute a suitable $\mathbf{p}_t$ by solving a simple convex program in the action space. Specifically, given predictions $\widehat{\mathbf{Y}}_t$ on round t, we define $\mathbf{p}_t \in \Delta_K$ as any solution to the following convex feasibility problem:

$$\mathbf{p}_t \text{ satisfies } \forall i \in [K] : \sum_b \widehat{Y}_t[i,b]p_t[b] + \frac{2}{\gamma}\frac{1}{p_t[i]} \leq \frac{5K}{\gamma}. \tag{8}$$

The proof of Lemma 3 shows that this program is always feasible[3] and that $\mathbf{p}_t$ provides the per-round inequality that we require.

Algorithm 1 `MinMaxDB`

input: Arm set: $[K]$, parameters $\gamma > 0$.
    An instance of `SqrReg` for function class $\mathcal{F}$
for $t = 1, 2, \ldots, T$ **do**
    Receive context x_t
    $\forall\, a < b,\ a, b \in [K] \times [K]$: Query $\hat{y}(x_t, a, b) \leftarrow \texttt{SqrReg}\Big(\{(x_\tau, a_\tau, b_\tau), y_\tau\}_{\tau=1}^{t-1}\Big)$
    Collect predictions into matrix $\widehat{\mathbf{Y}}_t$ and find $\mathbf{p}_t \in \Delta_K$ satisfying

$$\forall i \in [K] : \sum_{b \in [K]} \widehat{Y}_t[i,b]p_t[b] + \frac{2}{\gamma}\frac{1}{p_t[i]} \leq \frac{5K}{\gamma}.$$

    Sample $a_t, b_t \overset{iid}{\sim} \mathbf{p}_t$, play the duel (a_t, b_t) and receive feedback o_t.
    Update `SqrReg` with example (x_t, a_t, b_t) and label o_t.
end for

We put all the pieces together to obtain our final algorithm, with pseudocode displayed in Algorithm 1. The main guarantee for the algorithm is as follows:

Theorem 3 *Under Assumptions 1 and 2, Algorithm 1 with* $\gamma = \sqrt{\frac{20KT}{\mathrm{Reg}_{\mathrm{Sq}}(T)}}$ *ensures that* $\text{BR-Regret}_T \leq \sqrt{5KT\mathrm{Reg}_{\mathrm{Sq}}(T)}$ *for any* $T \geq 4K\mathrm{Reg}_{\mathrm{Sq}}(T)$.

The restriction on T arises since Lemma 3 applies only when $\gamma \geq 2K$. The complete proof of Theorem 3 is given in Appendix D.1. Moreover, we further note that given the choice of $\mathbf{p}_t$ in Eq. 8, there is actually a simpler argument for our reduction and derivation of Theorem 3. We provide this analysis in Appendix D.2.

By instantiating the square loss oracle appropriately, we obtain end-to-end guarantees for many function classes of interest. Some of these results are summarized in the next corollary.

Corollary 4 *Algorithm 1 yields the following best-response regret guarantees:*

- *For* $\mathcal{F}$ *with* $|\mathcal{F}| < \infty$, *instantiating* `SqrReg` *as the exponential weights algorithm guarantees* $\mathrm{Reg}_{\mathrm{Sq}}(T) \leq O(\log|\mathcal{F}|)$ *and hence* $\text{BR-Regret}(T) \leq O(\sqrt{KT\log|\mathcal{F}|})$. *For example in the* K*-armed (non-contextual) dueling bandit setting, one can construct* $\mathcal{F}$ *such that* $\log|\mathcal{F}| = \tilde{O}\big(K^2\big)$ *to obtain* $\text{BR-Regret}(T) \leq \tilde{O}\big(\sqrt{K^3T}\big)$.

3. Note that we can add an $O(K/\gamma)$ slack term in the RHS of Eq. (8) to the tolerate the approximations arising from numerical methods. This affects the final result only in constant factors.

- *For low dimensional linear predictors* $\mathcal{F} = \{(x, a, b) \mapsto \langle \theta, \phi(x, a, b)\rangle : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1\}$*, instantiating* `SqrReg` *as the Vovk-Azoury-Warmuth forecaster guarantees* $\mathrm{Reg}_{\mathrm{Sq}}(T) \leq O(d\log(T/d))$ *and hence* $\text{BR-Regret}(T) \leq O(\sqrt{dKT\log(T/d)})$.
- *Alternatively, for linear predictors, instantiating* `SqrReg` *as online gradient descent guarantees* $\mathrm{Reg}_{\mathrm{Sq}}(T) \leq O(\sqrt{T})$ *and hence* $\text{BR-Regret}(T) \leq O(K^{1/2}T^{3/4})$.

Please see Appendix D.3 for details and additional examples. Note that the non-contextual rate of $\tilde{O}\big(\sqrt{K^3T}\big)$ is sub-optimal by an $O(K)$ factor. We close this section with two final remarks.

Remark 3 *Formally, Theorem 3 only requires Eq. (4) to hold, and this may be possible in somewhat more general settings than the realizability condition stated in Assumption 2. One example is the time-varying or dynamic regret setting, where preferences at round t are governed by $f_t^\star \in \mathcal{F}$ and we assume the sequence $(f_1^\star, \ldots, f_T^\star)$ has small total variation or path length. In such cases, one can achieve Eq. (4) with non-trivial $\mathrm{Reg}_{\mathrm{Sq}}(T)$ (Raj et al., 2020; Baby and Wang, 2021), which can then be used in Theorem 3.*

Remark 4 (Computational Complexity) *The two main computational bottlenecks in Algorithm 1 are the square loss oracle itself and the computation of $\mathbf{p}_t$ in each iteration. The former is efficient for many function classes of interest, while the latter involves (approximately) solving a convex feasibility problem in K dimensions with $O(K)$ constraints, which can be done in $\mathrm{poly}(K)$ time. Thus the algorithm incurs a $\mathrm{poly}(K)$ computational overhead over the square loss oracle.*

5. Discussion: A Barrier for Oracle-Efficient Agnostic Algorithms

As we have seen, realizability of the payoff matrices permits computationally efficient algorithms for contextual dueling bandits with optimal $\sqrt{T}$-type regret. At the same time, the classical approach of sparring Exp4 achieves a similar regret guarantee in the *agnostic* setting (e.g., without realizability), but unfortunately it is not computationally efficient for most policy classes of interest (Dudík et al., 2015). It is thus natural to ask if we can design computationally tractable algorithms for contextual dueling bandits in the absence of realizability.

