Testification of Condorcet Winners in Dueling Bandits

Björn Haddenhorst[1] **Viktor Bengs**[1] **Jasmin Brandt**[1] **Eyke Hüllermeier**[2]

[1]Department of Computer Science, Paderborn University, Paderborn, Germany
[2]Institute of Informatics, University of Munich (LMU), Munich, Germany

Abstract

Several algorithms for finding the best arm in the dueling bandits setting assume the existence of a Condorcet winner (CW), that is, an arm that uniformly dominates all other arms. Yet, by simply relying on this assumption but not verifying it, such algorithms may produce doubtful results in cases where it actually fails to hold. Even worse, the problem may not be noticed, and an alleged CW still be produced. In this paper, we therefore address the problem as a "testification" task, by which we mean a combination of testing and identification: The online identification of the CW is combined with the statistical testing of the CW assumption. Thus, instead of returning a supposed CW at some point, the learner has the possibility to stop sampling and refuse an answer in case it feels confident that the CW assumption is violated. Analyzing the testification problem formally, we derive lower bounds on the expected sample complexity of any online algorithm solving it. Moreover, a concrete algorithm is proposed, which achieves the optimal sample complexity up to logarithmic terms.

1 INTRODUCTION

The standard multi-armed bandit (MAB) problem is a sequential decision problem with a finite set of choice alternatives (arms), in which arms are selected and corresponding (noisy) numerical rewards observed in a sequential manner. A practically motivated modification of the MAB problem is the dueling bandits [Sui et al., 2018] or preference-based multi-armed bandit problem [Bengs et al., 2021]. Here, instead of repeatedly pulling a single arm at a time and observing a numerical reward, the learner pulls two arms and observes the winner of the corresponding comparison (duel).

Thus, although both settings proceed from the assumption of stochastic feedback, the latter differs from the former with respect to the possible actions (pulling pairs vs. pulling single arms) as well as the provided feedback (qualitative vs. quantitative). Regardless of these differences, the goal of a learner is typically to find the "best" arm as quickly as possible. Yet, due to the absence of numerical rewards, the definition of a best arm is no longer straightforward in the dueling bandits setting. A natural and quite appealing definition of a best arm refers to the established notion of a *Condorcet Winner* (CW): An arm is a best arm if it outperforms all other arms in a pairwise comparison, in the sense of being more likely to win than to lose.

Compared to other assumptions made in the realm of dueling bandits, the existence of a Condorcet winner is a rather mild condition. Still, it guarantees the learning task to be well-defined, and allows for deriving performance guarantees of a learner [Bengs et al., 2021]. For all these reasons, the existence of a CW is a very desirable property. Nevertheless, in many practical applications, even the assumption of a CW cannot be assured. Learning algorithms relying on this assumption may then perform poorly, e.g., can reveal linear growth of regret rates. In the recent past, other notions of "best arm" (or winner) and corresponding learning algorithms have been studied, all coming with their own limitations, but still offering a reasonable alternative in the absence of a CW [Urvoy et al., 2013, Zoghi et al., 2015, Jamieson et al., 2015, Komiyama et al., 2016, Wu and Liu, 2016].

As a consequence, it would be desirable if a learner seeking to identify the (alleged) CW would be able to verify the validity of the CW assumption. By conducting a sanity check of its underlying assumptions, such a learner could make sure that its overall target is actually well defined, and if not, either terminate a presumably meaningless learning process or switch to an alternative criterion for identifying a best arm.

Inspired by these considerations, we introduce the "testification" (testing + identification) problem for the Condorcet

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

winner, that is, the learning task that combines the identification of the CW with the simultaneous statistical testing of the validity of the CW assumption. To be more precise, the learner's goal is to either identify the CW if it exists, or to stop the learning process in case it does not, both as quickly as possible while maintaining predefined error probabilities. Note that any testification algorithm can be used for finding the CW in a dueling bandits problem, but in contrast to existing algorithmic solutions for that problem, our algorithm is able to detect a violation of the CW assumption.

In this paper, we consider the dueling bandits framework under the low noise assumption (reviewed in Section 3). For this framework, we introduce the testification problem and prove an instance-wise lower bound on the expected sample complexity for any algorithm capable of solving this problem (Section 4). We show that the testification problem can be solved in a straightforward way by two separated identification and testing phases connected in series, each phase instantiated by an appropriate identification or testing procedure (Section 5.1). Although it can be proved that this straightforward approach already achieves almost asymptotically optimal worst-case expected sample complexity, provided that appropriate algorithms are used in each of the two phases (Section C), it is arguably more reasonable to interleave both testing and identification in the learning process. Indeed, the strict separation of testing and identification might result in a seemingly superfluous verification at the end of the learning process in cases where a CW does not exist. Guided by this consideration we suggest a more sophisticated testification algorithm, called *Noisy Tournament Sampling* (NTS), where testing and identifying goes hand in hand (Section 5.3). For its design we exploit a connection between tournaments in graph theory and the CW of a (binary) preference relation (Section 5.2), in order to emulate the mechanism of a deterministic CW tester, resulting in a learner that is nearly optimal for the testification problem with regard to its expected sample complexity if instantiated with a suitable CW tester (Section 5.5). As a side result, we show how NTS can be used in combination with any learning algorithm based on a CW assumption to passively monitor the statistical validity of the CW assumption (Section 5.4). Finally, we discuss the impact of our findings on other problems related to testification (Section 6), and demonstrate the superiority of NTS over the straightforward approach with separated identification and testing phases in an experimental study (Sections 7 and I).

All proofs of theoretical results are deferred to the supplementary material, which also contains a list of symbols used throughout the paper.

2 RELATED WORK

The existence of a CW in the dueling bandits problem is required in a variety of papers, either explicitly [Urvoy et al., 2013, Zoghi et al., 2014, Komiyama et al., 2015, Karnin, 2016, Chen and Frazier, 2017, Li et al., 2020] or implicitly as a consequence of stronger assumptions on the underlying preference relation, such as a total order of arms or some kind of stochastic transitivity [Yue and Joachims, 2011, Yue et al., 2012, Falahatgar et al., 2017b,a, 2018, Mohajer et al., 2017], by assuming latent utilities [Yue and Joachims, 2009, Ailon et al., 2014, Kumagai, 2017, Maystre and Grossglauser, 2017] or an underlying statistical ranking model [Busa-Fekete et al., 2014, Szörényi et al., 2015].

Yet, in real-life scenarios, the existence of a CW can not always be guaranteed, as specifically noted by Zoghi et al. [2015]. This observation initiated research on alternative best arm concepts that always exist, such as the Copeland winner [Zoghi et al., 2015, Komiyama et al., 2016, Wu and Liu, 2016], the Borda winner [Jamieson et al., 2015], or more general tournament solutions [Ramamohan et al., 2016]. However, it is important to note that the arguments put forward by Zoghi et al. [2015] are purely empirical, and the conclusion that a CW does not exists in the considered applications are only derived in hindsight, after having seen all the data. The authors do not provide a statistical framework to verify or reject the CW assumption, neither in an offline nor in an online manner.

Degenne and Koolen [2019] consider the pure exploration bandit problem with multiple correct answers in a quite general setting, which also covers the CW testification problem. From their results one can obtain instance-wise optimal lower and upper bounds on the asymptotics of CW testification algorithms, which we elaborate on in Section 3. Unfortunately, these bounds do not provide any information of the sample complexity of solving the CW testification task with a predefined level of confidence, which is probably the most common use case in reality.

Apart from this, the CW testification problem has merely been addressed in the deterministic scenario, in which the outcome of a duel between two arms, if queried repeatedly, is always the same. The question of the minimal number of queries necessary to decide the (non-)existence of a CW within a tournament, which has a natural connection to a CW in a strict binary preference relation, is investigated by Bollobás and Eldridge [1978], Balasubramanian et al. [1997] and Procaccia [2008].

