JMLR: Workshop and Conference Proceedings vol 49:1–24, 2016

Instance-dependent Regret Bounds for Dueling Bandits

Akshay Balsubramani[*] ABALSUBR@UCSD.EDU
UC San Diego, CA, USA

Zohar Karnin ZKARNIN@YAHOO-INC.COM
Yahoo! Research, New York, NY, USA

Robert E. Schapire SCHAPIRE@MICROSOFT.COM
Microsoft Research, New York, NY, USA

Masrour Zoghi[*] M.ZOGHI@UVA.NL
University of Amsterdam, Netherlands

Abstract

We study the multi-armed dueling bandit problem in which feedback is provided in the form of relative comparisons between pairs of actions, with the goal of eventually learning to select actions that are close to the best. Following Dudík et al. (2015), we aim for algorithms whose performance approaches that of the optimal randomized choice of actions, the von Neumann winner, expressly avoiding more restrictive assumptions, for instance, regarding the existence of a single best action (a Condorcet winner). In this general setting, the best known algorithms achieve regret $\tilde{O}(\sqrt{KT})$ in T rounds with K actions. In this paper, we present the first instance-dependent regret bounds for the general problem, focusing particularly on when the von Neumann winner is sparse. Specifically, we propose a new algorithm whose regret, relative to a unique von Neumann winner with sparsity s, is at most $\tilde{O}(\sqrt{sT})$, plus an instance-dependent constant. Thus, when the sparsity is much smaller than the total number of actions, our result indicates that learning can be substantially faster.

Keywords: dueling bandits, online learning, game theory.

1. Introduction

In many application domains, feedback can be more reliable and easier to obtain when given in the form of a preference between a pair of items or choices, rather than assigning a real-valued score or reward to individual options. For instance, humans may find it quite natural to select which of two news articles they are more interested in reading, and each such selection between a pair of articles can provide feedback to a learning system regarding user preferences. Similar examples include ranker evaluation in information retrieval (Joachims, 2002; Yue and Joachims, 2011; Hofmann et al., 2013), ad placement (Ciaramita et al., 2008; Pandey et al., 2007), and recommender systems (Park and Chu, 2009).

The *multi-armed dueling bandit problem* of (Yue et al., 2012) provides a model for studying such preference-learning problems. In this setting, the learner chooses from a pool of K possible *actions*, that is, options or choices, also called "arms." On each of a series of rounds, the learner selects just two of the actions, and a *duel* is held between these two actions in which one or the other is determined (stochastically) to be preferred. For instance, each round might model a website selecting two news articles for presentation to a user, with the duel and its outcome corresponding

[*] Part of this research was conducted during an internship with Microsoft Research.

to the user's selection between the two articles. The goal is for the learner over time to select the "best" actions on each round.

When individual actions yield scalar rewards, the best action is naturally the one producing highest expected reward. But in the dueling-bandits setting, which is based solely on relative comparisons between pairs of actions, merely defining what is meant by "best" is a critical challenge. One of the main approaches to this problem assumes the existence of a *Condorcet winner*, that is, a single action that wins any duel against any other action with probability exceeding $1/2$. When such an action exists, it is clearly the best. Several methods have been proposed for dueling bandits under this assumption, including Interleaved Filter (Yue et al., 2012), Beat the Mean (Yue and Joachims, 2011), Relative Confidence Sampling (Zoghi et al., 2014a), Relative Upper Confidence Bound (RUCB) (Zoghi et al., 2014b), Doubler and MultiSBM (Ailon et al., 2014), mergeRUCB (Zoghi et al., 2015b) and Relative Minimum Empirical Divergence (Komiyama et al., 2015). Often, additional assumptions are also needed, for instance, regarding transitivity among the actions. However, as observed by Dudík et al. (2015) and Zoghi et al. (2015a), real-world problems do not always have Condorcet winners.

Although there might not exist a *single* best action in the sense above, it was shown by Dudík et al. (2015) (and previously by Rivest and Shen (2010) in a related setting), that there always must exist a *distribution* over actions that is best in a similar sense. In other words, there must always exist a *randomized* way of choosing an action that will win a duel against any other action with probability at least $1/2$. The existence of such a distribution, called a *von Neumann winner*, follows from a game-theoretic view of dueling bandits together with a simple invocation of the minmax theorem for zero-sum games.

Roughly speaking, the goal now becomes that of learning to choose actions to achieve performance almost as good as that attainable using a von Neumann winner, that is, a best choice of actions. The difference between actual and optimal performance is the *regret*. For this problem, the sparring Exp3 algorithm (Ailon et al., 2014; Dudík et al., 2015) is known to achieve regret $\tilde{O}(\sqrt{KT})$ in T rounds. This bound is worst-case or *instance-independent* in the sense that it holds for all instances of this problem involving K actions.

In fact, under the strong assumption that a Condorcet winner exists (possibly along with other assumptions), the algorithms listed above for this case achieve a much more favorable bound of $O(K \ln T)$. However, these bounds are *instance-dependent* in the sense that they hide an additive constant that depends on the particular instance, that is, on the underlying probabilities controlling the outcomes of pairwise duels. Such bounds can give a more fine-grained understanding of how learning will progress on a single, fixed instance as T becomes large. They can be substantially better than instance-independent bounds for the same problem. Indeed, these much improved instance-dependent bounds under the restrictive Condorcet assumption suggest that the instance-independent bounds for the general case may be overly pessimistic on many instances.

In this paper, we prove the first such instance-dependent bounds for a very general case, without relying on restrictive Condorcet or transitivity assumptions. We focus specifically on the case in which the von Neumann winner is *sparse*, meaning that its support is concentrated on just a few of the actions. In other words, we suppose that there are a handful of "good" actions, with the rest being strictly inferior. For our analysis, we also assume that the von Neumann winner is unique, which is known to be the case in almost all instances (Owen, 1995, Exercise II.6). Under these conditions, we give an algorithm called SPAR2 whose instance-dependent regret is $\tilde{O}(\sqrt{sT})$, plus a term that is constant relative to T, though dependent on the particular instance. Here, s is the sparsity (number

of nonzero elements) of the unique von Neumann winner. Thus, when the sparsity s is much smaller than the total number of actions K, our result indicates that learning can be substantially faster.

Our algorithm is based on maintaining confidence bounds on relevant probability estimates, and using these to explicitly eliminate actions which cannot belong to the support of the von Neumann winner. For the remaining actions, previous methods, such as sparring Exp3, can be applied.

Although we present our results within the context of dueling bandits, they apply more generally to the problem of repeatedly playing any symmetric zero-sum game.

2. Dueling Bandits and von Neumann Winners

In the *multi-armed dueling bandits problem* (Yue et al., 2012), the learner has access to K actions, $1, \dots, K$. On each round $t = 1, \dots, T$, the learner must select two of these actions i and j. A *duel* is then held between these actions, with an outcome of $+1$ if i wins the duel, and -1 if j wins. We assume the outcome of each duel is independent, and that the probability of i beating j is fixed but unknown. In particular, we define a matrix $\mathbf{P}$ whose (i, j) entry P_{ij} is the expected outcome of a duel between i and j; this is equivalent to saying that the probability of i beating j is $(1 + P_{ij})/2$. For now, we assume only that this matrix $\mathbf{P}$ is skew-symmetric, meaning that $\mathbf{P} = -\mathbf{P}^\top$, which in this setting means simply that a duel (i, j) is equivalent to the negative of a duel (j, i), as is natural.

Roughly, the learner aims to select actions that are close to the best. As discussed in §1, there might not exist a single best action, a so-called Condorcet winner. Instead, Dudík et al. (2015) propose viewing the matrix $\mathbf{P}$ as defining a zero-sum game. Then von Neumann's minmax theorem, together with $\mathbf{P}$'s skew-symmetry, imply the existence of a probability vector $\mathbf{w} \in [0, 1]^K$ (with elements summing to 1) for which $\mathbf{w}^\top \mathbf{P} \mathbf{v} \geq 0$ for all probability vectors $\mathbf{v}$. Such a vector $\mathbf{w}$, which is really just a maxmin strategy for the game $\mathbf{P}$, is called a *von Neumann winner*. The randomized strategy defined by $\mathbf{w}$ is best in the sense that, for any action j, if i is chosen at random according to the distribution defined by $\mathbf{w}$, then i will beat j in a duel with probability at least $1/2$. (The same is true also if j is itself chosen at random from any fixed distribution.) Thus, we want the learner's performance to approach that of a von Neumann winner.

More precisely, let us write $(i(t), j(t))$ for the pair of actions selected on round t. Following Dudík et al. (2015), we define the *regret* for such a sequence of choices to be

$$\max_{k \in [K]} \sum_{t=1}^{T} \frac{P_{k,i(t)} + P_{k,j(t)}}{2} \tag{1}$$

where $[K] = \{1, \dots, K\}$. This regret is always nonnegative as can be seen by noting that if k is chosen randomly according to a von Neumann winner, then the expected value of each term of the sum, and therefore the entire sum, will all be nonnegative; thus, the maximum over k is as well. Furthermore, if $i(t)$ and $j(t)$ are themselves chosen by a von Neumann winner, then the expected value of the sum will be nonpositive. Therefore, the regret is minimized and equal to zero when actions are chosen according to a von Neumann winner.

Summarizing, this definition of regret thus measures how the learner's performance in its choice of actions varies from the optimal performance, which can always be realized, in expectation, using a von Neumann winner.

Algorithm 1 Sparse Sparring (SPAR2)

Input: A von Neumann dueling bandit problem, an exploration parameter $\alpha > \frac{1}{2}$, a horizon $T \in \{1, 2, \ldots\}$ and probability of failure $\delta > 0$

$\mathbf{W} = [W_{ij}] \leftarrow \mathbf{0}_{K \times K}$ // 2D array of wins: W_{ij} is the number of times action i beat action j

$\mathbf{N} = [N_{ij}] \leftarrow \mathbf{0}_{K \times K}$ // 2D array of plays: N_{ij} is the number of times action i was compared against action j

$\mathcal{A} \leftarrow \{1, \ldots, K\}$ // Active actions

Initialize two copies of Exp3.P with actions $\mathcal{A}$: a row copy Exp3.P$_R$ and a column copy Exp3.P$_C$

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

  $U_{ij} = \frac{W_{ij}}{N_{ij}} + \sqrt{\frac{\alpha \ln(2/\delta)}{N_{ij}}}$ for all $i, j \in \{1, \ldots, K\}$ with $i \neq j$

  $L_{ij} = \frac{W_{ij}}{N_{ij}} - \sqrt{\frac{\alpha \ln(2/\delta)}{N_{ij}}}$ for all $i, j \in \{1, \ldots, K\}$ with $i \neq j$

  // We use the convention that $\frac{0}{0} = 0$ and $\frac{x}{0} := \infty$ for any $x \neq 0$.