In the standard contextual bandit setting, computational tractability for the agnostic setting is formalized by providing the algorithm a policy class $\Pi : \mathcal{X} \to \mathcal{A}$ and an *optimization oracle* over Π. The optimization oracle serves as an abstraction of supervised learning, allowing the algorithm to efficiently search over the class Π, and it leads to algorithms that can be implemented via a reduction to supervised learning as a primitive (Dudik et al., 2011; Agarwal et al., 2014; Krishnamurthy et al., 2015; Rakhlin and Sridharan, 2016; Syrgkanis et al., 2016a,b; Luo et al., 2018). Such algorithms are called *oracle-efficient*.

In this section, we provide some evidence to suggest that significantly new techniques are required to develop oracle-efficient algorithms for agnostic contextual dueling bandits with $\sqrt{T}$ regret. The main observation is that all such algorithms for standard contextual bandits establish some concentration inequality on the regret of all policies $\pi \in \Pi$, but establishing such a guarantee in the dueling setting requires incurring large regret. Indeed, Lem. 13 in Agarwal et al. (2014) asserts that

$$\forall \pi \in \Pi : \mathrm{Reg}(\pi) \lesssim 2\widehat{\mathrm{Reg}}_t(\pi) + c_0\sqrt{\frac{1}{t}}, \qquad \widehat{\mathrm{Reg}}_t(\pi) \leq 2\,\mathrm{Reg}(\pi) + c_0\sqrt{\frac{1}{t}}, \tag{9}$$

where $\mathrm{Reg}(\pi)$ is the population regret for π, and $\widehat{\mathrm{Reg}}_t(\pi)$ is an importance weighted empirical estimate based on t rounds of interaction, and $c_0 > 0$ is some constant that captures other problem parameters (e.g., number of actions, size of policy class, etc.), but depends at most logarithmically on t. This guarantee is central to the regret analysis, and similar bounds appear in related works.

In the dueling setting, we define $\mathrm{Reg}(\pi) = \max_{\pi'} \mathbb{E}_{(x,\mathbf{P})}\left[P[\pi'(x), \pi(x)]\right]$ and extend this to distributions over policies in the obvious way. Then, as a step toward porting the proof technique from Agarwal et al. (2014) to the dueling setting, we can ask if there is an estimator $\widehat{\mathrm{Reg}}_t(\cdot)$ that achieves the guarantee in (9) for this definition. Unfortunately, this is not possible without the algorithm incurring $\Omega(T)$ regret.

Proposition 4 *Consider any algorithm* ALG *that produces estimates* $\widehat{\mathrm{Reg}}_t(\cdot)$ *that satisfy* (9) *for all* $t \leq T$ *and all policies* π *in some given class* Π*. Then there is a contextual dueling bandit instance where* ALG *incurs* $\Omega(T/c_0^2)$ *regret.*

Proof [Sketch] Consider the non-contextual 3×3 preference matrix instance parametrized by ϵ:

$$\mathbb{E}\left[P\right] = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & \epsilon \\ 0 & -\epsilon & 0 \end{pmatrix}.$$

Let us label the actions a, b, and c and define $\Pi = \{a, b, c\}$. Here action a is a Condorcet winner, while action c is near optimal, with $\mathrm{Reg}(c) = \epsilon$. However, to estimate the regret of action c, we require playing the pair (b, c). This poses a problem, since we can use action a as the comparator when calculating the dueling regret of ALG, which shows that ALG's regret is *lower bounded* by the number of times that it plays action b.

In more detail, at time T, the two guarantees in (9) imply that $\widehat{\mathrm{Reg}}_T(c) \in [\frac{1}{2}(\epsilon - c_0/\sqrt{T}), 2\epsilon + c_0/\sqrt{T}]$. Now, consider two instances, one where $\epsilon = \epsilon_1 := c_0/\sqrt{T}$ and the other, where $\epsilon = \epsilon_2 := 8c_0/\sqrt{T}$. The intervals for $\widehat{\mathrm{Reg}}_T(c)$ do not intersect, and so, if ALG guarantees (9) we can use the value of $\widehat{\mathrm{Reg}}_T(c)$ as a test statistic to distinguish between these two instances. On the other hand, testing between these two instances is equivalent to testing whether the mean of a Bernoulli random variable is $(1 + \epsilon_1)/2$ or $(1 + \epsilon_2)/2$ from iid samples. The number of samples available for this problem is the number of times ALG plays the pair (b, c). A standard lower bound argument reveals that, for T large enough, ALG must play action b at least $\Omega(T/c_0^2)$ times, which proves the claim. ■

We emphasize that the above claim only shows that establishing a certain intermediate guarantee is not possible in dueling contextual bandits. It is a somewhat weak form of hardness that does not rule out oracle-efficient agnostic algorithms. On the other hand, as all such algorithms for standard contextual bandits do make claims similar to (9), Proposition 4 suggests that fundamentally new techniques are required to obtain agnostic algorithms for this setting. We believe that this is quite an interesting open problem.

Acknowledgements

AK thanks Akshay Balsubramani, Alekh Agarwal, Miroslav Dudík, and Robert E. Schapire for fruitful discussions regarding the result in Section 5.

References

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 2014.

Nir Ailon, Zohar Shay Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, 2014.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 2002.

Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 2001.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. *arXiv preprint arXiv:2104.11824*, 2021.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, 2020.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 2021.