3 THE DUELING BANDITS PROBLEM

Consider a finite set of m arms identified by the index set $[m] \coloneqq \{1, \dots, m\}$. In the setting of the dueling bandits (preference-based multi-armed bandit) problem, two distinct arms $i, j \in [m]$ can be compared with each other at each time step $t \in \mathbb{N}$. Querying a pairwise preference, the learner is provided with binary feedback about the winner of the duel, which is assumed to be generated by a time-stationary

i.i.d. probabilistic process. The probability $\mathbb{P}(i \succ j)$ that arm i wins against arm j is given by some underlying (unknown) ground truth parameter $q_{i,j} \in [0,1]$. Excluding ties and setting (w.l.o.g.) $q_{i,i} = \frac{1}{2}$ for every $i \in [m]$, we can infer that $\mathbf{Q} = (q_{i,j})_{1 \leq i,j \leq m}$ is a reciprocal relation on $[m]$, i.e., $\mathbf{Q}$ is an element of the set of all *preference relations* formally defined by

$$\mathcal{Q}_m := \Big\{ \mathbf{Q} = (q_{i,j})_{1 \leq i,j \leq m} \in [0,1]^{m \times m} \mid q_{j,i} = 1 - q_{i,j} \text{ for every } i, j \in [m] \Big\}.$$

To assimilate the information available at time $t \in \mathbb{N}$, let us write $(\boldsymbol{n}_t)_{i,j}$ for the number of comparisons between i and j until time t, and $(\boldsymbol{w}_t)_{i,j}$ for the number of times i has won against j until time t. This obviously implies $(\boldsymbol{w}_t)_{i,j} + (\boldsymbol{w}_t)_{j,i} = (\boldsymbol{n}_t)_{i,j} = (\boldsymbol{n}_t)_{j,i}$. Let us write $[m]_2$ for the set containing all subsets of size 2 of $[m]$. Furthermore, if $\boldsymbol{w} \in \mathbb{N}_0^{m \times m}$ and $\boldsymbol{n} \in \mathbb{N}^{m \times m}$, we denote the matrix $(\frac{w_{i,j}}{n_{i,j}})_{1 \leq i,j \leq m} \in \mathbb{R}^{m \times m}$ by $\frac{\boldsymbol{w}}{\boldsymbol{n}}$. For convenience, we introduce the notations $(m)_2 := \{(i,j) \in [m] \times [m] : i < j\}$ as well as $\langle m \rangle_2 := \{(i,j) \in [m] \times [m] : i \neq j\}$. A specific learning algorithm in the realm of dueling bandits can be identified by a sampling strategy as defined in the following.

Definition 3.1. *A **sampling strategy** π is a family of random variables, which, depending on the time t and the observations $\boldsymbol{n}_0, \boldsymbol{w}_0, \ldots, \boldsymbol{n}_{t-1}, \boldsymbol{w}_{t-1}$ available before time t, determines a pair $(i(t), j(t)) \in \langle m \rangle_2$ to be compared at time $t \in \mathbb{N}$.*

Definition 3.2. *An arm $i \in [m]$ is a **Condorcet winner (CW) of** $\mathbf{Q} \in \mathcal{Q}_m$ (denoted by $\mathrm{CW}(\mathbf{Q})$) if $q_{i,j} > \frac{1}{2}$ for every $j \in [m] \setminus \{i\}$. The* CW preference relations *and* Non-CW preference relations *are*

$$\mathcal{Q}_m(\mathrm{CW}) := \{\mathbf{Q} \in \mathcal{Q}_m \mid \exists i \in [m] : i \text{ is a CW of } \mathbf{Q}\},$$
$$\mathcal{Q}_m(\neg\mathrm{CW}) := \mathcal{Q}_m \setminus \mathcal{Q}_m(\mathrm{CW}).$$

In this paper, we consider relations $\mathbf{Q}$ in

$$\mathcal{Q}_m^h := \big\{ \mathbf{Q} = (q_{i,j})_{1 \leq i,j \leq m} \in \mathcal{Q}_m \mid |q_{i,j} - 1/2| > h \text{ for every distinct } i, j \in [m] \big\},$$

for $h \in (0, 1/2)$. Preference relations in $\mathcal{Q}_m^h$ are said to satisfy the *low noise assumption* [Braverman et al., 2016, Korba et al., 2017]. To some extent, the parameter h determines the difficulty of the underlying dueling bandits problem, in the sense that an h near $1/2$ implies rather clear outcomes in the pairwise comparisons and consequently a rather easy problem scenario, while a small h implies that winning probabilities may be close to $1/2$, and hence difficult to distinguish. For the sake of convenience, we write $\mathcal{Q}_m(i)$ for the set of all $\mathbf{Q} \in \mathcal{Q}_m(\mathrm{CW})$ with $\mathrm{CW}(\mathbf{Q}) = i$ and define $\mathcal{Q}_m^h(\mathrm{X}) := \mathcal{Q}_m^h \cap \mathcal{Q}_m(\mathrm{X})$ for any $\mathrm{X} \in \{\mathrm{CW}, \neg\mathrm{CW}, 1, \ldots, m\}$.

4 THE CW TESTIFICATION PROBLEM

A first informal statement of the *testification problem for the CW* of an underlying (unknown) preference relation $\mathbf{Q}$ can be given as follows:

> Is $\mathbf{Q}$ in $\mathcal{Q}_m(\mathrm{CW})$? If so, determine the CW and return it, otherwise return $\neg\mathrm{CW}$.

In the course of the paper, we focus on algorithms $\mathcal{A}$ for the testification problem, which might be probabilistic and interact with the underlying dueling bandits environment, as stipulated by the definition of a sampling strategy π (Definition 3.1). In case an algorithm $\mathcal{A}$ terminates, it returns a decision denoted by $\mathbf{D}(\mathcal{A})$, which can be either

- an element $i^* \in [m]$, i.e., $\mathbf{D}(\mathcal{A}) = i^*$, which indicates that $\mathcal{A}$ predicts i^* to be the CW,
- or $\neg\mathrm{CW}$, i.e., $\mathbf{D}(\mathcal{A}) = \neg\mathrm{CW}$, which indicates that $\mathcal{A}$ predicts that no CW exists.

Moreover, we denote by $T^{\mathcal{A}}$ the sample complexity of an algorithm $\mathcal{A}$, i.e., the number of pairwise comparisons $\mathcal{A}$ has made before termination.

For given error probabilities $\alpha, \beta \in (0,1)$, we say that an algorithm $\mathcal{A}$ *solves the testification problem for the CW on $\mathcal{Q}_m^h$ for α and β* (short: $\mathcal{A}$ solves $\mathcal{P}^{m,h,\alpha,\beta}$) if $T^{\mathcal{A}}$ is almost surely finite and the following holds:

$$\begin{aligned} \inf_{i^* \in [m]} \inf_{\mathbf{Q} \in \mathcal{Q}_m^h(i^*)} \mathbb{P}_{\mathbf{Q}}\big(\mathbf{D}(\mathcal{A}) = i^*\big) &\geq 1 - \alpha, \\ \inf_{\mathbf{Q} \in \mathcal{Q}_m^h(\neg\mathrm{CW})} \mathbb{P}_{\mathbf{Q}}\big(\mathbf{D}(\mathcal{A}) = \neg\,\mathrm{CW}\big) &\geq 1 - \beta. \end{aligned} \tag{1}$$

The primary interest lies in constructing algorithms $\mathcal{A}$ capable of solving the testification problem for the CW on $\mathcal{Q}_m^h$ with an expected sample complexity $T^{\mathcal{A}}$ as small as possible. Obviously, the latter will strongly depend on the predefined error bounds α, β, the number of available arms m, as well as on the parameter h of the class of preference relations $\mathcal{Q}_m^h$ satisfying the low noise assumption.