  $U_{ii} = 0$ and $L_{ii} \leftarrow 0$ for each $i = 1, \ldots, K$

  $\mathbf{Q} := [Q_{ij}]$ with $Q_{ij} = \frac{W_{ij}}{N_{ij}}$ for $i \neq j$ and $Q_{ii} = 0$

  // Frequentist estimate of preference matrix underlying the dueling bandit problem

  $\mathcal{C} \leftarrow$ the set of preference matrices $\mathbf{R} = [R_{ij}]$ such that $L_{ij} \leq R_{ij} \leq U_{ij}$ for all i, j

  $\mathcal{E} \leftarrow$ output of Algorithm 2 with $\mathbf{Q}$ and $\mathcal{C}$ as the input

  if $\mathcal{E} \neq \varnothing$ **then**

    $\mathcal{A} \leftarrow \mathcal{A} \setminus \mathcal{E}$

    Modify Exp3.P$_R$ and Exp3.P$_C$ by setting the weights of the actions in $\mathcal{E}$ to zero and removing them from further consideration.

  end if

  Select actions $r, c \in \mathcal{A}$ by querying Exp3.P$_R$ and Exp3.P$_C$, respectively.

  Compare actions r and c and set $o = \begin{cases} 1 & \text{if } r \text{ won} \\ -1 & \text{if } c \text{ won} \end{cases}$

  Give feedback o to Exp3.P$_R$ and $-o$ to Exp3.P$_C$.

  Set $W_{rc} += o$ and $W_{cr} -= o$.

  Increment both N_{rc} and N_{cr}.

end for

3. The Algorithm

In this section, we give a basic description of our proposed algorithm, SPAR2. In the next section §4, the interested reader can find more detailed justifications for some of the specific choices made in the algorithm together with a summary of the proof techniques used in the analysis of SPAR2.

Our solution builds upon the existing, straightforward solution of sparring two independent copies of Exp3 (presented as *Sparring* in Ailon et al. (2014)). For clarity, we provide a short review of it. With this technique, we maintain two independent copies of any algorithm for the classic *Multi Armed Bandit* problem, e.g. the Exp3 algorithm. We refer to them as the row and column instances. In each round, the row instance plays action i and the column instance plays action j. The

Algorithm 2 Sub-procedure for action elimination

Input: A skew-symmetric matrix $\mathbf{Q} \in \mathbb{R}^{K \times K}$, and confidence region $\mathcal{C} \subseteq \mathbb{R}^{K \times K}$ around $\mathbf{Q}$.

Return: A subset $\mathcal{E} \subseteq [K]$ of actions that can be eliminated from being in the von Neumann support of *any* matrix $\mathbf{R} \in \mathcal{C}$.

1: Find a von Neumann winner $\mathbf{v}$ of $\mathbf{Q}$.

2: Set $S \leftarrow \mathrm{supp}(\mathbf{v})$, and $s \leftarrow |S|$.

3: Set $\mathbf{Q}^0 \leftarrow \mathbf{Q}_{S,S}$, the minor of $\mathbf{Q}$ formed by the rows and columns corresponding to indices that are in S.

4: For $i \in [s]$, let $\mathbf{Q}_i$ be the matrix obtained by replacing the i^{th} column of $\mathbf{Q}^0$ with the all 1 column vector.

5: Set $\sigma = \max_i \sigma_s(\mathbf{Q}_i)$ where σ_j is the j'th largest singular value of a matrix.

6: If $\sigma = 0$, return the empty set.

7: Set $\epsilon_i = v_i$ for $i \in S$ and $\epsilon_i = (\mathbf{v}^\top \mathbf{Q})_i$ for $i \notin S$, where $(\mathbf{v}^\top \mathbf{Q})_i$ denotes the i^{th} coordinate of the vector $\mathbf{v}^\top.\mathbf{Q}$

8: Set $\Delta = \frac{2}{9}\sigma \cdot \min\left\{\min_{i \in S} \epsilon_i, \min_{i \notin S} \frac{\epsilon_i}{\|\mathbf{Q}_{\{i\},S}\|}\right\}$

9: If there exists a $\mathbf{R} \in \mathcal{C}$ with $\sqrt{\sum_{i,j \in S} (Q_{ij} - R_{ij})^2} > \Delta$, return the empty set.

10: For $i \notin S$, let $\rho_i = \max_{\mathbf{R} \in \mathcal{C}} \sqrt{\sum_{j \in S} (Q_{ij} - R_{ij})^2}$. If $\rho_i > \epsilon_i/3$ for some $i \notin S$, return the empty set.

11: Return $[K] \setminus S$

pair played by the dueling bandit solver is (i, j). One of the actions wins the duel, and its instance is given a gain of 1; the other, losing instance gets a gain of -1. The two algorithms' inputs and outputs therefore depend on each other.

It is rather straightforward to show that as long as the algorithms have a regret guarantee of the form $\widetilde{O}(\sqrt{KT})$ against an adaptive adversary, the empirical distribution of actions played by both players is guaranteed to be close to the von-Neumann winner of $\mathbf{P}$ (Dudík et al., 2015, §5). This translates into a regret guarantee of the form $\widetilde{O}(\sqrt{KT})$ in our setting. The main observation in this paper is that although this bound may be tight in the worst-case scenario, typical game matrices $\mathbf{P}$ tend to be easier.

We focus on the case where the matrix has a unique and sparse von Neumann winner. As mentioned in the introduction, almost all preference matrices $\mathbf{P}$ have a unique von Neumann winner, and matrices that arise in practice often have one with sparse support. So the matrix of interest typically satisfies these criteria as well. In a setting with K actions but only $s \ll K$ non-zero values in the von Neumann winner, we manage to prove a high probability bound on the regret of $\widetilde{O}(\sqrt{sT} + C(\mathbf{P}))$, where $C(\mathbf{P})$ is an instance dependent parameter of $\mathbf{P}$.

The idea behind our algorithm is quite simple: during the run of the sparring algorithm we gradually get a clearer estimate of $\mathbf{P}$ formally defined as some confidence region $\mathcal{C}$. Once it becomes clear that an action is not in the support of any matrix $\mathbf{Q} \in \mathcal{C}$, the action can be excluded. This idea naturally gives rise to a computationally inefficient scheme: in each round, compute the union of

the supports of the von Neumann winners of all matrices in $\mathcal{C}$, excluding any action outside of this union. Now, if the non-supported actions are excluded at an early stage, then by existing bounds for sparring Exp3, the regret will be $\widetilde{O}(\sqrt{sT})$.

This is the basic idea behind our SPAR2 algorithm, outlined in Algorithm 1. However, in order to remedy the computational inefficiency of the above intractable algorithm, we replace the above condition on $\mathcal{C}$ by a more conservative condition that has the advantage of being easier to check. More specifically, the conditions in Lines 9 and 10 of Algorithm 2 ensure that the von Neumann winners of all preference matrices in the confidence region $\mathcal{C}$ have identical supports. Assuming that we know with high probability that our underlying matrix $\mathbf{P}$ lies inside $\mathcal{C}$, then we can safely discard actions that cannot be in the support of the von Neumann winner of $\mathbf{P}$. This is done by placing high-probability confidence intervals around our frequentist estimates of the entries of $\mathbf{P}$ (cf. Lines 6-8 of Algorithm 1) and using the resulting constraints as the confidence region (Line 10 of Algorithm 1).

Next, we provide a brief overview of Algorithm 2. As mentioned above, the goal of this subroutine is to facilitate verifying whether or not a confidence region $\mathcal{C}$ around a given preference matrix $\mathbf{Q}$ with von Neumann winner $\mathbf{v}$ (Line 1) contains a preference matrix $\mathbf{R}$ whose von Neumann winner is supported on a different set of actions than the support of $\mathbf{v}$: if not, we call $\mathcal{C}$ "pure." The algorithm provides two conditions (Lines 9 and 10) which can easily be checked if the conditions defining the confidence region $\mathcal{C}$ are given in terms of independent constraints on the entries of the matrix $\mathbf{R}$, as is the case with the constraints in Line 10 of Algorithm 1. There are two types of quantities used by the two conditions imposed by Algorithm 2: the first one is σ (Line 5), which is the smallest singular value of a modification of the submatrix of $\mathbf{Q}$ that corresponds to the actions on which $\mathbf{v}$ is supported (Lines 3-4); the second type of quantity used to test the purity of $\mathcal{C}$ are the ϵ_i (Line 7) which roughly speaking are a measure of how far an action is from leaving or entering the support of $\mathbf{v}$ if $\mathbf{Q}$ were to be perturbed. In §4.2.1, we provide a more intuitive explanation for why these quantities are useful in accomplishing what Algorithm 2 is tasked with.

In this paper, we prove the following regret bound for our algorithm SPAR2:

Theorem 1 *Suppose we are given a multi-armed dueling bandit problem, whose preference matrix $\mathbf{P}$ has a unique von Neumann winner that puts nonzero weight on s actions, a time horizon T and a probability of failure δ. Then, applying Algorithm 1 to this problem leads to a regret that is bounded by the minimum of $\widetilde{O}(\sqrt{sT\log(s/\delta)} + C(\mathbf{P})\log(1/\delta)^2)$ and $O(\sqrt{KT\log(K/\delta)})$ with probability at least $1-\delta$ for some constant $C(\mathbf{P})$.*

4. Our Techniques

In this section, we provide an outline of the main ingredients of the proof of Theorem 1, with the goal of assisting the reader in developing an intuition for both the algorithm and the analysis. For rigorous proofs and precise definitions we invite the reader to consult §5.

4.1. An Initial Idea

As has been shown in the literature (Dudík et al., 2015; Zoghi et al., 2015a), the dueling bandit problems that arise in practice tend to have von Neumann winners that have much smaller support size than the total number of actions. Therefore, if at some point we could discard the actions outside the von Neumann winner support, we could thereafter deal with a much smaller dueling

bandit problem, which has great computational and convergence benefits. Such reasoning suggests the following notional scheme for dueling bandits in the von Neumann setting:

1. Apply any algorithm $\mathbb{A}$ for the von Neumann dueling bandit problem with a regret bound of the form $\widetilde{O}(\sqrt{KT})$ (e.g. Sparring Exp3.P as in Ailon et al. (2014); Dudík et al. (2015)) to a dueling bandit problem defined by $\mathbf{P}$.

2. Use the outcomes of the comparisons carried out by the algorithm to maintain a high probability confidence region $\mathcal{C}$ around the empirical estimate $\mathbf{Q}$ of $\mathbf{P}$.

3. Continue running $\mathbb{A}$ until a time comes when for every preference matrix $\mathbf{R} \in \mathcal{C}$, the von Neumann winners of $\mathbf{Q}$ and $\mathbf{R}$ are supported on the same set of actions.

4. Eliminate the actions outside the common support of the von Neumann winners (call it S), and continue applying $\mathbb{A}$ to the smaller set S of actions.

Let us denote this modification of $\mathbb{A}$ by SPAR2$_{\mathbb{A}}$. Note that SPAR2$_{\mathbb{A}}$ can only improve upon $\mathbb{A}$, which can be seen by considering the following two scenarios: if Step 4 is never reached, then SPAR2$_{\mathbb{A}}$ will be identical to $\mathbb{A}$; alternatively, if Step 4 is reached, then the cumulative regret of SPAR2$_{\mathbb{A}}$ will be in $O(\sqrt{sT})$, where s is size of the support of $\mathbf{w}$: this is because of our assumption on the regret bound for $\mathbb{A}$. So this approach seems a plausible way to solve the problem with low regret.

However, there are two hurdles to overcome to make these ideas into an efficient algorithm.

1. **Computation:** It is not specified how to efficiently check that the von Neumann winners of matrices in $\mathcal{C}$ have the same support. There appears to be no element-wise way to do so, as existed for the Condorcet winner.