Brian Brost, Yevgeny Seldin, Ingemar J Cox, and Christina Lioma. Multi-dueling bandits and their application to online ranker evaluation. In *International on Conference on Information and Knowledge Management*, 2016.

Róbert Busa-Fekete and Eyke Hüllermeier. A survey of preference-based online learning with bandit algorithms. In *International Conference on Algorithmic Learning Theory*, 2014.

Róbert Busa-Fekete, Balazs Szorenyi, Weiwei Cheng, Paul Weng, and Eyke Hüllermeier. Top-k selection based on adaptive sampling of noisy preferences. In *International Conference on Machine Learning*, 2013.

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

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*. SIAM, 2018.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 2009.

Palash Dey. Lecture notes: Algorithmic game theory. 2019.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Uncertainty in Artificial Intelligence*, 2011.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, 2015.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. In *Advances in Neural Information Processing Systems*, 2017.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2020.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems*, 2020.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*, 2021.

Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *International Conference on Machine Learning*, 2015.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D. Lawrence. Preferential Bayesian optimization. In *International Conference on Machine Learning*, 2017.

Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Transactions on Information Systems*, 2013.

Kevin G Jamieson, Sumeet Katariya, Atul Deshpande, and Robert D Nowak. Sparse dueling bandits. In *Artificial Intelligence and Statistics*, 2015.

Sham Kakade, Adam Tauman Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, 2011.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on Learning Theory*, 2015.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *International Conference on Machine Learning*, 2016.

Akshay Krishnamurthy, Alekh Agarwal, and Miroslav Dudík. Contextual semibandits via supervised learning oracles. In *Advances in Neural Information Processing Systems*, 2015.

Wataru Kumagai. Regret analysis for continuous dueling bandit. In *Advances in Neural Information Processing Systems*, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, 2017.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 1994.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, 2018.

Soheil Mohajer, Changho Suh, and Adel Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *International Conference on Machine Learning*, 2017.

Francesco Orabona, Koby Crammer, and Nicolo Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 2015.

Anant Raj, Pierre Gaillard, and Christophe Saad. Non-stationary online regression. *arXiv preprint arXiv:2011.06957*, 2020.

Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning*, 2016.

Wenbo Ren, Jia Liu, and Ness B Shroff. PAC ranking from pairwise and listwise queries: Lower bounds and upper bounds. *arXiv preprint arXiv:1806.02970v2*, 2018.

Aadirupa Saha. Optimal algorithms for stochastic contextual dueling bandits. In *Advances in Neural Information Processing Systems*, 2021.

Aadirupa Saha and Aditya Gopalan. Active ranking with subset-wise preferences. *Artificial Intelligence and Statistics*, 2018.

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

Aadirupa Saha and Aditya Gopalan. PAC Battling Bandits in the Plackett-Luce Model. In *Algorithmic Learning Theory*, 2019b.

Aadirupa Saha and Aditya Gopalan. Best-item learning in random utility models with subset choices. In *Artificial Intelligence and Statistics*. PMLR, 2020.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, 2021.

Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, 2011.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *arXiv preprint arXiv:2003.12699v4*, 2020.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 2019.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in neural information processing systems*, 2011.

Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Multi-dueling bandits with dependent arms. In *Uncertainty in Artificial Intelligence*, 2017.

Yanan Sui, Masrour Zoghi, Katja Hofmann, and Yisong Yue. Advancements in dueling bandits. In *IJCAI*, 2018.

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert Schapire. Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, 2016a.

Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert E Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems*, 2016b.

Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for plackett-luce: A dueling bandits approach. In *Advances in Neural Information Processing Systems*, 2015.

Volodya Vovk. Competitive on-line linear regression. In *Advances in Neural Information Processing Systems*, 1998.

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

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, 2020.

D. Pal Y. Abbasi-Yadkori and C. Szepesvari. Improved algorithms for linear stochastic bandits. In *Neural Information Processing Systems*, 2011.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *International Conference on Machine Learning*, 2011.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 2012.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International Conference on Machine Learning*, 2014.

Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, 2015a.

Masrour Zoghi, Shimon Whiteson, and Maarten de Rijke. MergeRUCB: A method for large-scale online ranker evaluation. In *ACM International Conference on Web Search and Data Mining*, 2015b.

Supplementary: Efficient and Optimal Algorithms for Contextual Dueling Bandits under Realizability

Appendix A. Appendix for Sec. 1.2

A.1. Proof of Fact. 2

Fact 2 *Let Π be given, with $|\Pi| < \infty$. Then for any learner that chooses $(a_t, b_t) \sim \mathbf{p}_t$ at round t, we have*

$$\text{Policy-Regret}_T \leq \text{BR-Regret}_T + \sqrt{T \log(|\Pi|/\delta)},$$

with probability $1 - \delta$.

Proof Since $(a_t, b_t) \sim \mathbf{p}_t$, we can use Hoeffding's inequality and a union bound to deduce that $\forall \pi \in \Pi$:

$$\sum_{t=1}^{T} \frac{1}{2} \left[f^\star(x_t)[\pi(x_t), a_t] + f^\star(x_t)[\pi(x_t), b_t] \right] \leq \sum_{t=1}^{T} \frac{1}{2} \mathbb{E}_{\mathbf{p}_t} \left[f^\star(x_t)[\pi(x_t), a_t] + f^\star(x_t)[\pi(x_t), b_t] \right] + \mathcal{O}\left(\sqrt{T \log |\Pi|/\delta} \right),$$

with probability $1 - \delta$. Now we can easily translate from Policy-Regret to BR-Regret by pushing the $\max_{\pi \in \Pi}$ inside the summation only yields an upper bound, justifying the claim. ■