In Section J in the supplement, we reduce the CW testification problem to the pure exploration bandit problem with multiple correct answers as defined in [Degenne and Koolen, 2019]. This approach leads to the following results: If $\mathcal{A}(\gamma)$ solves $\mathcal{P}^{m,h,\gamma,\gamma}$, then

$$\liminf_{\gamma \to 0} \frac{\mathbb{E}_{\mathbf{Q}}[T^{\mathcal{A}(\gamma)}]}{\ln(\gamma^{-1})} \geq \frac{1}{D_m^h(\mathbf{Q})} \tag{2}$$

for some known constant $D_m^h(\mathbf{Q}) > 0$, and there exists a solution $\mathcal{A}(\gamma)$ to $\mathcal{P}^{m,h,\gamma,\gamma}$ with

$$\lim_{\gamma \to 0} \frac{\mathbb{E}_{\mathbf{Q}}[T^{\mathcal{A}(\gamma)}]}{\ln(\gamma^{-1})} \leq \frac{1}{D_m^h(\mathbf{Q})}. \tag{3}$$

In case $\mathbf{Q} \in \mathcal{Q}_m(\mathrm{X})$ (for $\mathrm{X} \in \{\neg\mathrm{CW}, 1, \ldots, m\}$), the complexity term $D_m^h(\mathbf{Q})$ is given as

$$\sup_{\mathbf{v} \in \Delta_{(m)_2}} \inf_{\mathbf{Q}' \in \mathcal{Q}_m^h(\neg\mathrm{X})} \sum\nolimits_{(i,j) \in (m)_2} v_{i,j} d_{\mathrm{KL}}(q_{i,j}, q'_{i,j}),$$

where $\Delta_{(m)_2}$ is the set of all $\mathbf{v} = (v_{i,j})_{1\leq i<j\leq m}$ with $\min_{i<j} v_{i,j} \geq 0$ and $\sum_{i<j} v_{i,j} = 1$ and $d_{\mathrm{KL}}(p,q) = p\ln(p/q)+(1-p)\ln((1-p)/(1-q))$ is the KL-divergence between two independent random variables $X \sim \mathrm{Ber}(p)$ and $Y \sim \mathrm{Ber}(q)$. We prove in the supplement (cf. Lemmata J.5 and J.6) that

$$\frac{(m-1)(1/4-h^2)}{4h^2} \leq \sup_{\mathbf{Q}\in\mathcal{Q}_m^h} \frac{1}{D_m^h(\mathbf{Q})} \leq \frac{m}{8h^2} \quad (4)$$

holds, hence any *optimal* solution $\mathcal{A}(\gamma)$ to $\mathcal{P}^{m,h,\gamma,\gamma}$ fulfills

$$\sup_{\mathbf{Q}\in\mathcal{Q}_m^h} \lim_{\gamma\to 0} \frac{\mathbb{E}_{\mathbf{Q}}[T^{\mathcal{A}(\gamma)}]}{\ln(\gamma^{-1})} \in \Theta(mh^{-2}) \quad (5)$$

as $\max\{m, h^{-1}\} \to \infty$. Unfortunately, these results do not yield any information for cases where γ is fixed. Moreover, the algorithmic solution $\mathcal{A}(\gamma)$ presented by Degenne and Koolen [2019] is very inefficient if not infeasible in practice, which is due to a hard min-max problem that has to be solved at each time step. In the following, we will discuss further lower and upper bounds on the worst-case sample complexity of solutions to $\mathcal{P}^{m,h,\alpha,\beta}$. Our results are to some extent stronger than (2) and (3) as they are also applicable if γ is fixed.

The following theorem provides an instance-wise lower bound on the sample complexity of any algorithm solving the testification problem for the CW on $\mathcal{Q}_m^h$.

Theorem 4.1. *For $h_0, \gamma_0 \in (0, 1/2)$ there exists a constant $c = c(h_0, \gamma_0) > 0$ with the following property: Let $h \in (0, h_0), \alpha, \beta \in (0, \gamma_0)$ and $\mathcal{A}$ be any solution to $\mathcal{P}^{m,h,\alpha,\beta}$. Then, for any $\mathbf{Q} \in \mathcal{Q}_m^h(\mathrm{CW})$, we have*

$$\mathbb{E}_{\mathbf{Q}}\left[T^{\mathcal{A}}\right] \geq c \sum\nolimits_{j\neq \mathrm{CW}(\mathbf{Q})} h_{\mathrm{CW}(\mathbf{Q}),j}^{-2} \ln\left(\gamma^{-1}\right),$$

where $\gamma := \max\{\alpha, \beta\}$ and $h_{i,j} := |q_{i,j} - 1/2|$ for every distinct i, j in $[m]$. In particular, $\mathcal{A}$ fulfills

$$\sup_{\mathbf{Q}\in\mathcal{Q}_m^h} \mathbb{E}_{\mathbf{Q}}[T^{\mathcal{A}}] \geq c(m-1)\ln\left(\gamma^{-1}\right)h^{-2}. \quad (6)$$

Theorem 4.1 (as well as (5)) reveals that the impact of the key quantities (i.e., α, β, m, h) on the order of the worst-case sample complexity for the testification problem is similar as for the worst-case sample complexity for the sole CW identification task under the stricter assumption of existence of a total order [Braverman et al., 2016]. Moreover, the dependency on the number of arms in (6) coincides with (4).

5 NEARLY OPTIMAL TESTIFICATION

In this section, we systematically provide practically feasible algorithmic solutions to the testification problem, starting with the straightforward approach that performs identification and testing separately one after the other. Next, we give a first naïve attempt to interleave identification and testing in an algorithmic solution whose obvious flaws together with some graph-theoretical considerations will help us to design a more sophisticated algorithmic solution.

5.1 NAÏVE APPROACHES

For the sake of convenience, let us consider first the symmetric case $\alpha = \beta =: \gamma$. To construct the first solution, suppose $\mathcal{A}$ to be an algorithm with parameters m, h, γ, which is able to find the Condorcet winner whenever[1] $\mathbf{Q} \in \mathcal{Q}_m^h(\mathrm{CW})$ with an error probability at most γ. Then, the algorithm $\mathcal{A}$-THEN-VERIFY, which executes $\mathcal{A}(m, h, \gamma/2)$, observes its output i (the alleged CW) and afterwards verifies with error probability at most $\gamma/2$ whether i is indeed the CW by querying each of the pairs $\{i, j\}, j \neq i$, sufficiently often, solves $\mathcal{P}^{m,h,\gamma,\gamma}$. Instantiating $\mathcal{A}$ with the SELECT algorithm [Mohajer et al., 2017], a state-of-the-art solution to the CW identification problem, results in a solution SELECT-THEN-VERIFY to $\mathcal{P}^{m,h,\gamma,\gamma}$. This approach achieves almost asymptotically optimal worst-case expected sample complexity (cf. Section C).

Despite the (almost) satisfactory theoretical guarantee of the latter approach it seems unfavorable to separate identification and testing in the learning process, as an unnecessary verification might be conducted at the end of the learning process. Quite naturally, the question arises how testing and identification could be interleaved in a suitable way, which formally boils down to construct an appropriate decision criterion.