2. **Convergence with Low Regret:** It is not even clear that it is possible to eliminate actions with certitude and drive this procedure to its final phase – Algorithm $\mathbb{A}$ is oblivious to the confidence region and so might not shrink $\mathcal{C}$ enough for us to be able to eliminate actions. Action elimination to find the von Neumann winner requires more care than is typically necessary with such algorithms (see §4.3).

4.2. Devising an Efficient Algorithm

We now give a detailed overview of how our proof proceeds to address the above computation and convergence issues of SPAR2$_{\mathbb{A}}$, to derive the efficient algorithm SPAR2 and prove Theorem 1. The interested reader can find the formal proofs in §5.

4.2.1. EFFICIENT COMPUTATION

To translate the action elimination idea to our situation, recall that the suboptimal actions are those outside of the support of the von Neumann winner. Therefore, given a confidence region $\mathcal{C}$ of possible preference matrices, we must be able to determine whether an action is outside of the support of the von Neumann winners of all $\mathbf{P} \in \mathcal{C}$. When the support size s is 1, each matrix in $\mathcal{C}$ must have a Condorcet winner and it is easy to do this: if there exists some pair i, j such that $\mathbf{P}_{ij} > 0$ for all $\mathbf{P} \in \mathcal{C}$, then action j can be eliminated. This element-wise scheme has the advantage

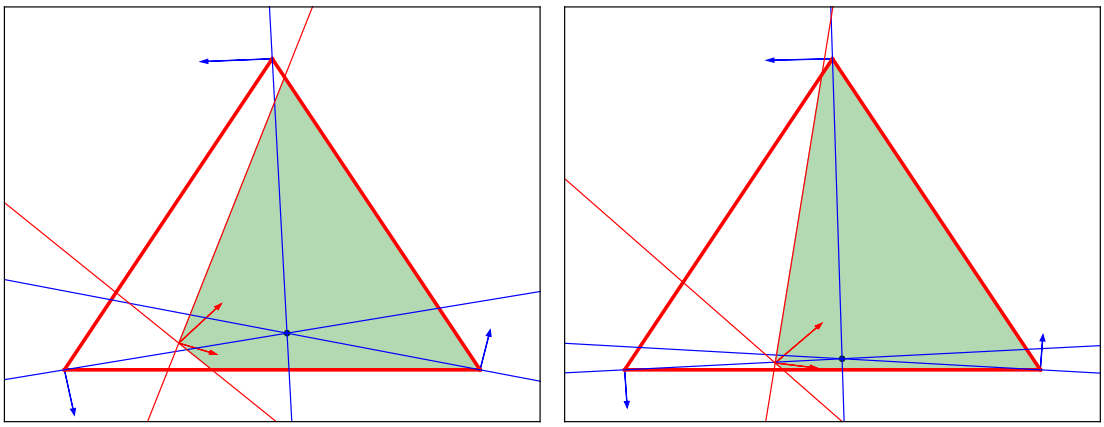

Figure 1: A pictorial representation of the conditions defining a von Neumann winner: the left plot depicts the von Neumann winner of a matrix $\mathbf{Q}$ and the right plot that of a small perturbation of $\mathbf{Q}$.

of being computationally tractable. However, when the support size $s > 1$, there is no obvious way to generalize it.

In this vein, our first contribution is an analogous result for preference matrices that have unique von Neumann winners, which is known to be the generic case encompassing almost all preference matrices (Owen, 1995, Exercise II.6). In §5.3, we show that, given a preference matrix $\mathbf{Q}$, the efficiently computable quantities Δ and ϵ_i, defined in Lines 9 and 10 of Algorithm 2, can be used to place proximity conditions on other preference matrices $\mathbf{R}$ to ensure that their von Neumann winners have the same support as $\mathbf{Q}$. So if we can empirically estimate the true preference matrix $\mathbf{P}$ with some matrix $\mathbf{Q}$ so that the easily checkable conditions in Algorithm 2 are satisfied, then we can indeed identify and retain only actions in the support of $\mathbf{Q}$'s von Neumann winner – they are exactly the actions in the support of the von Neumann winner of the true matrix $\mathbf{P}$.

We more precisely explain the ideas behind the quantity $\Delta(\mathbf{P})$ (in Lemma 6). Consider the matrix $\mathbf{P}$, its von Neumann winner $\mathbf{w}$ and its support, which we assume w.l.o.g. to be $[s]$. As discussed in §5.2, the vector $\mathbf{w}$ is a solution to the linear system of equations defined by $(\mathbf{w}^\top \mathbf{P})_i = 0$ for all $i \in [s]$, $\sum_{i=1}^{s} w_i = 1$, and $w_i = 0$ for all $i > s$. When considering the first s entries of $\mathbf{w}$, this can be viewed as a set of $s + 1$ equations with s variables. Since $\mathbf{w}$ is unique, it must be the only solution to this equation system, meaning the corresponding matrix is of rank s.

In our proof, we view the von Neumann winner as the solution to the above system of equations. This point is illustrated in the left plot of Figure 1, where the blue lines represent the linear equations determining the non-zero values of $\mathbf{w}$; their intersection is the point $\mathbf{w}$. The red lines represent the columns of $\mathbf{P}$ numbered $s + 1$ through K. For $\mathbf{w}$ to be a von Neumann winner, it must remain on the positive side of these lines: i.e. $\mathbf{w}^\top \mathbf{P} > 0$, the region shaded green.

When considering a matrix $\mathbf{Q}$ that is close enough to $\mathbf{P}$, as prescribed by Algorithm 2, all lines move an amount proportional to Δ and the ϵ_i's. Our key observation here is the following: as long as the blue lines describe a *sufficiently well-conditioned system*, the point $\mathbf{w}$ which is the intersection of the blue lines moves by a distance proportional to Δ; similarly, as long as the perturbation of the vectors defining the red lines are small enough, then $\mathbf{w}$ will remain a bounded distance away from

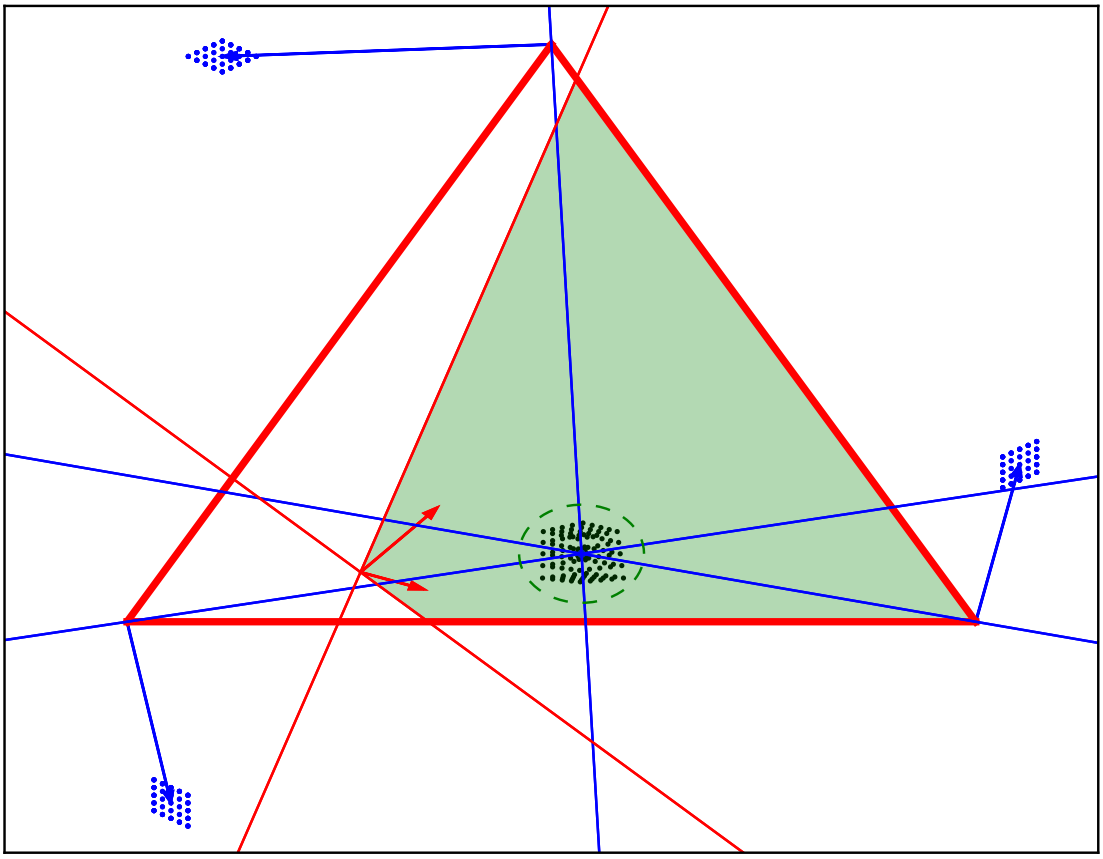

Figure 2: A pictorial representation of the perturbation of the sub-preference matrix corresponding to the arms in the support.

the red lines. This is illustrated in the right plot of Figure 1, which illustrates what would happen if we were to perturb the preference matrix depicted in the left plot by a small amount.

More specifically, Figure 2 provides a more detailed description of how the von Neumann winner $\mathbf{w}$ is affected by perturbations in the column vectors of the submatrix $\mathbf{P}^0$ (where $\mathbf{P}^0$ is the sub-preference matrix corresponding to the arms in the support of the von Neumann winner): the blue dots around the tips of the blue vectors represent an equispaced grid tiling a box around each column of $\mathbf{P}^0$ (recall that $\mathbf{P}^0$ is equal to zero along the diagonal, so each column has $s-1$ nonzero entries); on the other hand, the black dots near the intersection of the three blue lines represent the resulting perturbations in the von Neumann winner: each black dot corresponds to three blue dots chosen, each of them coming from a different cluster of blue dots. As this figure indicates, the resulting set of possible von Neumann winners can have a rather complicated shape, and so rather than dealing with it directly, we inscribe it inside a larger ball and insist for this larger ball to be contained inside the green region.

More precisely, the algebraic quantity determining how far $\mathbf{w}$ can move is the s'th singular value of the matrix of equations detailed above, that measures how far away this matrix is from a lower rank matrix. To conclude, the bound is obtained by making sure that the matrices are close enough

so that $\mathbf{w}$ does not cross any red lines out of the shaded region, and remains a valid distribution. Formalizing this intuition eventually leads to a measure of proximity between matrices that depends on (a) the values w_i for $i \leq s$, (b) the values $(\mathbf{w}^\top \mathbf{P})_i$ for $i > s$, and (c) the s'th singular value of the matrix corresponding to the linear equation system described by the blue lines.

This reasoning gives us the efficient prescription of Algorithm 2 for identifying actions outside the support of all preference matrices in $\mathcal{C}$.

4.2.2. CONVERGENCE WITH LOW REGRET

As we mentioned earlier, it is not clear that actions can be eliminated using this method simply because sparring Exp3.P might very well not shrink the confidence region $\mathcal{C}$ enough for Algorithm 2 to return anything other than the empty set. However, a careful inspection of the conditions in Lines 9 and 10 of Algorithm 2 reveals the following important fact: as far as the proximity of the matrices $\mathbf{Q}$ and $\mathbf{R}$ are concerned, we only care about the entries i, j with either i or j in $[s]$. Note that this pertains to an $s \times K$ submatrix of the full $K \times K$ preference matrix $\mathbf{P}$, so we simply need to show that this smaller submatrix is queries enough.