Appendix B. Appendix for Sec. 2

B.1. Examples: Some Specific Regression Function Classes

1. Any finite regression class $\mathcal{F}$ such that $|\mathcal{F}| < \infty$, one can choose `SqrReg` such that

$$\text{Reg}_{\text{Sq}}(T) \leq 2 \log |\mathcal{F}|.$$

2. Class of linear predictors $\mathcal{F} := \{(x, a) \mapsto \langle \theta, x_a \rangle \mid \theta \in \mathbb{R}^d, \|\theta\| \leq 1\}$. In this case choosing `SqrReg` to be the Vovk-Azoury-Warmuth forecaster, as proposed by Vovk (1998); Azoury and Warmuth (2001), we have

$$\text{Reg}_{\text{Sq}}(T) \leq d \log(T/d).$$

3. Class of generalized linear predictors $\mathcal{F} := \{(x, a) \mapsto \sigma(\langle \theta, x_a \rangle) \mid \theta \in \mathbb{R}^d, \|\theta\| \leq 1\}$ where $\sigma : \mathbb{R} \mapsto [0, 1]$ is a fixed non-decreasing 1-Lipschitz link function. Here using GLMtron algorithm of Kakade et al. (2011) as `SqrReg` leads to $\text{Reg}_{\text{Sq}}(T) \leq \sqrt{T}$. Alternatively a second order variant of GLMtron leads to an instance dependent guarantee $\text{Reg}_{\text{Sq}}(T) \leq O(d \log T/\kappa_\sigma^2)$, further assuming a lower bound κ_σ on the gradient of σ, more precisely $\sigma' \geq \kappa_\sigma > 0$.

4. Reproducing kernel hilbert space (RKHS) $\mathcal{F} := \{f \mid \|f\|_{\mathcal{H}} \leq 1, \mathcal{K}(x_a, x_a) \leq 1\}$: Using (kernelized) Online Gradient Descent, one can obtain $\mathrm{Reg}_{\mathrm{Sq}}(T) \leq O(\sqrt{T})$.

5. Banach Spaces $\mathcal{F} := \{(x, a) \mapsto \langle \theta, x_a \rangle \mid \theta \in \mathcal{B}, \|\theta\| \leq 1\}$, where $(\mathcal{B}, \|\cdot\|)$ is a separable Banach space and x belongs to the dual space $(\mathcal{B}, \|\cdot\|_*)$: For this setting, whenever $\mathcal{B}$ is $(2, D)$-uniformly convex, using 'Online Mirror Descent' (for example see Orabona et al. (2015)) as `SqrReg` can be configured to have $\mathrm{Reg}_{\mathrm{Sq}}(T) \leq (T/D)^{1/2}$ (Srebro et al., 2011).

B.2. An Example for Remark 2

Consider the scenario where the leaner is not allowed to randomize and evaluated on a fixed sequence of actions $\{(a_t, b_t)_{t\in[T]}\}$ as defined in (5). The claim is for the regret definition in (5) it is impossible for the learner to achieve $o(T)$ best-response regret in the worst case. To see why, consider the following counter example with a single context:

$$f^\star := \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

This matrix is skew symmetric and hence zero-sum. However, for any choice (a, b) of the learner, the adversary has a choice that can guarantee value of $1/2$. Specifically, if learner chooses $(1, 2)$ then adversary chooses 1, if learner chooses $(1, 3)$ then adversary chooses 3 and if learner chooses $(2, 3)$ then adversary chooses 2. This shows that we must allow the learner to randomize.

Appendix C. Appendix for Sec. 3

C.1. `CCE-DB`: Algorithm Pseudocode for Standard "Non-Contextual" K-armed Dueling Bandits

C.2. Regret Analysis of Algorithm 2

Theorem 5 (Restatement of Thm. 2: Expected regret of `CCE-DB` on StdDB(K)) *For the setting of standard K-armed dueling bandit (StdDB(K)), the best-response regret of `CCE-DB` (Alg. 2) satisfies:*

$$\text{BR-Regret}_T \leq O(K \log(KT)\sqrt{T}).$$

Proof The proof relies on two main results: Confidence bounding $\mathbf{P}$ through $\mathbf{U}(t)$ (Lem. 6) and analyzing the instantaneous regret of the column and the row player (Lem. 7). Lem. 6 simply guarantees that with high probability at least $(1 - 1/T)$, $\mathbf{P}$ can be sandwithced inside $\hat{\boldsymbol{P}}(t) \pm \mathbf{C}(t)$, which is crucially used in the later part of the proof.

Lemma 6 (Confidence Bounding $\mathbf{P}$) *Setting $\delta = \frac{1}{T}$ in Alg. 2, we get*

$$Pr\big(\exists t \in [T], \exists (i, j) \in [K] \times [K],\ \hat{P}_t[i, j] - C_t[i, j] \leq P_t[i, j] \leq \hat{P}_t[i, j] + C_t[i, j]\big) \leq 1/T$$

.

Algorithm 2 `CCE-DB`

input: Arm set: $[K]$, parameters $\delta \in (0,1)$
init: $W_1[i,j] \leftarrow 0, \forall i,j \in [K]$. Use $N_t[i,j] := W_t[i,j] + W_t[j,i] \, \forall t \in [T]$
for $t = 1, 2, \ldots, T$ **do**
  $\tilde{P}_t[i,j] := \frac{W_t[i,j]}{N_t[i,j]}$, $\hat{P}_t[i,j] := 2\tilde{P}_t[i,j] - 1$, $C_t[i,j] \leftarrow \sqrt{\frac{\log(K^2t^2/\delta)}{N_t[i,j]}}$, $\forall i,j \in [K]$ (we assume $x/0 := 0.5, \ \forall x \in \mathbb{R}$)
  $U_t[i,j] \leftarrow \hat{P}_t[i,j] + C_t[i,j], U_t[i,j] \leftarrow 0, \ \ \forall i,j \in [K]$
  Denoting $p_t^\ell[\cdot] := \sum_{b=1}^K p_t[\cdot, b]$ and $p_t^r[\cdot] := \sum_{a=1}^K p_t[a, \cdot]$, find a policy $\mathbf{p}_t \in \Delta_{K\times K}$ (CCE of $\mathbf{U}_t$) such that:

$$\sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[a,b] \geq \max_{a^\star\in[K]} \Big[\sum_{b\in[K]} p_t^r[b]U_t[a^\star,b] \Big]$$
$$\sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[b,a] \geq \max_{b^\star\in[K]} \Big[\sum_{a\in[K]} p_t^\ell[a]U_t[b^\star,a] \Big], \tag{10}$$