As a gentle start for the development of our decision criterion, consider the following naïve criterion: For each pair $(i, j) \in (m)_2$, sample repeatedly noisy pairwise comparisons of the corresponding pairwise probability $q_{i,j}$ until being confident enough whether $q_{i,j}$ is above or below $1/2$ with confidence $\geq 1 - \gamma'$ for some $\gamma' \in (0, 1)$, based on the pairwise probability estimates $(\hat{\boldsymbol{q}}_t)_{i,j} := \frac{(\boldsymbol{w}_t)_{i,j}}{(\boldsymbol{n}_t)_{i,j}}$. Then, decide for $\neg\mathrm{CW}$ in case $\hat{\boldsymbol{q}}_t := ((\hat{\boldsymbol{q}}_t)_{i,j})_{1\leq i,j\leq m}$ is in $\mathcal{Q}_m(\neg\mathrm{CW})$, and otherwise decide for $i^* = \mathrm{argmax}_{i\in[m]} \sum_{j\neq i} \mathbf{1}_{\{(\hat{\boldsymbol{q}}_t)_{i,j}>1/2\}}$, which necessarily exists in this case. Let us denote the resulting algorithm by $\mathcal{A}^{\mathrm{naive}}$ and let $\gamma' = \gamma/\binom{m}{2}$. Then, by virtue of independence of the individual stopping decisions in the pairwise samplings, and Bernoulli's inequality,

$$\begin{aligned}\mathbb{P}_{\mathbf{Q}}(\mathbf{D}(\mathcal{A}^{\mathrm{naive}}) = \neg\mathrm{CW}) &= \mathbb{P}_{\mathbf{Q}}(\hat{\boldsymbol{q}}_t \in \mathcal{Q}_m(\neg\mathrm{CW})) \\ &\geq (1-\gamma')^{\binom{m}{2}} \geq 1-\gamma\end{aligned}$$

holds for any $\mathbf{Q} \in \mathcal{Q}_m^h(\neg\mathrm{CW})$. Similarly, the first inequality in (1) holds, i.e., $\mathcal{A}^{\mathrm{naive}}$ is a solution for the testification problem for CW on $\mathcal{Q}_m^h$ for α and β. Evidently, however,

[1] In particular, $\mathcal{A}$ is not required to have any guarantees for $\mathbf{Q} \in \mathcal{Q}_m^h(\neg\mathrm{CW})$ here.

this algorithm has an expected sample complexity that depends quadratically on the number of arms m. In addition, it is not clearly specified what it means that $\mathcal{A}^{\text{naive}}$ is "confident enough" about the sign of $q_{i,j} - 1/2$.

In order to overcome the obvious flaws of $\mathcal{A}^{\text{naive}}$, we formulate the following questions, the answers of which will lead us to a more sophisticated algorithm for the testification problem:

(i) How can we decide, as early as possible and with confidence $\geq 1-\gamma'$, based on $(\hat{\boldsymbol{q}}_t)_{i,j}$, whether $q_{i,j} > 1/2$ or $q_{i,j} < 1/2$ holds?
(ii) Do we need to be sure about the sign of $q_{i,j} - 1/2$ for *all* pairs $(i,j) \in (m)_2$?
(iii) Is the choice $\gamma' = \gamma/\binom{m}{2}$ necessary, or is $\gamma' = \gamma/m'$ with some $m' < \binom{m}{2}$ sufficient?

We answer the first question by using suitable anytime confidence bounds on the pairwise winning probabilities $q_{i,j}$, depending on the desired confidence level γ' and low noise parameter h, and quit the repeated sampling of noisy pairwise comparisons as soon as $1/2$ is not an element of the confidence bounds around the empirical estimate $(\hat{\boldsymbol{q}}_t)_{i,j}$ anymore. To be more precise, we choose

$$C_{h,\gamma'}(n) \coloneqq \frac{1}{2n}\left\lceil \frac{\ln((1-\gamma')/\gamma')}{\ln\left((1/2+h)/(1/2-h)\right)} \right\rceil, \qquad (7)$$

sample until

$$(\hat{\boldsymbol{q}}_t)_{i,j} \notin \Big[1/2 - C_{h,\gamma'}((\boldsymbol{n}_t)_{i,j}), 1/2 + C_{h,\gamma'}((\boldsymbol{n}_t)_{i,j})\Big],$$

and decide for "$q_{i,j} > 1/2$" if $(\hat{\boldsymbol{q}}_t)_{i,j} - C_{h,\gamma'}((\boldsymbol{n}_t)_{i,j}) > 1/2$ holds at this time t, and for "$q_{i,j} < 1/2$" otherwise. Provided $|q_{i,j} - 1/2| > h$ holds, our estimation of the sign of $q_{i,j} - 1/2$ is correct with probability $\geq 1-\gamma'$ and requires, in expectation, in the worst case at most $\mathcal{O}(h^{-2}\ln(\gamma^{-1}))$ observations (i.e., i.i.d. samples) of a duel between the pair of arms (i,j). The choice of $C_{h,\gamma'}$ is based on the sequential probability ratio test, which is known to be optimal in some sense in case $|q_{i,j} - 1/2| \approx h$ (cf. Lemma B.1).

For questions (ii) and (iii), it will turn out to be fruitful to exploit a connection of the testification problem to graph-theoretical concepts of tournaments. In particular, we will see that question (ii) can be answered negatively, while the answer to question (iii) is $\gamma' = \gamma/m$ in the symmetric case and $\gamma' = \min\{\frac{\alpha}{m}, \frac{\beta}{m-1}\}$ in the asymmetric case.

5.2 GRAPH-THEORETICAL CONCEPTS

We write $\mathcal{G}_m$ for the set of all simple digraphs on $[m]$ without loops and with at most one edge between each two nodes. In other words, $\mathcal{G}_m$ contains all directed graphs $G = ([m], E_G)$ with $E_G \subseteq \langle m\rangle_2$ such that $(i,j) \notin E_G$ or $(j,i) \notin E_G$ holds for every distinct $i,j \in [m]$. Let $\overline{\mathcal{G}}_m$ be the set of tournaments on $[m]$, i.e., $\overline{\mathcal{G}}_m \subseteq \mathcal{G}_m$ contains all digraphs $G = ([m], E_G)$, where for every $(i,j) \in \langle m\rangle_2$ either $(i,j) \in E_G$ (we write $i \to j$ in G, or $i \xrightarrow{G} j$) or $(j,i) \in E_G$. With this, it is possible to establish a connection between tournaments and binary preference relations.

Fact 5.1. *There is a one-to-one connection given by a mapping $\Phi : \mathcal{Q}_m^{\text{Bin}} \to \overline{\mathcal{G}}_m$ between* binary preference relations $\mathbf{Q} \in \mathcal{Q}_m^{\text{Bin}} \coloneqq \{\mathbf{Q}' \in \mathcal{Q}_m : q'_{i,j} \in \{0,1\}\, \forall (i,j) \in \langle m\rangle_2\}$ *and tournaments $G \in \overline{\mathcal{G}}_m$ such that $q_{i,j} = 1$ iff $i \to j$ in $G = \Phi(\mathbf{Q})$.*

Due to Fact 5.1, we may say that $G \in \overline{\mathcal{G}}_m$ *has a CW* if $\Phi^{-1}(G) \in \mathcal{Q}_m(\mathrm{CW})$, i.e., G has a CW iff there exists some $i \in [m]$ with $i \to j$ in G for every $j \in [m] \setminus \{i\}$. We define $\overline{\mathcal{G}}_m(\mathrm{CW})$ as the set of all $G \in \overline{\mathcal{G}}_m$ which have a CW and $\overline{\mathcal{G}}_m(\neg\mathrm{CW}) \coloneqq \overline{\mathcal{G}}_m \setminus \overline{\mathcal{G}}_m(\mathrm{CW})$. For any $G \in \overline{\mathcal{G}}_m(\mathrm{CW})$, we denote by $\mathrm{CW}(G)$ the Condorcet winner of G, i.e., $\mathrm{CW}(G) \to j$ holds in G for all $j \in [m] \setminus \{\mathrm{CW}(G)\}$. For every $i^* \in [m]$, we write $\overline{\mathcal{G}}_m(i^*) \coloneqq \{G \in \overline{\mathcal{G}}_m(\mathrm{CW}) \,|\, \mathrm{CW}(G) = i^*\}$.