We show that the low regret guarantees of Exp3.P translate into a guarantee that the actions in $[s]$ are queried a constant fraction of the time (Lemmas 12 and 13). This, combined with an assumption that all actions are queried at least a fraction of (roughly) $1/\sqrt{T}$ of the time allows us to find a finite upper bound to the time required for two copies of Exp3.P to obtain a confidence bound in which all matrices share the same von Neumann support (Corollary 17).

4.3. Related Work

Our reasoning follows the intuition behind the *action elimination* family of algorithms (see e.g. Even-Dar et al. (2002)), where suboptimal actions are excluded once it is clear with high probability that they are indeed suboptimal. These algorithms, when applied to the classic MAB problem, track a confidence interval around each action's reward, and eliminate an action once all reward settings within these confidence intervals are such that the action is suboptimal. Typically, the action elimination algorithms use a uniform sampling approach, but one can consider algorithms that use other sampling techniques that may require more time to eliminate actions, but eventually result in smaller regret bounds. Although such an approach is not required in the MAB setting, in our setting we must sample the actions non-uniformly in order to guarantee a min-max regret bound of $\sqrt{T}$ rather than $T^{2/3}$ for the problem, as is sometimes the case for ϵ-first approaches.

When removing actions from contention, the poorest-performing actions are the most natural ones to choose to eliminate. However, it is not clear how this can be done in our setting. For instance, there could be an action a_1 which is included in the von Neumann winner solely because it is uniquely capable of beating some other action a_2, while a_1 itself performs poorly against almost all other actions (and contending von Neumann winners). For instance, consider the case of three actions, with a_1 always beating a_2, and a_2 beating a_3 with some small probability δ, and a_1 tying with a_3, so that the von Neumann winner is just a_1. Then a_3 is δ-close to being the von Neumann winner, so it needs to be played; but the worse-than-von-Neumann action a_2 cannot be eliminated, because it is the only way of distinguishing a_3 from being a von Neumann winner. In short, the algorithm should not eliminate actions like a_2, whose deceptively poor performance belies their importance to achieving low regret.

Our algorithm sidesteps this concern, because our arguments motivating $\Delta(\mathbf{P})$ make it clear that we are not necessarily removing actions just based on their individual performance. Rather, we depend in a more complex way on actions' interactions through the geometry of the constraint set represented by $\Delta(\mathbf{P})$.

Furthermore, there is another group of algorithms that do not assume the existence of a Condorcet winner and instead try to find a different notion of winner, such as Copeland (Urvoy et al., 2013; Zoghi et al., 2015a; Komiyama et al., 2016), Borda (Busa-Fekete et al., 2013, 2014; Jamieson et al., 2015) or Random Walk (Negahban et al., 2012; Busa-Fekete et al., 2013; Chen and Suh, 2015). As argued in (Dudík et al., 2015), from a preference learning point of view, the von Neumann winner is more preferable as a solution concept than these other notions that have their roots in other fields such as social choice theory: this is because in a head-to-head comparison, the von Neumann winner is preferred to any other choice of arms.

5. Analysis

5.1. Notation

Let us begin by fixing some notation: $\mathbb{R}^k$ denotes the vector space of k-dimensional vectors or real numbers. In what follows all vectors are given in bold letters and are row vectors. Also, $\mathbf{e}_j$ denotes the j^{th} basis vector that is 1 in the j^{th} coordinate and zero elsewhere. The dimension of $\mathbf{e}_j$ will be clear from context or stated otherwise. We also denote by $\mathbf{0}_k$ and $\mathbf{1}_k$ the k-dimensional vectors, whose coordinates are all zeros and ones, respectively. The *k-simplex* Δ^k is the subset of $\mathbb{R}^k$ consisting of vectors whose coordinates are all non-negative and sum to 1. Recall that Δ^k is the convex hull of the k basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_k$

Definition 2 *Given a matrix $\mathbf{M}$ and an ordered subset of indices $S \subseteq \{1, \dots, K\}$, we denote by $\mathbf{M}_S$ the matrix obtained from extracting the i, j entries of $\mathbf{M}$ for all $i, j \in S$. If S is equal to consecutive indices $S = \{i, i+1, \dots, i'\}$, we also use the notation $\mathbf{M}_{i:i'} := \mathbf{M}_S$. Also, more generally, given ordered sets of indices S and S', the notation $\mathbf{M}_{S,S'}$ will denote the matrix consisting of the numbers $M_{i,j}$ for all $(i, j) \in S \times S'$.*

Definition 3 *Given a preference matrix $\mathbf{P}$ with a unique von Neumann winner $\mathbf{w}$, we define the following entities:*

1. *We define the* support *of $\mathbf{P}$ to be the set of actions k such that the k^{th} entry of $\mathbf{w}$ is non-zero and use the notation*

$$\operatorname{supp} \mathbf{P} := \{k \mid s.t.\ w_k > 0\}.$$

   Furthermore, we denote by $s(\mathbf{P})$ the number of actions in $\operatorname{supp} \mathbf{P}$.

2. *Relabeling the actions so that the actions in* $\operatorname{supp} \mathbf{P}$ *are the first $s := s(\mathbf{P})$ actions, we get the following decomposition of $\mathbf{P}$ and its von Neumann winner:*

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}^0 & \mathbf{P}^{01} \\ -\mathbf{P}^{01\top} & \mathbf{P}^1 \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}^0 & \mathbf{0} \end{bmatrix} \tag{2}$$

   with $\mathbf{P}^0$ and $\mathbf{P}^{01}$ being matrices of size $s \times s$ and $s \times (K - s)$ respectively, and similarly $\mathbf{w}^0$ a s-dimensional row vector with positive elements and $\mathbf{0}$ the $(K - s)$-dimensional zero vector.

5.2. Facts about the von Neumann winner

Let us now recall Theorem II.4.4 in (Owen, 1995), restricted to case of a unique von Nuemann winner $\mathbf{w}$, which in our notation states that for $\mathbf{l} = \mathbf{w}^\top \mathbf{P}$,

$$\min_i (\mathbf{w} + \mathbf{l})_i > 0 \quad \text{and} \quad \mathbf{w} \circ \mathbf{l} = \mathbf{0} \tag{3}$$

where $\circ$ denotes the element-wise product of vectors. The second equality is a consequence of complementary slackness in KKT. In other words, the above states that the supports of $\mathbf{w}, \mathbf{w}^\top \mathbf{P}$ form a partition of $[K]$.

Lemma 4 *Suppose we are given a preference matrix $\mathbf{P}$, whose $s \times s$ minor (i.e. $\mathbf{P}_{1:s}$ using the notation of Definition 2) has rank $s-1$ and its 1-dimensional kernel intersects the interior of the s-simplex Δ^s at a point $\mathbf{w}^0$ (i.e. $\mathbf{w}^{0\top} \mathbf{P}_{1:s} = \mathbf{0}$, $\sum_i w_i^0 = 1$ and $\mathbf{w}^0 > \mathbf{0}$) which additionally satisfies $\mathbf{w}^{0\top} \mathbf{P}_{1:s,s+1:K} > 0$. Then, $\mathbf{w} = [\mathbf{w}^0 \ \mathbf{0}_{K-s}]$ is the unique von Neumann winner of $\mathbf{P}$.*

Proof First of all, let us note that $\mathbf{w}$ is a von Neumann winner for $\mathbf{P}$ since we have

$$\begin{aligned}
\mathbf{w}^\top \mathbf{P} &= \begin{bmatrix} \mathbf{w}^0 & \mathbf{0}_{K-s} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{1:s} & \mathbf{P}_{1:s,s+1:K} \\ \mathbf{P}_{s+1:K,1:s} & \mathbf{P}_{s+1:K} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{w}^{0\top} \mathbf{P}_{1:s} & \mathbf{w}^{0\top} \mathbf{P}_{1:s,s+1:K} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{w}^{0\top} \mathbf{P}_{1:s,s+1:K} \end{bmatrix} \geq 0.
\end{aligned}$$

Now, let us assume that $\mathbf{P}$ has a von Neumann winner $\mathbf{v} \neq \mathbf{w}$. By Equation (3), we know that we have $v_{s+1:K} = \mathbf{0}$ since by assumption the last $K-s$ coordinates of $\mathbf{l} := \mathbf{w}^\top \mathbf{P}$ are non-zero, so we also have $\mathbf{v} = [\mathbf{v}^0 \ \mathbf{0}_{K-s}]$, which means that $\mathbf{v}^0$ satisfies $\mathbf{v}^{0\top} \mathbf{P}_{1:s} = \mathbf{0}$ and $\mathbf{v}^0 \in \Delta^s$. But, the assumption that $\mathbf{v} \neq \mathbf{w}$ implies that we also have $\mathbf{v}^0 \neq \mathbf{w}^0$ and since we have $\mathbf{v}^0, \mathbf{w}^0 \in \Delta^s$ we can conclude that $\mathbf{v}^0$ is not a constant multiple of $\mathbf{w}^0$, but this would mean $\mathbf{P}_{1:s}$ has a rank of at most $s-2$ as $\mathbf{v}^{0\top} \mathbf{P}_{1:s} = \mathbf{w}^{0\top} \mathbf{P}_{1:s} = 0$, and $\mathbf{v}^0$ and $\mathbf{w}^0$ are linearly independent. This contradicts our assumption that $\mathbf{P}_{1:s}$ has rank $s-1$ and so $\mathbf{v} \neq \mathbf{w}$ cannot be true. ■

5.3. The Stability of the von Neumann Support

In this section we analyze the proximity requirement of a pair of matrices for them to have the same von Neumann support. We begin with the case of a $s \times s$ matrix $\mathbf{P}$ with a full von Neumann support.

Definition 5 *Given a preference matrix $\mathbf{P}$ of size $s \times s$ and an index $i \in \{1, \ldots, s\}$, we define the modified matrix $\mathbf{P}_i$ to be the $s \times s$ matrix (with $s > 1$) obtained by removing the i^{th} row of $\mathbf{P}$ and replacing it with $\mathbf{1}_s$, i.e. the s-dimensional row vector whose coordinates are all equal to 1.*

Lemma 6 *Suppose we are given the following for some parameter $1/3 > \alpha > 0$:*

1. *a preference matrix $\mathbf{P}$ of size $s \times s$ that has a unique von Neumann winner $\mathbf{w}$ with no vanishing coordinates, i.e. with full support (so s is necessarily odd)*
2. *$\epsilon \in (0, \min_i w_i]$*

3. *a preference matrix* $\mathbf{Q}$ *satisfying* $\|\mathbf{P}-\mathbf{Q}\| < \Delta_\alpha(\mathbf{P})$*, where* $\Delta_\alpha(\mathbf{P}) := \alpha\epsilon \max_i \sigma_s(\mathbf{P}_i)$ *and* $\sigma_j(\cdot)$ *denotes the* j^{th} *largest singular value of the matrix.*

Then, we can deduce that $\mathbf{Q}$ *also has a unique von Neumann winner* $\mathbf{v}$ *with no non-zero coordinates such that* $\|\mathbf{v}-\mathbf{w}\|_2 < 3\alpha\epsilon/2$*, and furthermore* $\Delta_\alpha(\mathbf{Q}) \geq \Delta_\alpha(\mathbf{P})/3$*.*