  Play $(a_t, b_t) \sim \mathbf{p}_t$
  Receive preference feedback $o_t \in \{-1, 1\}$. $\tilde{o}_t \leftarrow (o_t + 1)/2$
  Update $W_{t+1}[a_t, b_t] \leftarrow W_t[a_t, b_t] + \tilde{o}_t$; $W_{t+1}[b_t, a_t] \leftarrow W_t[b_t, a_t] + (1 - \tilde{o}_t)$
end for

Recall that $\mathbf{p}_t^\ell$ and $\mathbf{p}_t^r$ respectively denotes the marginal distribution of the left and right arm of the dueling pair (a_t, b_t) when sampled as $(a_t, b_t) \sim \mathbf{p}_t$. The next claim upper bounds the regret of both left (row) and the right (column) action at each round t. Precisely, Lem. 7 shows the learner's instantaneous regret can be bounded by the expected confidence bounds of played duel as follows:

Lemma 7 (Learner's Instantaneous Regret) *For any $\mathbf{q} \in \Delta_K$, and $(a_t, b_t) \sim \mathbf{p}_t$*

$$\mathbf{q}^\top \mathbf{P}(\mathbf{p}_t^\ell + \mathbf{p}_t^r) \leq 2 \sum_{(a,b)\in[K]\times[K]} p_t[a,b]C_t[a,b].$$

The regret bound of Alg. 2 (Thm. 5) now follows by summing the instantaneous regret upper bound of Lem. 7 over $t \in [T]$. Precisely,

$$\begin{aligned}
\text{BR-Regret} &= \sum_{t=1}^T \max_{q_t\in\Delta_K} \mathbb{E}_{a\sim q_t}\mathbb{E}_{(a_t,b_t)\sim\mathbf{p}_t} \frac{[P[a,a_t] + P[a,b_t]]}{2} \\
&\leq \sum_{t=1}^T \frac{\left[\max_{\mathbf{q}\in\Delta_K} \mathbf{q}^\top \mathbf{P}(\mathbf{p}_t^\ell + \mathbf{p}_t^r)\right]}{2} \overset{(i)}{\leq} \sum_{t=1}^T \sum_{(a,b)\in[K]\times[K]} p_t[a,b]C_t[a,b] \\
&\overset{(ii)}{\leq} \sum_{t=1}^T C_t[a_t,b_t] + 4\sqrt{T}\log(KT) = 2\sqrt{\ln(KT)} \sum_{t=1}^T \sqrt{\frac{1}{N_t[a_t,b_t]}} + 4\sqrt{T}\log(KT) \\
&= 2\sqrt{\ln(KT)} \sum_{a<b} \sum_{\tau=1}^{N_t[a,b]} \sqrt{\frac{1}{\tau}} + 4\sqrt{T}\log(KT)
\end{aligned}$$

$$\overset{(iii)}{\leq} 4\sqrt{\ln(KT)}\sum_{a<b}\sqrt{N_t[a,b]} + 4\sqrt{T}\log(KT) \overset{(iv)}{\leq} 4\sqrt{\ln(KT)}\sqrt{K^2T} + 4\sqrt{T}\log(KT)$$

where (i) follows from Lem. 7, (ii) applies Azuma-Hoeffding's Inequality, (iii) uses $\sum_{i=1}^{n} 1/\sqrt{i} \leq 2\sqrt{n} - 1$, and (iv) applies Cauchy's Scharwz inequality. This concludes the proof. ■

C.3. Technical Lemmas for Thm. 5

C.3.1. PROOF OF LEM. 6

Lemma 6 (Confidence Bounding P) *Setting $\delta = \frac{1}{T}$ in Alg. 2, we get*

$$Pr\big(\exists t \in [T], \exists (i,j) \in [K]\times[K],\ \hat{P}_t[i,j] - C_t[i,j] \leq P_t[i,j] \leq \hat{P}_t[i,j] + C_t[i,j]\big) \leq 1/T$$

.

For any $\delta > 0$, then, with probability at least $1-\delta$, for any $i, j \in [K]$

$$\hat{P}_t[i,j] - C_t[i,j] \leq P[i,j] \leq U_t[i,j] := \hat{P}_t[i,j] + C_t[i,j], \qquad \forall t \in [T]\,.$$

Proof Suppose $\mathcal{G}_t[i,j]$ denotes the event that at time $t \in [T]$ and item-pair $i, j \in [K]$, $\hat{P}_t[i,j] - C_t[i,j] \leq P[i,j] \leq \hat{P}_t[i,j] + C_t[i,j]$. Note for any such that pair (i,i), $\mathcal{G}_t[i,i]$ always holds true for any $t \in [T]$ and $i \in [n]$, as $P_t[i,i] = U_t[i,i] = 0$ by definition. We can thus assume $i \neq j$. Moreover, for any t and i, j, $\mathcal{G}_t[i,j]$ holds if and only if $\mathcal{G}_t[i,j]$, thus we will restrict our focus only to pairs $i < j$ for the rest of the proof. Hence, to prove the lemma it suffices to show

$$\mathbf{P}\Big(\exists t \in [T], \exists i < j, \text{ such that } \mathcal{G}_t^c[i,j]\Big) \leq \frac{1}{T}\ ,$$

which we do now. $\mathcal{G}_t[i,j]$ can be rewritten as:

$$|\hat{P}_t[i,j] - P[i,j]| \leq C_t[i,j]\,.$$

Let $\tau_{ij}(n)$ the time step $t \in [T]$ when the pair (i,j) was updated (i.e. i and j was compared) for the n^{th} time. We now bound the probability of the confidence bound $(\mathcal{G}_t[i,j])$ getting violated at any round $t \in [T]$ for some duel (i,j) as follows:

$$\begin{aligned}
&\mathbf{P}\Big(\exists t \in [T], i < j, \text{ such that } \mathcal{G}_t^c[i,j]\Big) \\
&\leq \sum_{i<j} \mathbf{P}\left(\exists n \geq 0, |P[i,j] - \hat{P}_{\tau_{ij}(n)}[i,j]| > C_t[i,j]\right) \\
&= \sum_{i<j} \mathbf{P}\left(\exists n \in [0,T],\ |P[i,j] - \hat{P}_n[i,j]| > C_t[i,j]\right),
\end{aligned}$$

where $\hat{P}_t[i,j] = \frac{W_t[i,j]}{W_t[i,j]+W_t[i,j]}$ is the frequentist estimate of $p[i,j]$ at round t (after $n = N_t[i,j] \in [0,T]$ comparisons between arm i and j). Noting $N_{\tau_{ij}(n)}[i,j] = n$, $\tau_{ij}(n) > n$, and using Hoeffding's inequality, we further get

$$\mathbf{P}\Big(\exists t \in [T], i < j, \text{ such that } \mathcal{G}_t^c[i,j]\Big) \leq \sum_{i<j}\left[\sum_{n=1}^{T} 2e^{-2n\frac{\ln(K^2\tau_{ij}(n)^2/\delta)}{2n}}\right]$$

$$\leq \sum_{i<j}\left[\sum_{n=1}^{T}\frac{\delta}{K^2n^2}\right] < \sum_{i<j}\left[\sum_{n=1}^{\infty}\frac{\delta}{K^2n^2}\right]$$
$$< \left[\frac{K(K-1)}{2}\frac{\delta\pi^2}{K^26}\right] < \delta = \frac{1}{T}.$$

This concludes the claim. ■

C.3.2. PROOF OF LEM. 7

Lemma 7 (Learner's Instantaneous Regret) *For any* $\mathbf{q} \in \Delta_K$, *and* $(a_t, b_t) \sim \mathbf{p}_t$

$$\mathbf{q}^\top \mathbf{P}(\mathbf{p}_t^\ell + \mathbf{p}_t^r) \leq 2 \sum_{(a,b)\in[K]\times[K]} p_t[a,b]C_t[a,b].$$

Proof Note for any $\mathbf{q} \in \Delta_K$,

$$\mathbf{q}^\top \mathbf{P}\mathbf{p}_t^\ell = \sum_{a^\star=1}^{K}\sum_{a=1}^{K} q[a^\star]p_t^\ell[a]P[a^\star,a] \leq \sum_{a^\star=1}^{K}\sum_{a=1}^{K} q[a^\star]p_t^\ell[a]U_t[a^\star,a]$$
$$= \max_{a^\star\in[K]}\sum_{a=1}^{K} p_t^\ell[a]U_t[a^\star,a] \leq \sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[a,b]$$

where the first inequality follows from Lem. 6 and last inequality by the second inequality constraint of the CCE equations (see Eqn. (10)).

On the other hand for the right action (column player), similarly again for any $\mathbf{q} \in \Delta_K$:

$$\mathbf{q}^\top \mathbf{P}\mathbf{p}_t^r = \sum_{a^\star=1}^{K}\sum_{a=1}^{K} q[a^\star]p_t^r[a]P[a^\star,a] \overset{\text{Lem.6}}{\leq} \sum_{a^\star=1}^{K}\sum_{a=1}^{K} q[a^\star]p_t^r[a]U_t[a^\star,a]$$
$$= \max_{a^\star\in[K]}\sum_{a=1}^{K} p_t^r[a]U_t[a^\star,a] \leq \sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[a,b]$$

where the last inequality follows from first inequality constraint of the CCE equations (see Eqn. (10)).

Finally combining above two results and noting that for any $(a,b) \in [K]\times[K]$, $U_t[b,a] = \hat{P}_t[b,a] + C_t[b,a] = -(\hat{P}_t[a,b] + C_t[a,b]) + C_t[a,b] + C_t[b,a] = -U_t[a,b] + 2C_t[a,b]$, the claim follows. ■

C.4. Other Structured Function Classes

`CCE-DB` (Alg. 2), analyzed above, can be extended to other parametric structured function classes as well which may support a statistical estimation based techniques, such as generalized linear function classes, etc. We briefly discuss the case for linear function classes here:

Setting: Dueling-Bandits with Linear Realizability (LinDB(d)) Consider $\mathcal{X} \subseteq [-1,1]^{K\times K\times d}$, such that if $\mathbf{x}_t$ is the context received at time t then $f(x_t)[a,b] := \mathbf{w}^\top \mathbf{x}_t[a,b] \in [-1,1]$ for any pair $(a,b) \in [K]\times[K]$, for some unknown $\mathbf{w} \in [-1,1]^d$. Note $f(\mathbf{x}) \in [-1,1]^{K\times K}$ for any $\mathbf{x} \in \mathcal{X}$. Considering the same setup of Sec. 2, the goal is to again minimize the BR-Regret. We detail the complete pseudocode in Alg. 3 (`CCE-linDB`), and analyze its regret guarantee as follows:

Theorem 8 (Expected regret for LinDB) *For the setting of dueling bandits with linear-realizability LinDB(d) class, we have* $\text{BR-Regret}_T \le O(d\log(KT)\sqrt{T})$.