We call a tournament $G' \in \overline{\mathcal{G}}_m$ an *extension of* $G \in \mathcal{G}_m$ if $E_G \subseteq E_{G'}$ holds. Further, define for $\mathrm{X} \in \{\mathrm{CW}, \neg\mathrm{CW}, i^*\}$ the set $\mathcal{G}_m(\mathrm{X})$ as the set of all $G \in \mathcal{G}_m$, for which $G' \in \overline{\mathcal{G}}_m(X)$ for every extension G' of G. As $\mathbf{Q} \in \mathcal{Q}_m^h$ has i^* as CW iff the graph $G_{\mathbf{Q}} \in \overline{\mathcal{G}}_m$ defined via "$i \to j$ in $G_{\mathbf{Q}}$ iff $q_{i,j} > 1/2$" fulfills $\mathrm{CW}(G_{\mathbf{Q}}) = i^*$, we obtain: If we know $\forall j \in [m] \setminus \{i^*\} : i^* \to j$ in $G_{\mathbf{Q}}$ — which is equivalent to $G_{\mathbf{Q}}$ having a subgraph $\tilde{G} \in \mathcal{G}_m(i^*)$ (Lemma A.4) — with confidence $\geq 1-\alpha$, then deciding for i^* in the testification problem is correct with probability $\geq 1-\alpha$. Similarly, if $G_{\mathbf{Q}}$ contains a subgraph $\tilde{G} \in \mathcal{G}_m(\neg\mathrm{CW})$ with confidence $\geq 1-\beta$, then the decision $\neg$CW is correct with probability $\geq 1-\beta$. Note that $G \in \mathcal{G}_m(\neg\mathrm{CW})$ is equivalent to $\forall\, i \in [m]\, \exists\, j \in [m] \setminus \{i\} : j \xrightarrow{G} i$ (Proposition A.3).

The notion of $\mathcal{G}_m(\mathrm{CW})$ is not required for the testification problem, but it will play a major role for testing whether a CW exists (Section 6). Note that due to $\mathcal{G}_m(\mathrm{CW}) \supsetneq \bigcup_{i^* \in [m]} \mathcal{G}_m(i^*)$, the notion is not redundant, and we also obtain an appropriate characterization of it (Proposition A.1).

5.3 NOISY TOURNAMENT SAMPLING

We incorporate the above graph-theoretical observations into a more sophisticated testification algorithm, which we call the *Noisy Tournament Sampling* (NTS) and denote by $\mathcal{A}^{\text{NTS}}$. The name stems from the resemblance of its underlying sampling idea to noisy sorting algorithms [Braverman and Mossel, 2008], which will be described more thoroughly in the following.

The algorithm $\mathcal{A}^{\text{NTS}}$ maintains a graph $\hat{G}_t \coloneqq ([m], \hat{E}_t)$ and successively adds edges (corresponding to pairs $(i,j) \in \langle m\rangle_2$) to $\hat{G}_t$, for which at time t the algorithm $\mathcal{A}^{\text{NTS}}$ is

Algorithm 1 $\mathcal{A}^{\mathrm{NTS}}$: Noisy tournament sampling

Input: α, β, h, π
Initialization: $\boldsymbol{n}_0 \leftarrow \boldsymbol{w}_0 \leftarrow (0)_{1 \leq i,j \leq m}$, $\hat{E}_0 \leftarrow \emptyset$, $\gamma' \leftarrow \min\{\frac{\alpha}{m}, \frac{\beta}{m-1}\}$, $C_{h,\gamma'}$ as in (7)

1: **for** $t \in \mathbb{N}$ **do**
2: $(i,j) \sim \pi(t, (\boldsymbol{n}_{t'}, \boldsymbol{w}_{t'})_{0 \leq t' \leq t-1})$
3: Observe $X_{i,j}^{[t]} \sim \mathrm{Ber}(q_{i,j})$
4: Define $\boldsymbol{w}_t$ via $(\boldsymbol{w}_t)_{k,l} \leftarrow (\boldsymbol{w}_{t-1})_{k,l} + \mathbf{1}_{\{\{k,l\}=\{i,j\} \text{ and } X_{k,l}^{[t]}=1\}}$ $\forall 1 \leq k,l \leq m$
5: Define $\boldsymbol{n}_t$ via $(\boldsymbol{n}_t)_{k,l} \leftarrow (\boldsymbol{n}_{t-1})_{k,l} + \mathbf{1}_{\{\{k,l\}=\{i,j\}\}}$ $\forall 1 \leq k,l \leq m$
6: $\hat{E}_t \leftarrow \hat{E}_{t-1}$
7: **if** $(\hat{\boldsymbol{q}}_t)_{i,j} > \frac{1}{2} + C_{h,\gamma'}((\boldsymbol{n}_t)_{i,j})$ **then**
8: $\hat{E}_t \leftarrow \hat{E}_t \cup \{(i,j)\}$
9: **if** $(\hat{\boldsymbol{q}}_t)_{i,j} < \frac{1}{2} - C_{h,\gamma'}((\boldsymbol{n}_t)_{i,j})$ **then**
10: $\hat{E}_t \leftarrow \hat{E}_t \cup \{(j,i)\}$
11: $\hat{G}_t \leftarrow ([m], \hat{E}_t)$
12: **if** $\exists i^* \in [m] : \hat{G}_t \in \mathcal{G}_m(i^*)$ **then**
13: **return** i^*
14: **if** $\hat{G}_t \in \mathcal{G}_m(\neg\mathrm{CW})$ **then return** $\neg\mathrm{CW}$

confident with level $1-\gamma'$ that $q_{i,j} > \frac{1}{2}$ holds (lines 7–10). $\mathcal{A}^{\mathrm{NTS}}$ stops only in two cases: One in which the graph $\hat{G}_t$ is in $\mathcal{G}_m(\neg\mathrm{CW})$, i.e., none of its tournament extensions can bring forth a CW (line 14), the other in which the graph $\hat{G}_t$ is in $\mathcal{G}_m(i^*)$ for some $i^* \in [m]$, i.e., all tournament extensions are preference relations with i^* as CW. According to which event caused the termination, either the supposed CW (i.e., i^*) or $\neg\mathrm{CW}$ is returned (lines 12–14). Formally, we have $\mathbf{D}(\mathcal{A}^{\mathrm{NTS}}) = i^*$ if $\hat{G}_t \in \mathcal{G}_m(i^*)$ and $\mathbf{D}(\mathcal{A}^{\mathrm{NTS}}) = \neg\mathrm{CW}$ if $\hat{G}_t \in \mathcal{G}_m(\neg\mathrm{CW})$. Regarding the definition of $\mathcal{G}_m(i^*)$ and $\mathcal{G}_m(\neg\mathrm{CW})$ (as well as Lemma E.1 below), termination is only reasonable if $\hat{G}_t$ is in $\bigcup_{i^* \in [m]} \mathcal{G}_m(i^*) \cup \mathcal{G}_m(\neg\mathrm{CW})$. Proposition A.3 and Lemma A.4 indicate that the used correction term in the choice of γ' is optimal.