Proof Define $\mathbf{D} := \mathbf{Q}-\mathbf{P}$ and $\mathbf{D}_i := \mathbf{Q}_i - \mathbf{P}_i$: note that since $\mathbf{D}_i$ is obtained by replacing the i^{th} the row of $\mathbf{D}$ with zeros we get that $\mathbf{D}_i^\top \mathbf{D}_i \preceq \mathbf{D}^\top \mathbf{D}$ and in particular we have

$$\|\mathbf{D}_i\| \leq \|\mathbf{D}\| < \Delta_\alpha(\mathbf{P}) \tag{4}$$

Fix j as the maximizer of $\sigma_s(\mathbf{P}_j)$. We proceed to analyze $\mathbf{Q}_j$. We first observe that it is full rank as by Weyl's inequality we have

$$\begin{aligned}\sigma_s(\mathbf{Q}_j) &\geq \sigma_s(\mathbf{P}_j) - \|\mathbf{D}_j\| \geq \sigma_s(\mathbf{P}_j) - \Delta_\alpha(\mathbf{P}) \\ &\geq \sigma_s(\mathbf{P}_j) - \max_i \sigma_s(\mathbf{P}_i)/3 = 2\sigma_s(\mathbf{P}_j)/3 > 0\end{aligned} \tag{5}$$

where the third inequality is due to the fact that $\epsilon \leq 1, \alpha < 1/3$ and hence $\mathbf{Q}_j$ is full rank. It follows that the vector $\mathbf{v} := \mathbf{e}_j^\top \mathbf{Q}_j^{-1}$ is well defined. We proceed to show that $\mathbf{v}$ is the von Neumann winner of $\mathbf{Q}$. First, notice that $(\mathbf{v}^\top\mathbf{Q})_i = 0$ for all $i \neq j$ by the definition of $\mathbf{v}$. To prove that $(\mathbf{v}^\top\mathbf{Q})_j = 0$, we define z as the unit vector orthogonal to $\mathbf{Q}$, meaning $\mathbf{Q}z^\top = \mathbf{0}^\top$ (z is guaranteed to exist since $\mathbf{Q}$ is not a full rank matrix). Notice that if $z_j = 0$ then the columns other than j are linearly dependent and $\mathbf{Q}_i$ cannot be full rank, leading to a contradiction. It follow that $z_j \neq 0$ and

$$0 = \mathbf{v}\cdot\mathbf{0}^\top \overset{(i)}{=} \mathbf{v}^\top\mathbf{Q}z^\top = \sum_i z_i(\mathbf{v}^\top\mathbf{Q})_i \overset{(ii)}{=} z_j(\mathbf{v}^\top\mathbf{Q})_j$$

Equality (i) is since $\mathbf{Q}z^\top = \mathbf{0}^\top$. Equality (ii) holds since $(\mathbf{v}^\top\mathbf{Q})_i = 0$ for all $i \neq j$. We get that since $z_j \neq 0$, $(\mathbf{v}^\top\mathbf{Q})_j = 0$ meaning that $\mathbf{v}^\top\mathbf{Q} = \mathbf{0}$ as required.

In order to prove that $\mathbf{v}$ is the von Neumann winner of $\mathbf{Q}$ it now remains to show that $v_i \geq 0$ for all i. To do so, we bound its Euclidean distance from $\mathbf{w}$.

$$\begin{aligned}
\|\mathbf{v}-\mathbf{w}\| &= \|(\mathbf{Q}_j^{-1}-\mathbf{P}_j^{-1})\mathbf{e}_j\| \\
&= \|((\mathbf{P}_j+\mathbf{D}_j)^{-1}-\mathbf{P}_j^{-1})\mathbf{e}_j\| \\
&= \|\mathbf{P}_j^{-1}(I+\mathbf{D}_j\mathbf{P}_j^{-1})^{-1}\mathbf{D}_j\mathbf{P}_j^{-1}\mathbf{e}_j\| && \text{by the binomial inverse theorem} \\
&= \|\mathbf{P}_j^{-1}(I+\mathbf{D}_j\mathbf{P}_j^{-1})^{-1}\mathbf{D}_j\mathbf{w}\| \\
&\leq \|\mathbf{P}_j^{-1}(I+\mathbf{D}_j\mathbf{P}_j^{-1})^{-1}\mathbf{D}_j\| \\
&\leq \frac{\|\mathbf{D}_j\|}{\sigma_s(\mathbf{P}_j)}\|(I+\mathbf{D}_j\mathbf{P}_j^{-1})^{-1}\| \\
&\leq \frac{\|\mathbf{D}_j\|}{\sigma_s(\mathbf{P}_j)\sigma_s(I+\mathbf{D}_j\mathbf{P}_j^{-1})} \\
&\leq \frac{\|\mathbf{D}_j\|}{\sigma_s(\mathbf{P}_j)(1-\|\mathbf{D}_j\mathbf{P}_j^{-1}\|)} && \text{by Weyl's inequality} \\
&\leq \frac{\Delta_\alpha(\mathbf{P})}{\sigma_s(\mathbf{P}_j)(1-\Delta_\alpha(\mathbf{P})\|\mathbf{P}_j^{-1}\|)} && \text{by Equation (4)} \\
&= \frac{\frac{\Delta_\alpha(\mathbf{P})}{\sigma_s(\mathbf{P}_j)}}{1-\frac{\Delta_\alpha(\mathbf{P})}{\sigma_s(\mathbf{P}_j)}} \\
&\leq \frac{3\Delta_\alpha(\mathbf{P})}{2\sigma_s(\mathbf{P}_j)} && \text{since } \Delta_\alpha(\mathbf{P}) \leq \sigma_s(\mathbf{P}_j)/3 \text{ and so } 1-\frac{\Delta_\alpha(\mathbf{P})}{\sigma_s(\mathbf{P}_j)} \geq 2/3 \\
&< 3\alpha\epsilon/2.
\end{aligned}$$

Now, according to the definition of ϵ and the fact that $\alpha < 1/3$ we have for all i that

$$v_i \geq w_i - \|\mathbf{v}-\mathbf{w}\| \geq \epsilon/2 \tag{6}$$

and we conclude that $\mathbf{v}$ is indeed a von Neumann winner of the matrix $\mathbf{Q}$.

We proceed to analyze the uniqueness of $\mathbf{v}$. To this end we apply Lemma 4. Since $\mathbf{Q}_j$ is full rank and is obtained by replacing exactly one column of $\mathbf{Q}$ we get that $\mathrm{rank}(\mathbf{Q}) \geq \mathrm{rank}(\mathbf{Q}_j)-1 = s-1$ and $\mathbf{Q}$ is of rank $s-1$ as required. By Equation (6) we have that $v_i > 0$ for all i, hence Lemma 4 can indeed be applied and $\mathbf{Q}$ has a unique von Neumann winner. The lower bound for $\Delta_\alpha(\mathbf{Q})$ follows immediately from Equations (5) and (6). ■

We are now ready to deal with the more general case of a $K \times K$ matrix with a von Neumann winner with support size s. The following Lemma will later be used to show that after shifting the matrix $\mathbf{P}$, the columns of it for which $\mathbf{w}^\top\mathbf{P} > 0$, meaning those whose indices are not in the support of $\mathbf{w}$, remain to be outside the support after perturbing $\mathbf{P}$. That is, if the perturbed version of $\mathbf{P}$ is $\mathbf{Q}$ and its von Neumann winner is $\mathbf{v}$ then for these indices we still have $\mathbf{v}^\top\mathbf{Q} > 0$.

Lemma 7 *Let $\mathbf{p},\mathbf{q}$ be vectors with $\|\mathbf{p}-\mathbf{q}\| < \epsilon_1$, and let $\mathbf{w},\mathbf{v}$ be such that $\|\mathbf{w}-\mathbf{v}\| < \epsilon_2$. We have that*

$$|\langle \mathbf{p},\mathbf{w}\rangle - \langle \mathbf{q},\mathbf{v}\rangle| < \epsilon_2\|\mathbf{p}\| + \epsilon_1\|\mathbf{v}\|$$

Proof

$$|\langle \mathbf{p}, \mathbf{w} \rangle - \langle \mathbf{p}, \mathbf{v} \rangle| \leq \|\mathbf{p}\| \|\mathbf{w} - \mathbf{v}\| < \epsilon_2 \|\mathbf{p}\|$$
$$|\langle \mathbf{p}, \mathbf{v} \rangle - \langle \mathbf{q}, \mathbf{v} \rangle| \leq \|\mathbf{v}\| \|\mathbf{p} - \mathbf{q}\| < \epsilon_1 \|\mathbf{v}\|$$

The result follows from the triangle inequality ■

We are now ready to provide the final result of the section characterizing the proximity requirement of a pair of matrices for them to have the same von Neumann support.

Definition 8 *Suppose we are given a preference matrix* $\mathbf{P}$ *with a unique von Neumann winner* $\mathbf{w}$ *with support on the first* s *actions and the decomposition in Definition 3.2: in particular,* $\mathbf{P}^0$ *denotes the* $s \times s$ *minor that determines the non-zero weights of* $\mathbf{w}$*. We define the following quantities:*

1. $\epsilon_i := w_i$ *for* $i \leq s$
2. $\mathbf{p}_i := \mathbf{P}_{1:s,i}$ *using the notation of Definition 2*
3. $\epsilon_i := \langle \mathbf{w}, \mathbf{p}_i \rangle = (\mathbf{w}^\top \mathbf{P})_i$ *for* $i > s$
4. $\Delta(\mathbf{P}) := 2 \min \{\min_{i>s} \epsilon_i / \|\mathbf{p}_i\|, \min_{i \leq s} \epsilon_i\} \cdot \max_i \sigma_s(\mathbf{P}_i^0)/9$

In the following lemma, we adopt the notation where for a matrix $\mathbf{M}$, $\mathbf{m}_i := \mathbf{M}_{1:s,i}$.

Lemma 9 *Let* $\mathbf{P}$ *be a preference matrix with a unique von Neumann winner,* $\mathbf{w}$*, supported on the first* s *indices and consider the decomposition in Definition 3.2. Let* $\mathbf{Q}$ *be a* $K \times K$ *preference matrix satisfying*

1. $\|\mathbf{Q}_{1:s} - \mathbf{P}^0\| < \Delta(\mathbf{P})$*, using the notation of Definition 8,*
2. $\|\mathbf{p}_i - \mathbf{q}_i\| < \epsilon_i/3$ *for all* $i > S$*.*

Then, $\mathbf{Q}$ *has a unique von Neumann winner* $\mathbf{v}$*, supported on the first* s *actions. Furthermore, it holds that* $i > s$ *that* $\langle \mathbf{v}, \mathbf{q}_i \rangle > \epsilon_i/3$*, and that* $\Delta(\mathbf{Q})$ *defined as in Definition 8 is lower bounded by* $\Delta(\mathbf{Q}) \geq \Delta(\mathbf{P})/9$*.*

Proof Applying Lemmas 6 to $\mathbf{P}^0$ and $\mathbf{Q}_{1:s}$ and using the first bound above, we can conclude that $\mathbf{Q}_{1:s}$ has a unique von Neumann winner $\mathbf{v}^0$ with no vanishing coordinates and satisfying $\|\mathbf{w}^0 - \mathbf{v}^0\| < \epsilon/2$. For $i > s$, applying Lemma 7 to $\mathbf{w}^0$ and $\mathbf{p}_i$ and using the second bound above, we can conclude that

$$\langle \mathbf{q}_i, \mathbf{v} \rangle > \langle \mathbf{p}_i, \mathbf{w} \rangle - \Delta_i/3 - \epsilon \|\mathbf{p}_i\|/2 =$$
$$2\Delta_i/3 - \min \left\{ \min_{j>s} \Delta_j / \|\mathbf{p}_j\|, w_{\min} \right\} \|\mathbf{p}_i\|/3 \geq 2\Delta_i/3 - \Delta_i/3 = \Delta_i/3$$

It follows that $\mathbf{v}$ defined as $\mathbf{v}^0$ padded with zeros is the unique von Neumann winner of $\mathbf{Q}$. The claim regarding $\Delta(\mathbf{Q})$ follows via an elementary calculation. ■

The previous lemma can now be used for two separate claims. The first is the correctness of the algorithm. For a matrix $\mathbf{Q}$ and a confidence region $\mathcal{C}$ around it, if we conclude that any $\mathbf{P}$ inside $\mathcal{C}$

is close enough to $\mathbf{Q}$, in the terms stated above then we can conclude that all the matrices in $\mathcal{C}$ share the same support for their von Neumann winner. The second claim is for the convergence. Once our hypothesis matrix $\mathbf{Q}$ is close enough to the true matrix $\mathbf{P}$, it is guaranteed to have sufficiently large Δ and ϵ measures, meaning that an algorithm excluding actions based on the confidence region and the Δ and ϵ measures will be able to do so within a bounded amount of time.