Algorithm 3 `CCE-linDB`

input (tuning parameters): Regularizer $\lambda > 0$, Exploration length $t_0 > 0$
init: Set $V_1 := \lambda \mathbf{I}_d$, $Y_1 \leftarrow 0$, $X_1 \leftarrow \mathbf{0}_d \in \mathbb{R}^d$
for $t = 1, 2, \dots T$ **do**
  $\hat{\mathbf{w}} \leftarrow (X_t^\top X_t + \lambda \mathbf{I})^{-1} X_t^\top Y_t$
  Receive the context vector $\mathbf{x}_t \in [-1,1]^{K\times K\times d}$
  Pairwise preference estimates: $\hat{P}_t[a,b] \leftarrow \hat{\mathbf{w}}^\top \mathbf{x}_t[a,b]$, for all $[a,b] \in [K]\times[K]$
  UCB estimates: $U_t[a,b] \leftarrow \hat{P}_t[a,b] + \eta\sqrt{\mathbf{x}_t[a,b]^\top V_t^{-1} \mathbf{x}_t[a,b]}$. $U_t(a,a) \leftarrow 0$, for all $(a,b) \in [K]\times[K]$
  Find a policy $\mathbf{p}_t \in \Delta_{K\times K}$ (CCE of $U(t)$) such that:

$$\sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[a,b] \ge \max_{a^\star\in[K]} \Big[\sum_{b\in[K]} p_t^r[b]U_t[a^\star,b]\Big]$$
$$\sum_{a,b\in[K]\times[K]} p_t[a,b]U_t[b,a] \ge \max_{b^\star\in[K]} \Big[\sum_{a\in[K]} p_t^\ell[a]U_t[b^\star,a]\Big],$$

  where $p_t^\ell[\cdot] := \sum_{b=1}^K p_t[\cdot,b]$ and $p_t^r[\cdot] := \sum_{a=1}^K p_t[a,\cdot]$.
  Play $(a_t, b_t) \sim \mathbf{p}_t$.
  Receive preference feedback $o_t \in \{-1, 1\}$.
  Update $V_{t+1} = V_t + \mathbf{x}_t[a_t,b_t]\mathbf{x}_t[a_t,b_t]^\top \in \mathbb{R}^{d\times d}$
  Update $Y_{t+1} \leftarrow [Y_t; o_t] \in \mathbb{R}^{t+1}$, $X_{t+1} \leftarrow [X_t; \mathbf{x}_t[a_t,b_t]] \in \mathbb{R}^{t+1\times d}$
end for

Thm. 8 gives the BR-Regret of `CCE-linDB` (Alg. 3 for the LinDB setup). The regret analysis of Thm. 8 follows exactly same as the proof of Thm. 5 along with applying the standard concentration techniques from the linear bandits literature with proper tuning of λ and η (see Y. Abbasi-Yadkori and Szepesvari (2011); Li et al. (2017); Lattimore and Szepesvári (2020) for details on concentration results of linear bandits).

Appendix D. Appendix for Sec. 4

D.1. Proof of Thm. 3

Proof Start by noting that when the learner plays (a_t, b_t) from a product distribution s.t. $(a_t, b_t) \sim \mathbf{p}_t \times \mathbf{p}_t$. Then the best-response regret becomes:

$$\begin{aligned}\text{BR-Regret}_T &:= \sum_{t=1}^{T} \max_{\mathbf{q}\in\Delta_K} \mathbb{E}_{a\sim\mathbf{q}}\mathbb{E}_{a_t\sim\mathbf{p}_t}\left[f^\star(x_t)[a, a_t]\right] = \sum_{t=1}^{T} \max_{\mathbf{q}\in\Delta_K} \mathbf{q}^\top f^\star \mathbf{p}_t \\ &\le \Big[\frac{\gamma \text{Reg}_{\text{Sq}}(T)}{4} + \sum_{t=1}^{T} \frac{5K}{\gamma}\Big] = \sqrt{5KT\text{Reg}_{\text{Sq}}(T)}\end{aligned}$$

where the last inequality follows from Lem. 3, and last equality is due to choosing $\gamma = \sqrt{\frac{20KT}{\text{Reg}_{\text{Sq}}(T)}}$, which leads to the desired regret guarantee of Thm. 3. Further, note we need the constraint $T \ge 4K\text{Reg}_{\text{Sq}}(T)$ since Lem. 3 requires $\gamma > 2K$. And since we set $\gamma = \sqrt{\frac{20KT}{\text{Reg}_{\text{Sq}}(T)}}$, this is satisfied only if $T \ge 4K\text{Reg}_{\text{Sq}}(T)$. ■

D.2. Simpler Analysis of Thm. 3 (using Eqn. 8)

Given the choice of $\mathbf{p}_t$ as shown in Eqn. 8, we now give simpler and more direct proof of Thm. 3. Recall we assume the that dueling arms (a_t, b_t) are drawn from a product measure $(a_t, b_t) \sim \mathbf{p}_t \times \mathbf{p}_t$ for some $\mathbf{p}_t \in \Delta_K$ at each round. Now suppose we find $\mathbf{p}_t$ that satisfies Eqn. (8).

Then for any $f^\star$ and any $\mathbf{q}$, using Eqn. (8) we have:

$$\begin{aligned}\mathbf{q}^\top f^\star \mathbf{p}_t &\le \mathbf{q}^\top (f^\star - \hat{Y}_t)\mathbf{p}_t - \frac{2}{\gamma}\sum_a \frac{q[a]}{p_t[a]} + \frac{5K}{\gamma} \\ &= \sum_{a=1}^{K} \frac{q[a]}{\sqrt{p_t[a]\gamma}} \cdot \sqrt{p_t[a]\gamma}(f^\star[a,\cdot] - \hat{Y}_t[a,\cdot])\mathbf{p}_t - \frac{2}{\gamma}\sum_a \frac{q[a]}{p_t[a]} + \frac{5K}{\gamma} \\ &= \sum_{a=1}^{K} \sqrt{\frac{q[a]^2}{p_t[a]\gamma} \cdot p_t[a]\gamma\big((f^\star[a,\cdot] - \hat{Y}_t[a,\cdot])\mathbf{p}_t\big)^2} - \frac{2}{\gamma}\sum_a \frac{q[a]}{p_t[a]} + \frac{5K}{\gamma} \\ &\le \frac{1}{2\gamma}\sum_a \frac{q[a]^2}{p_t[a]} + \frac{\gamma}{2}\sum_a p_t[a]\left((f^\star[a,\cdot] - \hat{Y}_t[a,\cdot])\mathbf{p}_t\right)^2 - \frac{2}{\gamma}\sum_a \frac{q[a]}{p_t[a]} + \frac{5K}{\gamma} \\ &\le \frac{2}{\gamma}\sum_a \frac{q[a]}{p_t[a]} + \frac{\gamma}{2}\sum_a p_t[a]\left((f^\star[a,\cdot] - \hat{Y}_t[a,\cdot])\mathbf{p}_t\right)^2 - \frac{2}{\gamma}\sum_a \frac{q[a]}{p_t[a]} + \frac{5K}{\gamma} \\ &= \frac{\gamma}{2}\sum_a p_t[a]\left((f^\star[a,\cdot] - \hat{Y}_t[a,\cdot])\mathbf{p}_t\right)^2 + \frac{5K}{\gamma} \\ &= \frac{\gamma}{2}\sum_a p_t[a]\left(\mathbb{E}_{b\sim\mathbf{p}_t}(f^\star[a,b] - \hat{Y}_t[a,b])\right)^2 + \frac{5K}{\gamma} \\ &\le \frac{\gamma}{2}\sum_{a,b} p_t[a]p_t[b](f^\star[a,b] - \hat{Y}_t[a,b])^2 + \frac{5K}{\gamma},\end{aligned}$$