5.4 THE PASSIVE SCENARIO

In this section we analyze the passive testification scenario, where the sampling strategy π might *not* be specifically designed in order to ensure a quick termination of the testing algorithm. In other words, π might be any sampling strategy that interacts with the underlying dueling bandit problem as stipulated by Definition 3.1.

In light of this, we let Π be the set of all sampling strategies (Definition 3.1) and denote by Π_∞ the family of sampling strategies π that sample every pair $\{i,j\}$ almost surely (a.s.) infinitely often, which means that $(\boldsymbol{n}_t)_{i,j} \to \infty$ a.s. as $t \to \infty$.

Note that if $\pi \in \Pi \setminus \Pi_\infty$, a sampling strategy $\hat{\pi} \in \Pi$ that chooses $\hat{\pi}(t, (\boldsymbol{n}_{t'}, \boldsymbol{w}_{t'})_{0 \leq t' \leq t-1}) = \pi(t, (\boldsymbol{n}_{t'}, \boldsymbol{w}_{t'})_{0 \leq t' \leq t-1})$ with probability $1 - \frac{1}{t}$, and otherwise $\hat{\pi}(t, (\boldsymbol{n}_{t'}, \boldsymbol{w}_{t'})_{0 \leq t' \leq t-1})$ picks a pair (i,j) uniformly at random from $\langle m \rangle_2$ with probability $\frac{1}{t}$, fulfills $\hat{\pi} \in \Pi_\infty$ and $\mathbb{P}(\pi(t, (\boldsymbol{n}_{t'}, \boldsymbol{w}_{t'})_{0 \leq t' \leq t-1}) \neq \hat{\pi}(t, (\boldsymbol{n}_{t'}, \boldsymbol{w}_{t'})_{0 \leq t' \leq t-1})) \leq \frac{1}{t} \to 0$ as $t \to \infty$. Hence, $\hat{\pi}$ and π behave similarly in the limit. This shows that the assumption $\pi \in \Pi_\infty$, which is required for theoretical results in our framework, is rather mild.

Theorem 5.2. *Let $\pi \in \Pi_\infty$, $h \in (0, 1/2)$ and $\alpha, \beta \in (0,1)$ be fixed. Let $\mathcal{A}$ be Algorithm 1, called with the parameters h, α, β and π as sampling strategy. Then, $\mathcal{A}$ solves the testification problem for the CW on $\mathcal{Q}_m^h$ for α and β.*

By (passively) monitoring the statistical validity of the CW assumption, the algorithmic framework presented in Theorem 5.2 can be utilized in order to justify the usage of dueling bandits algorithms focusing on alternative best arm concepts for the goal of regret minimization, if the test component detects a violation of the CW assumption. Finally, it is worth noting that $\mathcal{A}$-THEN-VERIFY cannot be used in a sensible way for this passive scenario due to the strictly separated identification and testing phases.

5.5 THE ACTIVE SCENARIO

The key question is how to construct a sampling strategy π such that $\mathcal{A}^{\mathrm{NTS}}$ terminates as soon as possible. Apparently, one needs to construct the internal tournament $\hat{G}_t$ in Algorithm 1 such that it quickly becomes clear whether each extension admits a CW or not (cf. lines 12 and 14). Thus, a natural approach would be to build this tournament according to a deterministic sequential testing algorithm (DSTA) for testification of the CW in a tournament, as those are commonly designed specifically for that purpose. However, as the outcome of a duel in the underlying problem is in general not deterministic, one has to conduct the duels several times until having enough confidence on the actual pairwise probability.

Based on these considerations, we define an epoch-based sampling strategy (implicitly defined by lines 1, 10 and 14 in Algorithm 6 in the supplement) using $\mathcal{A}_{\mathrm{Bin}}$ to determine which pair shall be sampled repeatedly during an epoch. To be more precise, at the beginning of each epoch, the noisy tournament sampling strategy queries the black-box DSTA $\mathcal{A}_{\mathrm{Bin}}$ to provide a pair, say (i,j), for a duel. This duel is repeated until the sign of $q_{i,j} - 1/2$ is determined with a specific confidence (based on α, β and h) leading to both, the end of the current epoch, and no consideration of the pair in any upcoming epoch (lines 3–15). If the sign is assumed to be positive resp. negative, $\mathcal{A}_{\mathrm{Bin}}$ is provided with the feedback $i \to j$ resp. $j \to i$, as if no randomness was involved, leading $\mathcal{A}_{\mathrm{Bin}}$ either to suggest the next pair to be queried (lines 10 and 14) or to terminate. If $\mathcal{A}_{\mathrm{Bin}}$ terminates before $\mathcal{A}^{\mathrm{NTS}}$ came to a decision, we suppose

$\mathcal{A}^{\mathrm{NTS}}$ to continue until its termination by choosing the duels uniformly at random from $\langle m \rangle_2$ (lines 19–21). As a result, we obtain Algorithm 6, which is essentially a modification of Algorithm 1, where line 2 is replaced by the just described sampling mechanism based on interaction with $\mathcal{A}_{\mathrm{Bin}}$.

In order to state our main findings, we introduce some further terminology. We say that a DSTA $\mathcal{A}_{\mathrm{Bin}}$ is *correct for the testification of the CW* (or simply *testification-correct*) if it outputs the correct decision for any tournament $G \in \overline{\mathcal{G}}_m$, i.e., whether a CW exists, and if so, it returns the CW. Further, denote by $\mathfrak{A}^*_m$ the set of all testification-correct DSTAs, and for any DSTA $\mathcal{A}_{\mathrm{Bin}}$, let $T^{\mathcal{A}_{\mathrm{Bin}}}$ be its worst-case sample complexity over all tournaments $\overline{\mathcal{G}}_m$, that is, $T^{\mathcal{A}_{\mathrm{Bin}}} = \max_{G \in \overline{\mathcal{G}}_m} T_G^{\mathcal{A}_{\mathrm{Bin}}}$, where $T_G^{\mathcal{A}_{\mathrm{Bin}}}$ is the sample complexity of $\mathcal{A}_{\mathrm{Bin}}$ for G.

Lemma E.1 ensures that, as soon as any testification-correct black-box DSTA $\mathcal{A}_{\mathrm{Bin}}$ used in the noisy tournament sampling strategy described above terminates, $\mathcal{A}^{\mathrm{NTS}}$ terminates, too (i.e., returns $\neg$CW or a candidate for the CW).

Theorem 5.3. *Let $\mathcal{A}_{\mathrm{Bin}}$ be a DSTA and $\gamma_0 \in (0, 1/2)$ be fixed. Then, for any $\alpha, \beta \in (0, \gamma_0)$ and $h \in (0, 1/2)$, the noisy sorting algorithm $\mathcal{A}^{\mathrm{NTS}}$ (Algorithm 6) called with the parameters h, α, β and $\mathcal{A}_{\mathrm{Bin}}$ as its black-box DSTA, solves the testification problem for the CW on $\mathcal{Q}^h_m$ for α and β. If $\mathcal{A}_{\mathrm{Bin}} \in \mathfrak{A}^*_m$ and $\gamma' = \min\{\frac{\alpha}{m}, \frac{\beta}{m-1}\}$, then*

$$\mathbb{E}_{\mathbf{Q}}[T^{\mathcal{A}^{\mathrm{NTS}}}] \in \mathcal{O}(T^{\mathcal{A}_{\mathrm{Bin}}} h^{-2} \ln(\gamma'^{-1}))$$

as $\max\{h^{-1}, \gamma'^{-1}\} \to \infty$ for any $\mathbf{Q} \in \mathcal{Q}^h_m$.