5.4. Tying the Loss in a Game to the Distance From the von Neumann Winner

In this section we prove that a low regret policy must have its strategy converge to that of the von Neumann winner not only in terms of its regret, but only in terms of the strategy it plays. The rate of convergence relies on the problem specific parameters. We start by bounding the loss obtained by playing any vector that is orthogonal to the von Neumann winner, in the setting where the von Neumann winner has a full support.

Lemma 10 *Let $\mathbf{P}$ be a $s \times s$ skew-symmetric matrix with a unique von Neumann winner $\mathbf{w}$ with full support. Let σ be the $s-1$ singular value of the matrix $\mathbf{P}$. Let $\mathbf{x}$ be a Euclidean unit vector orthogonal to $\mathbf{w}$. Then it must be the case that $\min_i(\mathbf{x}^\top \mathbf{P})_i \leq -\frac{\sigma}{2\sqrt{s}}$*

Proof Let $\epsilon = -\min_i(\mathbf{x}^\top \mathbf{P})_i$. Set $\mathbf{r}$ as the vector taking a zero value in the indices where $(\mathbf{x}^\top \mathbf{P})_i \geq 0$ and $-(\mathbf{x}^\top \mathbf{P})_i$ elsewhere. We clearly have that $\|\mathbf{r}\| \leq \sqrt{s}\epsilon$ and $\|\mathbf{x}^\top \mathbf{P} + \mathbf{r}\|_1 \leq \|\mathbf{x}^\top \mathbf{P}\|_1$. Now, using the notation $\langle \cdot, \cdot \rangle$ for the dot product of two vectors, we have

$$
\begin{aligned}
-\left\langle \mathbf{x}^\top \mathbf{P}, \frac{\mathbf{x}^\top \mathbf{P} + \mathbf{r}}{\|\mathbf{x}^\top \mathbf{P} + \mathbf{r}\|_1} \right\rangle &= \frac{\|\mathbf{x}^\top \mathbf{P}\|^2 - \mathbf{x}^\top \mathbf{P}\mathbf{r}}{\|\mathbf{x}^\top \mathbf{P} + \mathbf{r}\|_1} \geq \frac{\|\mathbf{x}^\top \mathbf{P}\|^2 - \|\mathbf{x}^\top \mathbf{P}\|\sqrt{s}\epsilon}{\|\mathbf{x}^\top \mathbf{P}\|_1} \\
&\geq \frac{\|\mathbf{x}^\top \mathbf{P}\| - \sqrt{s}\epsilon}{\sqrt{s}} \geq \frac{\sigma - \sqrt{s}\epsilon}{\sqrt{s}}
\end{aligned}
$$

Assume for the sake of contradiction that we have $\epsilon < \frac{\sigma}{2\sqrt{s}}$. Notice that $\sum_i (\mathbf{x}^\top \mathbf{P})_i^2 = \|\mathbf{x}^\top \mathbf{P}\|^2 \geq \sigma^2$, hence our assumption on ϵ dictates that $\mathbf{x}^\top \mathbf{P}$ has entries with a positive value. It follows that $\mathbf{x}^\top \mathbf{P} + \mathbf{r}$ is non-zero, meaning that the inner product in the above equation is in fact a weighted averaging of the entries of $\mathbf{x}^\top \mathbf{P}$. We conclude that

$$
\epsilon = \max_i -(\mathbf{x}^\top \mathbf{P})_i \geq -\left\langle \mathbf{x}^\top \mathbf{P}, (\mathbf{x}^\top \mathbf{P} + \mathbf{r})/\|\mathbf{x}^\top \mathbf{P} + \mathbf{r}\|_1 \right\rangle \geq \frac{\sigma - \sqrt{s}\epsilon}{\sqrt{s}}
$$

Rearranging we get $\epsilon \geq \sigma/2\sqrt{s}$ contradicting our assumption. We conclude that it must be the case that $\epsilon \geq \sigma/2\sqrt{s}$. ■

We are now ready to lower bound the loss suffered by playing any vector that is orthogonal to the von Neumann winner in the setting where the von Neumann winner does not necessarily have a full support. We adopt the convention that for any matrix $\mathbf{M}$, we use the notation $\mathbf{m}_i$ to denote the transpose of the i^{th} column of $\mathbf{M}$. Moreover, for any K-dimensional (row) vector $\mathbf{x}$, we denote by $\mathbf{x}^0$ the s-dimensional vector consisting of the first s coordinates and $\mathbf{x}^1$ the $K-s$-dimensional vector consisting of the remaining coordinates. Furthermore, we use the notation

$$
\overline{\mathbf{x}}^0 := [\mathbf{x}^0\ \mathbf{0}] \text{ and } \overline{\mathbf{x}}^1 := [\mathbf{0}\ \mathbf{x}^1]. \tag{7}
$$

Note that we have $\mathbf{x} = \overline{\mathbf{x}}^0 + \overline{\mathbf{x}}^1$ and that for any p we have $\|\mathbf{x}^i\|_p = \|\overline{\mathbf{x}}^i\|_p$, where $\|\cdot\|_p$ denotes the L^p norm of a vector.

Lemma 11 *Let $\mathbf{P}$ be a $K \times K$ preference matrix with a unique von Neumann winner $\mathbf{w}$ with support on the first s actions and consider the decomposition in Definition 3.2: in particular, $\mathbf{P}^0$ is the $s \times s$ minor and we have $\mathbf{w} = [\mathbf{w}^0 \ \mathbf{w}^1]$ with $\mathbf{w}^0$ of length s. Let us also define $\epsilon := \min_{i \notin S}(\mathbf{w}^\top \mathbf{P})_i$. Denote by σ the smallest non-zero singular value of the matrix $\mathbf{P}^0$. Let $\mathbf{x}$ be a vector that is orthogonal to $\mathbf{w}$. We have that*

$$-\min_{i \le s} \mathbf{x}\mathbf{p}_i \ge \|\mathbf{x}\|_1 \frac{\sigma\epsilon}{2s}$$

Proof Using the notation of Equation (7), we have $\mathbf{x} = \overline{\mathbf{x}}^0 + \overline{\mathbf{x}}^1$ and $\|\mathbf{x}\|_1 = \|\overline{\mathbf{x}}^0\|_1 + \|\overline{\mathbf{x}}^1\|_1$, since the nonzero coordinates of $\overline{\mathbf{x}}^0$ and $\overline{\mathbf{x}}^1$ are non-overlapping. Also, we have the following two facts:

$$\overline{\mathbf{x}}^{0\top}\mathbf{P}\mathbf{w} = \mathbf{x}^{0\top}\begin{bmatrix}\mathbf{P}^0 & \mathbf{P}^{01}\end{bmatrix}\begin{bmatrix}\mathbf{w}^{0\top} \\ \mathbf{0}\end{bmatrix} = (-\mathbf{w}^{0\top}\mathbf{P}^0 + \mathbf{0})(\mathbf{x}^0)^\top = 0 \quad \text{as } \mathbf{P}^0 = -\mathbf{P}^{0\top} \text{ and } \mathbf{w}^{0\top}\mathbf{P}^0 = \mathbf{0}$$

$$\overline{\mathbf{x}}^{1\top}\mathbf{P}\mathbf{w} \le -\epsilon\|\overline{\mathbf{x}}^1\|_1 \qquad \text{by the definition of } \epsilon$$

Using this and the notation $\widehat{\imath} := \arg\min_{i \le s} \mathbf{x}\mathbf{p}_i^\top$, we have

$$\begin{aligned}
\min_{i \le s} \mathbf{x}\mathbf{p}_i^\top &= \sum_{i \le s} \left(\mathbf{x}\mathbf{p}_{\widehat{\imath}}^\top\right) w_i && \text{as } \mathbf{w} \text{ is supported on the first } s \text{ coordinates and so } \sum_{i \le s} w_i = 1 \\
&\le \sum_{i \le s} \left(\mathbf{x}\mathbf{p}_i^\top\right) w_i && \text{since } \mathbf{x}\mathbf{p}_{\widehat{\imath}}^\top \le \mathbf{x}\mathbf{p}_i^\top \text{ for all } i \\
&= \mathbf{x}^\top\mathbf{P}\mathbf{w} \\
&= \overline{\mathbf{x}}^{0\top}\mathbf{P}\mathbf{w} + \overline{\mathbf{x}}^{1\top}\mathbf{P}\mathbf{w} \\
&\le -\epsilon\|\overline{\mathbf{x}}^1\|_1
\end{aligned}$$

Multiplying everything in the above chain of inequalities by -1, we get

$$-\min_{i \le s} \mathbf{x}\mathbf{p}_i^\top \ge \epsilon\|\overline{\mathbf{x}}^1\|_1 \tag{8}$$

On the other hand, according to Lemma 10 we have that

$$-\min_{i \le s} \overline{\mathbf{x}}^0\mathbf{p}_i^\top \ge \frac{\|\overline{\mathbf{x}}^0\|_2\sigma}{2\sqrt{s}} \ge \frac{\|\overline{\mathbf{x}}^0\|_1\sigma}{2s}$$

and by the definition of $\mathbf{x} = \overline{\mathbf{x}}^0 + \overline{\mathbf{x}}^1$ we get

$$\begin{aligned}
-\min_{i\le s} \mathbf{x}\mathbf{p}_i^\top &= -\min_{i\le s} \left(\overline{\mathbf{x}}^0\mathbf{p}_i^\top + \overline{\mathbf{x}}^1\mathbf{p}_i^\top\right) \\
&= \max_{i\le s} -\left(\overline{\mathbf{x}}^0\mathbf{p}_i^\top + \overline{\mathbf{x}}^1\mathbf{p}_i^\top\right) \\
&\ge -\min_{i\le s} \overline{\mathbf{x}}^0\mathbf{p}_i^\top - \max_{i\le s} |\overline{\mathbf{x}}^1\mathbf{p}_i^\top| \\
&\ge -\min_{i\le s} \overline{\mathbf{x}}^0\mathbf{p}_i^\top - \|\overline{\mathbf{x}}^1\|_1 && \forall i,j,\ \ P_{ij}\in[-1,1] \\
&\ge \frac{\|\overline{\mathbf{x}}^0\|\sigma}{2\sqrt{s}} - \|\overline{\mathbf{x}}^1\|_1 && \text{Lemma 10} \\
&\ge \frac{\|\overline{\mathbf{x}}^0\|_1\sigma}{2s} - \|\overline{\mathbf{x}}^1\|_1 && \text{Cauchy-Schwarz inequality} \\
&\ge \frac{\sigma(\|\mathbf{x}\|_1 - \|\overline{\mathbf{x}}^1\|_1)}{2s} - \|\overline{\mathbf{x}}^1\|_1 && \|\mathbf{x}\|_1 = \|\overline{\mathbf{x}}^0\|_1 + \|\overline{\mathbf{x}}^1\|_1 \\
&= \frac{\sigma\|\mathbf{x}\|_1 - (\sigma+2s)\|\overline{\mathbf{x}}^1\|_1}{2s} && (9)
\end{aligned}$$