where the first inequality follows from Eqn. (8), the second inequality uses AM-GM inequality: For any $x, y \in \mathbb{R}_+$, $\frac{x+y}{2} > \sqrt{xy}$, and lastly we use $q[a]^2 < q[a]$ in the third inequality.

Then proceeding similar to the proof of Thm. 3, the best-response regret of any learner playing $(a_t, b_t) \sim \mathbf{p}_t \times \mathbf{p}_t$ becomes:

$$\begin{aligned}\text{BR-Regret}_T &:= \sum_{t=1}^{T} \max_{\mathbf{q} \in \Delta_K} \mathbb{E}_{a \sim \mathbf{q}} \mathbb{E}_{a_t \sim \mathbf{p}_t} \left[f^\star(x_t)[a, a_t] \right] = \sum_{t=1}^{T} \max_{\mathbf{q} \in \Delta_K} \mathbf{q}^\top f^\star \mathbf{p}_t \\ &\leq \left[\frac{\gamma \text{Reg}_{\text{Sq}}(T)}{2} + \sum_{t=1}^{T} \frac{5K}{\gamma} \right] = \sqrt{10KT\text{Reg}_{\text{Sq}}(T)}\end{aligned}$$

setting $\gamma = \sqrt{\frac{10KT}{\text{Reg}_{\text{Sq}}}}$ following the similar line of argument as shown in the proof of Thm. 3. This yields the desired $O(\sqrt{KT\text{Reg}_{\text{Sq}}})$ regret guarantee of Alg. 1.

D.3. Regret Bound of Alg. 1 for some special realizability function classes

We analyze some special function classes and derive the BR-Regret of Alg. 1 for each cases using the specific regression oracles (recall the details and notations from Rem. 1 and Appendix B.1):

1. Any finite regression class $\mathcal{F}$ such that $|\mathcal{F}| < \infty$: Since we can have `SqrReg` oracle such that $\text{Reg}_{\text{Sq}}(T) \leq 2 \log |\mathcal{F}|$, this yields $\text{BR-Regret}_T[\texttt{MinMaxDB}] = O(\sqrt{KT \log |\mathcal{F}|})$. For example in the standard K-armed (non-contextual) dueling bandit setting, one can construct $\mathcal{F}$ such that $\log |\mathcal{F}| = O\big(K^2 \log(KT)\big)$, and hence implying the BR-Regret(T) of Alg. 1 to be $O\big(\sqrt{K^3 T \log(KT)}\big)$ in this case, which is though $O(K)$ multiplicative factor worse than the optimal rate Dudík et al. (2015).

2. Class of linear predictors $\mathcal{F} := \{(x, a) \mapsto \langle \theta, x_a \rangle \mid \theta \in \mathbb{R}^d, \|\theta\| \leq 1\}$. Here we can have `SqrReg` oracle with $\text{Reg}_{\text{Sq}}(T) \leq d \log(T/d)$, which yields $\text{BR-Regret}_T[\texttt{MinMaxDB}] = O(\sqrt{dKT \log(T/d)})$.

3. Alternatively, for linear predictors, instantiating `SqrReg` as online gradient descent guarantees $\text{Reg}_{\text{Sq}}(T) \leq O(\sqrt{T})$ and hence BR-Regret$(T) \leq O(K^{1/2} T^{3/4})$.

4. Class of generalized linear predictors $\mathcal{F} := \{(x, a) \mapsto \sigma(\langle \theta, x_a \rangle) \mid \theta \in \mathbb{R}^d, \|\theta\| \leq 1\}$. Since we can have `SqrReg` oracle such that $\text{Reg}_{\text{Sq}}(T) \leq d \log T / \kappa_\sigma^2$, one can achieve $\text{BR-Regret}_T[\texttt{MinMaxDB}] = O\Big(\sqrt{\frac{dKT \log T}{\kappa_\sigma^2}}\Big)$.

5. Reproducing Kernel Hilbert Space (RKHS) $\mathcal{F} := \{f \mid \|f\|_{\mathcal{H}} \leq 1, \mathcal{K}(x_a, x_a) \leq 1\}$: Since we can have `SqrReg` oracle such that $\text{Reg}_{\text{Sq}}(T) \leq O(\sqrt{T})$, one can achieve $\text{BR-Regret}_T[\texttt{MinMaxDB}] = O(\sqrt{KT^{3/4}})$.

6. Banach Spaces $\mathcal{F} := \{(x, a) \mapsto \langle \theta, x_a \rangle \mid \theta \in \mathcal{B}, \|\theta\| \leq 1\}$. Since we can have `SqrReg` oracle such that $\text{Reg}_{\text{Sq}}(T) \leq (T/D)^{1/2}$, one can achieve $\text{BR-Regret}_T[\texttt{MinMaxDB}] = O(\sqrt{KD^{-1/2} T^{3/4}})$.