With slightly more effort we obtain an instance-wise version of Theorem 5.3, in which the runtime on any instance $\mathbf{Q}$ depends – similarly as in our lower bound stated in Theorem 4.1 – on some of the terms $|q_{i,j} - 1/2|$, $(i,j) \in (m)_2$. As the worst-case bound from Theorem 5.3 is easier and sufficient for our further discussion, we provide the details of the instance-wise version in the supplement (Theorem H.1).

It is well known that the optimal worst-case sample complexity for a testification-correct DSTA is $2m - \lfloor \log_2 m \rfloor - 2$ [Procaccia, 2008, Balasubramanian et al., 1997, Bollobás and Eldridge, 1978]. For the sake of completeness, we state this algorithm and the result in the supplement (Algorithm 3, Proposition E.2). As a direct consequence of Theorem 5.3, we thus obtain the following result.

Corollary 5.4. *Let $\gamma_0 \in (0, 1/2)$ be fixed. The noisy tournament sampling algorithm $\mathcal{A}^{\mathrm{NTS}}$ used with the parameters $h \in (0, 1/2)$, $\alpha = \beta = \gamma \in (0, \gamma_0)$ and $\mathcal{A}_{\mathrm{Bin}}$ as defined in Algorithm 3 as its black-box DSTA, solves the testification problem for the CW on $\mathcal{Q}^h_m$ for $\alpha = \beta = \gamma$ such that*

$$\sup_{\mathbf{Q} \in \mathcal{Q}^h_m} \mathbb{E}_{\mathbf{Q}}[T^{\mathcal{A}^{\mathrm{NTS}}}] \in \mathcal{O}\left(m \ln(m) h^{-2} \ln(\gamma^{-1})\right)$$

as $\max\{m, h^{-1}, \gamma^{-1}\} \to \infty$.

According to Theorem 4.1, the algorithm $\mathcal{A}^{\mathrm{NTS}}$ from Corollary 5.4 is optimal w.r.t. the worst-case expected sample complexity up to a factor of $\ln m$ for the CW testification problem on instances $\mathbf{Q} \in \mathcal{Q}^h_m$.

In contrast to (5), Theorem 4.1 and Corollary 5.4 specifically yield asymptotic lower and upper sample complexity bounds for solving $\mathcal{P}^{m,h,\gamma,\gamma}$ if γ is fixed. Moreover, we are able to show that the algorithmic solution in Corollary 5.4 is in expectation superior over the straightforward SELECT-THEN-VERIFY approach (see Lemma C.2).

6 OTHER CW-RELATED PROBLEMS

Any solution to the CW testification problem can easily be modified to solve the following binary classification problem with classes CW and $\neg$CW:

> **Check**_CW: Is $\mathbf{Q}$ in $\mathcal{Q}_m(\mathrm{CW})$? If so, return CW, otherwise return $\neg$CW.

Here, a type I/II error of the testing algorithm corresponds to a false positive/negative classification.

Another practically relevant problem related to the testification problem is the scenario in which one has a guess for the CW beforehand and would like to check its validity:

> **Verify_i_as**_CW (with input $i \in [m]$): Is $\mathrm{CW}(\mathbf{Q}) = i$? If so, return i, otherwise return $\neg i$.

Note that any algorithmic solution for **Verify_i_as**_CW is a candidate for the testing procedure of the $\mathcal{A}$-THEN-VERIFY approach in Section 5.1.

Another important use case of the **Verify_i_as**_CW problem is in verifying the validity of a ranking over the arms, say $\succ$, which shall be consistent with $\mathbf{Q}$ in the sense that $q_{i,j} > 1/2$ iff $i \succ j$. Such a ranking could be the output of some ranking learning algorithm in the realm of the dueling bandits setting, for instance. By iteratively verifying the arm with rank $i \in [m]$ to be the CW among the arms with lower ranks, one needs at most $m - 1$ executions of **Verify_i_as**_CW to decide, up to some adjustable confidence, whether $\succ$ is correct.

It is worth noting that the **Verify_i_as**_CW problem can, in contrast to **Check**_CW, be considered in two variants: One variant with already assuming $\mathbf{Q} \in \mathcal{Q}_m(\mathrm{CW})$ (which will be denoted by $\exists$CW for convenience) and in the other without any assumption. The former variant has also been considered by Urvoy et al. [2013] and Karnin [2016] for constructing an algorithmic solution to identify the CW.

Moreover, the testification problem for the CW under the assumption of its existence corresponds to the well-known *best-arm-identification* problem in the realm of dueling bandits, for which lower and (matching) upper bounds have already been established by Braverman et al. [2016] for

Table 1: Sample complexity bounds for **Check**_CW, **Verify_i_as**_CW and testification for the CW († : [Braverman et al., 2016]; ‡ : [Procaccia, 2008, Balasubramanian et al., 1997, Bollobás and Eldridge, 1978])

	probabilistic setting		deterministic setting	
Problem	no assumption	∃CW	no assumption	∃CW
Check_CW	$\widetilde{\Theta}(mh^{-2}\ln\gamma^{-1})$	-	$2m - \lfloor \log_2 m \rfloor - 2$ (‡)	-
Verify_i_as_CW	$\widetilde{\Theta}(mh^{-2}\ln\gamma^{-1})$	$\widetilde{\Theta}(mh^{-2}\ln\gamma^{-1})$	$m-1$	$m-1$
Testification for the CW	$\widetilde{\Theta}(mh^{-2}\ln\gamma^{-1})$	$\widetilde{\Theta}(mh^{-2}\ln\gamma^{-1})$ (†)	$2m - \lfloor \log_2 m \rfloor - 2$ (‡)	$m-1$

preference relations in $\mathcal{Q}_m^h$.

Table 1 summarizes upper and lower bounds for the worst-case (expected) sample complexities on $\mathcal{Q}_m^h$-instances for the different problems in the deterministic setting, i.e., $\mathbf{Q} \in \mathcal{Q}_m^{\mathrm{Bin}}$ and DSTAs, as well as for the probabilistic setting (with a tolerated error probability $\leq \gamma$). Therein, we conveniently write $\widetilde{\Theta}$ for Θ, which hides $\ln m$-factors in the probabilistic setting. Note that Table 1 contains *sharp* bounds for the runtime of the deterministic variants of the problems. The corresponding proofs can be found in Sections E, F and G.

For all of these problems, a reduction to the pure exploration bandit problem with multiple correct answers from Degenne and Koolen [2019] yields a solution with asymptotic optimality in some sense (cf. Section J). Just like for testification of the CW, we also provide instance-wise lower bounds for **Check**_CW and **Verify_i_as**_CW in the supplement.

7 EXPERIMENTAL STUDY

7.1 THE ACTIVE SCENARIO

In this section, we present a small experimental study to illustrate the performance of our algorithm $\mathcal{A}^{\mathrm{NTS}}$ for CW testification, while further experiments can be found in the supplement.[2] To this end, we start with a comparison between $\mathcal{A}^{\mathrm{NTS}}$ and SELECT-THEN-VERIFY. To guarantee stability of the results, we average over 25000 runs, each time sampling the ground truth relation $\mathbf{Q}$ at random from $\mathcal{Q}_5^{0.05}$. Since $\mathbf{Q}$ is usually unknown in practice, we try out and compare different values of h as parameters for $\mathcal{A}^{\mathrm{NTS}}$ and SELECT-THEN-VERIFY, as well as different values for the guaranteed error probability γ. These free parameters cause a variation in the used confidence bounds and thus in the number of iterations before the algorithms terminate.

The curves in Figure 1, which have been produced through variation of the parameter γ, illustrate the compromise between the success rate and the number of iterations of the algorithms (decreasing γ increases the success rate but also the sample complexity). As can be seen, the curves

[2]Our implementation is provided at `https://github.com/bjoernhad/CondorcetWinnerTestification`.