Equations (8) and (9) provide two lower bounds on $-\min_{i\le s} \mathbf{x}\mathbf{p}_i^\top$, both of them linear functions of $\|\overline{\mathbf{x}}^1\|_1$, one with negative slope and the other with positive slope. But, the upper envelope of two linear functions is bounded from below by the height of the point where the two lines intersect. So, setting the right-hand sides of Equations (8) and (9) and so solving for $\|\overline{\mathbf{x}}^1\|_1$, we get

$$\|\overline{\mathbf{x}}^1\|_1 = \frac{\sigma\|\mathbf{x}\|_1}{2s(1+\epsilon)-\sigma}$$

and since $\epsilon \le 1$,

$$-\min_{i\le s} \mathbf{x}\mathbf{p}_i^\top \ge \|\mathbf{x}\|_1 \frac{\sigma\epsilon}{2s(1+\epsilon)-\sigma} \ge \|\mathbf{x}\|_1 \frac{\sigma\epsilon}{4s}$$

as required. ■

We are now ready for the main result of this section stating that by playing a vector $\mathbf{v}$ that is far away from $\mathbf{w}$, the suffered loss must be large with linear proportion to the distance of $\mathbf{v}$ from $\mathbf{w}$.

Lemma 12 *Let $\mathbf{v}$ be a probability vector in $\mathbb{R}^K$. Let $\mathbf{P}$ be a skew-symmetric matrix with entires in $[-1,1]$ and a unique von Neumann winner $\mathbf{w}$. Let S be the support of $\mathbf{w}$ and let $\epsilon = \min_{i\notin S}(\mathbf{w}^\top\mathbf{P})_i$. Denote by σ the $(|S|-1)^{th}$ singular value of the matrix $\mathbf{P}_S$ (cf. Definition 2). Then it holds that*

$$\min_i (\mathbf{v}^\top\mathbf{P})_i \le -\frac{\sigma\epsilon\|\mathbf{w}-\mathbf{v}\|}{2\sqrt{K}|S|\|\mathbf{w}\|}$$

Proof Since $\mathbf{w}$ is a probability vector we have that $\left\langle \mathbf{w}, \frac{\mathbf{1}}{\sqrt{K}} \right\rangle = \frac{1}{\sqrt{K}}$ and in particular, $\left\| \mathbf{w} - \frac{\mathbf{1}}{|K|} \right\| = \sqrt{\|\mathbf{w}\|^2 - 1/|K|}$. It follows that for any vector $\mathbf{z} \perp \mathbf{1}$, we have

$$\begin{aligned}
\langle \mathbf{w}, \mathbf{z} \rangle &= \left\langle \mathbf{w} - \frac{\mathbf{1} \langle \mathbf{1}, \mathbf{w} \rangle}{|K|}, \mathbf{z} - \frac{\mathbf{1} \langle \mathbf{1}, \mathbf{z} \rangle}{|K|} \right\rangle + \frac{\langle \mathbf{1}, \mathbf{w} \rangle \langle \mathbf{1}, \mathbf{z} \rangle}{|K|} \\
&= \left\langle \mathbf{w} - \frac{\mathbf{1}}{|K|}, \mathbf{z} \right\rangle \\
&\leq \|\mathbf{z}\| \sqrt{\|\mathbf{w}\|^2 - \frac{1}{K}}
\end{aligned}$$

Now, we can write

$$\|\mathbf{w}\|^2 \|\mathbf{w} + \mathbf{z}\|^2 = \|\mathbf{w}\|^4 + 2 \langle \mathbf{w}, \mathbf{z} \rangle \|\mathbf{w}\|^2 + \|\mathbf{z}\|^2 \|\mathbf{w}\|^2$$

and

$$\langle \mathbf{w}, \mathbf{w} + \mathbf{z} \rangle^2 = \|\mathbf{w}\|^4 + 2 \langle \mathbf{w}, \mathbf{z} \rangle \|\mathbf{w}\|^2 + \langle \mathbf{w}, \mathbf{z} \rangle^2$$

meaning that

$$\|\mathbf{w}\|^2 \|\mathbf{w} + \mathbf{z}\|^2 - \langle \mathbf{w}, \mathbf{w} + \mathbf{z} \rangle^2 = \|\mathbf{z}\|^2 \|\mathbf{w}\|^2 - \langle \mathbf{w}, \mathbf{z} \rangle^2 \geq \frac{\|\mathbf{z}\|^2}{K}$$

Now, since both $\mathbf{w}, \mathbf{v}$ are probability vectors, we have that $\mathbf{z} = \mathbf{v} - \mathbf{w} \perp \mathbf{1}$ meaning that

$$\|\mathbf{w}\|^2 \|\mathbf{v}\|^2 (1 - \cos^2 \theta) = \|\mathbf{w}\|^2 \|\mathbf{v}\|^2 - \langle \mathbf{w}, \mathbf{v} \rangle^2 \geq \frac{\|\mathbf{w} - \mathbf{v}\|^2}{K}$$

with θ being the angle between the vectors $\mathbf{w}, \mathbf{v}$. We conclude that

$$\sin \theta \geq \frac{\|\mathbf{w} - \mathbf{v}\|}{\sqrt{K} \|\mathbf{w}\| \|\mathbf{v}\|}$$

This implies that if we write $\mathbf{v} = \alpha \mathbf{w} + \mathbf{u}$ with $\mathbf{u} \perp \mathbf{w}$, it must be the case that

$$\|\mathbf{u}\| = \|\mathbf{v}\| \sin \theta \geq \frac{\|\mathbf{w} - \mathbf{v}\|}{\sqrt{K} \|\mathbf{w}\|}$$

We now apply Lemma 11 to conclude that for some $i \in S$, with S being the support of $\mathbf{w}$,

$$-(\mathbf{v}^\top \mathbf{P})_i = -\alpha (\mathbf{w}^\top \mathbf{P})_i - (\mathbf{u}^\top \mathbf{P})_i = -(\mathbf{u}^\top \mathbf{P})_i \geq \frac{\sigma \epsilon \|\mathbf{u}\|_1}{2|S|} \geq \frac{\sigma \epsilon \|\mathbf{w} - \mathbf{v}\|}{2|S| \sqrt{K} \|\mathbf{w}\|}$$

■

5.5. Low Regret Algorithms

In this section we use the fact that low regret implies converging to the von Neumann winner to prove that when sparring two low-regret algorithms, after a finite (in term of T) amount of time, the matrix will be queried sufficiently many times in order to determine its von Neumann support.

In the following we fix $\mathbf{P}$ as a skew-symmetric $K \times K$ matrix with entries in $[-1, 1]$ with a unique von Neumann winner $\mathbf{w}$. W.l.o.g. we assume the support of $\mathbf{w}$ is in entries 1 through s. We analyze a two player game played for T steps where in each round the players choose $\mathbf{x}(t), \mathbf{y}(t) \in \Delta^K$ respectively. The gain for player #1 is $\mathbf{x}(t)^\top \mathbf{P}\mathbf{y}(t)$ and for player #2 is $\mathbf{y}(t)^\top \mathbf{P}^\top \mathbf{x}(t)$. We assume that both players are playing a strategy with a high probability regret guarantee stating that with probability at least $1 - \delta$,

$$\min_{\mathbf{z} \in \Delta^K} \sum_{t=1}^{T} (\mathbf{x}(t)^\top \mathbf{P}\mathbf{z}) \geq -c\sqrt{KT \log(K/\delta)}$$

and the respective guarantee for $\mathbf{y}$. We assume here that for both algorithms, for all time points t and indices i we have $\mathbf{x}(t)_i \geq 1/\sqrt{KT}$.

Lemma 13 *Let $\mathbf{x} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}(t)$. Let $\sigma = \sigma(\mathbf{P}), \epsilon = \epsilon(\mathbf{P})$ be as defined in Lemma 12. It holds for any $\rho > 0$ and*

$$T \geq \frac{4c^2 K^2 s^2 \|\mathbf{w}\|^2 \log(K/\delta)}{\sigma^2 \epsilon^2 \rho^2} = O\left(\frac{K^2 s^2 \log(K/\delta)}{\sigma^2 \epsilon^2 \rho^2}\right)$$

that $\|\mathbf{x} - \mathbf{w}\|_\infty < \rho$ with probability at least $1 - \delta$.

Proof By Lemma 12 we have that w.p. at least $1 - \delta$,

$$\frac{\sigma\epsilon\|\mathbf{x} - \mathbf{w}\|}{2\sqrt{K}|S|\|\mathbf{w}\|} \leq c\sqrt{KT \log(1/\delta)}/T$$

Since $\|a\|_\infty \leq \|a\|_2$ for any vector the claim follows. ■

Lemma 14 *Let i be an action in the support of the von Neumann ($\mathbf{w}$) winner of $\mathbf{P}$, and let $j \in [K]$. There exist some*

$$T_0 = O\left(\frac{K^2 s^2 \log(K/\delta)}{\sigma^2 \epsilon^2 w_i^2} + \frac{K \log(1/\delta)^2}{w_i^4}\right)$$

Such that if $T \geq T_0$ then with probability at least $1 - O(\delta)$ the pair i, j is queried at least $\sqrt{T} w_i^2/10$ many times.

Proof We prove the lemma under the event in which $\|\mathbf{x} - \mathbf{w}\|_\infty < w_i/4$. By Lemma 13, this event occurs w.p. at least $1 - \delta$ given that $T \geq T_0$ for some

$$T_0 = O\left(\frac{K^2 s^2 \log(K/\delta)}{\sigma^2 \epsilon^2 w_i^2}\right)$$

Given the bound on $\|\mathbf{w}-\mathbf{x}\|$ we would like to bound the number of times in which $\mathbf{x}(t)_i$ is too small. To this end we consider the random variable obtained by choosing t uniformly from $[T]$ and considering the quantity $1-\mathbf{x}(t)_i$. This random variable is non-negative and its expected value is at most $1-3w_i/4$. We apply Markov's inequality and obtain that for at most $1/(1+w_i/4) \leq 1-w_i/5$ fraction of the time it may hold that

$$1-\mathbf{x}(t)_i \geq (1-3w_i/4)(1+w_i/4) \geq 1-w_i/2$$

In other words, there are at least $Tw_i/5$ many values of $t \in [T]$ in which $\mathbf{x}(t)_i \geq w_i/2$. According to the property mentioned above we have that for all such t values, $\mathbf{y}(t)_j \geq \frac{1}{\sqrt{KT}}$. It follows that

$$\sum_{t=1}^{T} \mathbf{x}(t)_i \mathbf{y}(t)_j \geq \frac{w_i^2\sqrt{T}}{10\sqrt{K}}$$

Consider now the random variable $Z(t)$ taking the value of 1 at round t if player #1's random process of choosing a single action based on $\mathbf{x}(t)$ chose action i and player #2 chose action j. We have that all such $Z(t)$'s are Bernoulli random variables with

$$\sum_{t} \mathbb{E}[Z(t)] \geq \frac{w_i^2\sqrt{T}}{10\sqrt{K}}$$

The following auxiliary lemma provides a concentration bound, via standard techniques, that allows to lower bound the sum $\sum_{t=1}^{T} Z(t)$

Lemma 15 (Bernstein's inequality) *Let $Z(1),\ldots,Z(T)$ be a sequence of Bernoulli random variables. Let $\delta > 0$ and assume that $\sum_{t=1}^{T} \mathbb{E}[Z(t)] \geq 2\log(1/\delta)$. Then*

$$\Pr\left[\sum_{t=1}^{T} Z(t) < \sum_{t=1}^{T} \mathbb{E}[Z(t)] - \sqrt{3\sum_{t} \mathbb{E}[Z(t)]\log(1/\delta)}\right] < \delta$$

In particular, if $\sum_{t=1}^{T} \mathbb{E}[Z(t)] \geq 12\log(1/\delta)$ then

$$\Pr\left[\sum_{t=1}^{T} Z(t) < \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}[Z(t)]\right] < \delta$$

Proof By Bernstein's inequality applied on the independent sequence of $\mathbb{E}[Z(t)] - Z(t)$, for any positive α it holds

$$Pr\left[\sum_{t=1}^{T}(\mathbb{E}[Z(t)] - Z(t)) > \alpha\right] < \exp\left(-\frac{\alpha^2}{2\sum_{t=1}^{T}\mathbb{E}[(Z(t)-\mathbb{E}[Z(t)])^2] + 2\alpha/3}\right) \leq$$

$$\exp\left(-\frac{\alpha^2}{2\sum_{t=1}^{T}\mathbb{E}[Z(t)] + 2\alpha/3}\right)$$

The second inequality is since $Z(t)$ is a Bernouli random variable and its variance is smaller than its expected value. Set $\alpha = \sqrt{3\sum_{t=1}^{T}\mathbb{E}[Z(t)]\log(1/\delta)}$. Since $\sum_{t=1}^{T}\mathbb{E}[Z(t)] \geq 2\log(1/\delta)$ we have

$$\frac{2\alpha/3}{\sum_t \mathbb{E}[Z(t)]} = \frac{2}{3}\cdot\sqrt{\frac{3\log(1/\delta)}{\sum_t \mathbb{E}[Z(t)]}} \leq \frac{2}{3}\cdot\sqrt{\frac{3\log(1/\delta)}{2\log(1/\delta)}} < 1$$

Hence,

$$\Pr\left[\sum_{t=1}^{T}(\mathbb{E}[Z(t)] - Z(t)) > \alpha\right] < \exp\left(-\frac{\alpha^2}{3\sum_{t=1}^{T}\mathbb{E}[Z(t)]}\right) = \delta$$

■

We continue with the proof of Lemma 14. By the above, since $\frac{w_i^2\sqrt{T}}{10\sqrt{K}} \geq 12\log(1/\delta)$, with large enough constants in the $O()$ term, we get that w.p. at least $1-\delta$ it holds that

$$\sum_t Z(t) \geq \frac{w_i^2\sqrt{T}}{20\sqrt{K}}$$

Since the analog can be said for the probability of player #1 choosing j and player #2 choosing i the claim follows. ■

Theorem 16 *Let* $\mathbf{P}$ *be a preference matrix with von Neumann winner* $\mathbf{w}$ *with support on the first* s *actions and denote by* $w_{\min}$ *the smallest non-zero coordinate of* $\mathbf{w}$*. Let* n *be some natural number, and let* $\delta > 0$*. For* $T \geq T_0$ *with*

$$T_0 = O\left(\frac{K^2 s^2\log(K/\delta)}{\sigma^2\epsilon^2 w_{\min}^2} + \frac{Kn^2 + \log(K/\delta)^2}{w_{\min}^4}\right)$$

we have that w.p. at least $1-\delta$*, every point* i, j *with* $i \leq s$ *being in the support of* $\mathbf{w}$ *is queried at least* n *times.*

Proof By Lemma 14 we get that for any fixed i, j w.p. at least $1-\delta/Ks$ that the point i, j is queried at least

$$\frac{w_i^2\sqrt{T}}{10\sqrt{K}} \geq n$$

times. The claim follows via union bound. ■

Corollary 17 *For* $\delta > 0$*,*

$$T_0 = \tilde{O}\left(\frac{K^2 s^2\log(K/\delta)}{\sigma^2\epsilon^2 w_{\min}^2} + \frac{\log(K/\delta)^2}{w_{\min}^4} + \frac{Ks^2\log(K/\delta)^4}{\epsilon^4 w_{\min}^4} + \frac{Ks^4\log(K/\delta)^4}{\Delta(P)^4 w_{\min}^4}\right)$$

and $T \geq T_0$ *we have that w.p. at least* $1-\delta$ *starting from time* T_0 *and onwards any invocation of Subroutine 2 will result in the exclusion of all actions outside of* $[s]$*.*

Proof In what follows we prove the result under the occurrence of (1) the event that the confidence region of the algorithm contain $\mathbf{P}$ and (2) the event leading to the correctness of Theorem 16. These event occur together by union bound w.p. at least $1-\delta$, with large enough constants in the above $O(\cdot)$ term.

According to Theorem 16 we have that by time T_0, any pair i,j with $i \in [s]$ is queried at least

$$O\left(\frac{s\log(KT_0/\delta)^2}{\epsilon^2} + \frac{s^2\log(KT_0/\delta)^2}{\Delta(P)^2}\right)$$

many times. It follows that that the confidence interval around it is of length at most $c\min\{\Delta(\mathbf{P})/s, \epsilon/\sqrt{s}\}$ for some sufficiently small constant c. Also, by the definition of the confidence intervals we know that the true value of P_{ij} is contained in it. We conclude that at time T_0, w.p. at least $1-\delta$, the empirical estimate $\mathbf{Q}$ of $\mathbf{P}$ and the confidence region around it $\mathcal{C}$ have the following properties

1. $\mathbf{P} \in \mathcal{C}$
2. $\|\mathbf{Q}_{1:s} - \mathbf{P}_{1:s}\| \leq \|\mathbf{Q}_{1:s} - \mathbf{P}_{1:s}\|_F < s \cdot c\min\{\Delta(\mathbf{P})/s, \epsilon/\sqrt{s}\} < c\Delta(\mathbf{P})$, using the notation of Definition 8
3. $\forall i > s,\ \|\mathbf{q}_i - \mathbf{p}_i\| < \sqrt{s}\cdot c\min\{\Delta(\mathbf{P})/s, \epsilon/\sqrt{s}\} < c\epsilon_i(\mathbf{P})$, using the notation of Definition 8

Hence, we may applying Lemma 9 to conclude that additionally, assuming c is sufficiently small,

1. $\mathbf{Q}$ has a unique von Neumann winner with the same support as that of $\mathbf{P}$ (assumed w.l.o.g. to be $[s]$)
2. $\forall \mathbf{P}' \in \mathcal{C}$, $\|\mathbf{Q}_{1:s} - \mathbf{P}'_{1:s}\| < \Delta(\mathbf{Q})$
3. $\forall \mathbf{P}' \in \mathcal{C}$, $\|\mathbf{q}_i - \mathbf{p}'_i\| < \epsilon_i(\mathbf{Q})$

It is an easy exercise to show that given this, Subroutine 2 will indeed determine that $[s]$ is the support of the von Neumann winner of all matrices $\mathbf{P}' \in \mathcal{C}$, as required. ■

Corollary 18 *There exists some* $C(P) = \tilde{O}\left(Ks^2\min\{\min_{i\leq s} w_i, \min_{i>s}(\mathbf{w}^\top\mathbf{P})_i, \sigma(\mathbf{P}_S), \Delta(\mathbf{P})\}^{-4}\right)$ *such that the regret obtained by sparring two copies of Exp3.P with success probability of* $1-\delta$ *is at most*

$$O\left(\min\left\{C(P)\log(1/\delta)^2 + \sqrt{sT\log(s/\delta)}, \sqrt{KT\log(K/\delta)}\right\}\right)$$

6. Conclusion

In this work, we have provided the first instance-dependent regret bound for the multi-armed dueling bandit problem in the absence of a Condorcet winner. The bound takes the form $\widetilde{O}(\sqrt{sT})$, where s is a sparsity parameter that was shown to be small in practice, even when the number of actions is large. This improves upon the worst-case bound of $\widetilde{O}(\sqrt{KT})$, where K is the number of actions. Moreover, the result holds for almost all dueling bandit problem. This result is obtained by using a mixture of multiplicative weights methods and those using confidence bounds for action elimination.

A natural direction for future work is applying the ideas put forth in this work to the contextual dueling bandit problem, which could lead to the first online contextual dueling bandit algorithm. Another avenue for further research is the lifting of the remaining restrictions placed on the multi-armed dueling bandit problem.

Acknowledgments

We would like to thank Alekh Agarwal, Miroslav Dudík, and Akshay Krishnamurthy for many formative discussions on the problem.

References

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

R. Busa-Fekete, B. Szörényi, P. Weng, W. Cheng, and E. Hüllermeier. Top-k selection based on adaptive sampling of noisy preferences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

R. Busa-Fekete, B. Szörényi, and E. Hüllermeier. PAC rank elicitation through adaptive sampling of stochastic pairwise preferences. In *National Conference on Artificial Intelligence (AAAI)*, 2014.

Yuxin Chen and Changho Suh. Spectral mle: Top-k rank aggregation from pairwise comparisons. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 371–380, 2015.

Massimiliano Ciaramita, Vanessa Murdock, and Vassilis Plachouras. Online learning from click data for sponsored search. In *World Wide Web*. ACM, 2008.

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

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Conference on Learning Theory (COLT)*. Springer, 2002.

Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Information Retrieval*, 16(1):63–90, 2013.

Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse dueling bandits. In *Artificial Intelligence and Statistics*, pages 416–424, 2015.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

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

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

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *NIPS*, 2012.

Guillermo Owen. *Game Theory*. Emerald Group Publishing Limited, 3rd edition, 1995.

Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for taxonomies: A model-based approach. In *SIAM Conference on Data Mining (SDM)*. SIAM, 2007.

Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*. ACM, 2009.

Ronald L. Rivest and Emily Shen. An optimal single-winner preferential voting system based on game theory. In *Proceedings Third International Workshop on Computational Social Choice*, 2010.

Tanguy Urvoy, Fabrice Clerot, Raphael Féraud, and Sami Naamane. Generic exploration and k-armed voting bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 91–99, 2013.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

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, September 2012.

Masrour Zoghi, Shimon Whiteson, Maarten de Rijke, and Rémi Munos. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, 2014a.

Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten de Rijke. Relative upper confidence bound for the k-armed dueling bandits problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014b.

Masrour Zoghi, Zohar Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland dueling bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015a.

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