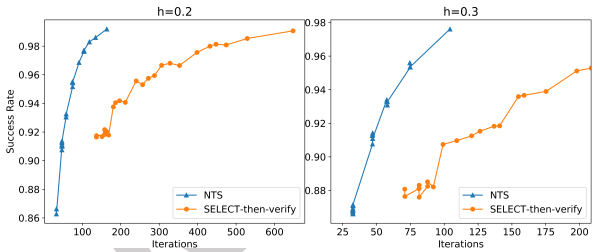

Figure 1: Success rate and total number of comparisons until termination for the proposed $\mathcal{A}^{\mathrm{NTS}}$ and SELECT-THEN-VERIFY for different values of the gap h to $1/2$.

of $\mathcal{A}^{\mathrm{NTS}}$ dominate the curves of SELECT-THEN-VERIFY for $h = 0.2$ and $h = 0.3$. Indeed, with the same number of comparisons, $\mathcal{A}^{\mathrm{NTS}}$ achieves a higher success rate than SELECT-THEN-VERIFY, regardless of the parameter h. In fact, $\mathcal{A}^{\mathrm{NTS}}$ seems to be quite robust towards incorrect choices of this parameter.

In the supplementary material, we repeat this experiment for a larger number of arms as well as for the case where $\mathbf{Q}$ is sampled uniformly at random from $\mathcal{Q}_5^{0.05}(\mathrm{CW})$ resp. $\mathcal{Q}_5^{0.05}(\neg\mathrm{CW})$, with the result that $\mathcal{A}^{\mathrm{NTS}}$ outperforms SELECT-THEN-VERIFY in any considered case as well (see Section I). There, we also compare $\mathcal{A}^{\mathrm{NTS}}$ and SELECT-THEN-VERIFY in case h is smaller than 0.2.

7.2 THE PASSIVE SCENARIO

Finally, we demonstrate how the passive setting described in Section 5.4 can be utilized in order to justify the usage of dueling bandits algorithms focusing on alternative best arm concepts for the goal of regret minimization, if the test component detects a violation of the CW assumption. For this purpose, we consider two sampling strategies:

- Relative Upper Confidence Bound (RUCB) from [Zoghi et al., 2014], which is a dueling bandit algorithm based on the Condorcet Winner assumption.
- Double Thompson Sampling (DTS) from [Wu and Liu, 2016], which is a dueling bandit algorithm not relying on the Condorcet Winner assumption, but instead focusing on the set of Copeland winners.

Further, we consider the regret based on the difference in the

normalized Copeland scores of the Copeland winner and the two chosen arms (cf. [Zoghi et al., 2015]). It is well known that RUCB can achieve linear regret in case no Condorcet winner exists, while DTS provably only suffers sub-linear regret (with respect to the Copeland scores) in such cases.

In light of this, we consider a two-staged algorithm (denoted by RUCB→DTS), which executes in its first stage Algorithm 1 (with parameters $\gamma = 0.1$ and $h = 0.1$) instantiated with RUCB as its sampling strategy, and in its second stage simply DTS in case that no Condorcet winner exists in the ground truth relation and otherwise RUCB. That is, RUCB→DTS switches from the CW-based sampling strategy RUCB to the Copeland winner based DTS algorithm if the CW assumption is likely violated.

For evaluating the algorithms we choose as the underlying ground-truth preference relation the Hudry-tournament $\mathbf{Q}_{\text{Hudry}} \in \mathcal{Q}_{13}$, which does not have a CW and has a Copeland set of size 1 (cf. [Ramamohan et al., 2016]); for the sake of convenience, it is deferred to the appendix (see Section I.3). Figure 2 illustrates the benefit of changing from RUCB to DTS with regard to the regret. In particular, RUCB→DTS does not suffer the linear regret of RUCB but instead its cumulative regret appears only by a constant term larger than that of DTS.

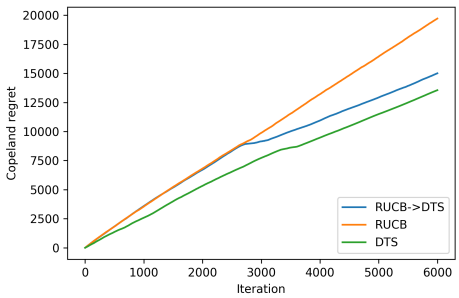

Figure 2: Copeland regret of DTS, RUCB and RUCB→DTS on $\mathbf{Q}_{\text{Hudry}}$

8 CONCLUSION

We introduced the testification problem for the Condorcet winner in dueling bandits. This problem extends the best arm identification problem and asks for simultaneously testing the CW assumption and identifying the CW in case it exists. Thus, instead of taking this assumption for granted, the learner is supposed to detect a possible violation (as quickly as possible). We provided an instance-wise lower bound on the expected number of samples needed by any learning algorithm solving the CW testification problem for preference relations satisfying the low noise assumption. Further, by exploiting a connection between tournaments and the CW of a (binary) preference relation, we proposed the noisy tournament sampling algorithm for the testification problem, which is shown to decide correctly with arbitrarily high confidence under mild assumptions on the underlying sampling strategy. By mimicing the query behavior of an optimal deterministic CW testification algorithm, noisy tournament sampling is shown to achieve an expected sample complexity that is optimal up to logarithmic terms.

Our novel problem setting opens up several lines for future work. In the first place, it would be tempting to extend the testification problem to stochastic types of transitivities of the preference relation, which are explicitly or implicitly assumed by several dueling bandits learning algorithms. Next, theoretical properties of the suggested testing component should be analyzed for structural properties of the preference relation other than the low noise assumption. Finally, the very idea of testification, i.e., to equip a learner with the ability to provide a sanity check of its own assumptions, is of course not restricted to the dueling bandits setting and could also be applied to other machine learning problems.

Acknowledgements

The authors gratefully acknowledge financial support by the German Research Foundation (DFG).

References

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 856–864, 2014.

Ramachandran Balasubramanian, Venkatesh Raman, and G. Srinivasaragavan. Finding scores in tournaments. *Journal of Algorithms*, 24(2):380–394, 1997.

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*, 22(7):1–108, 2021.

Béla Bollobás and Stephen E. Eldridge. Packings of graphs and applications to computational complexity. *Journal of Combinatorial Theory, Series B*, 25(2):105–124, 1978.

Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In *Proceedings of Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 268–276, 2008.

Mark Braverman, Jieming Mao, and S. Matthew Weinberg. Parallel algorithms for select and partition with noisy comparisons. In *Proceedings of the ACM symposium on Theory of Computing*, pages 851–862, 2016.

Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models:

The case of Mallows. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1071–1079, 2014.

Bangrui Chen and Peter I. Frazier. Dueling bandits with weak regret. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 731–739, 2017.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 7060–7070, 2017a.

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1088–1096, 2017b.

Moein Falahatgar, Ayush Jain, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. The limits of maxing, ranking, and preference learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1426–1435, 2018.

Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse dueling bandits. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 416–424, 2015.

Zohar Karnin. Verification based solution for structured MAB problems. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 145–153, 2016.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Proceedings of Annual Conference on Learning Theory (COLT)*, pages 1141–1154, 2015.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1235–1244, 2016.

Anna Korba, Stephan Clémençon, and Eric Sibony. A learning theory of ranking aggregation. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1001–1010, 2017.

Wataru Kumagai. Regret analysis for continuous dueling bandit. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 1488–1497, 2017.

Chang Li, Ilya Markov, Maarten De Rijke, and Masrour Zoghi. MergeDTS: A method for effective large-scale online ranker evaluation. *ACM Transactions on Information Systems (TOIS)*, 38(4):1–28, 2020.

Lucas Maystre and Matthias Grossglauser. Just sort it! A simple and effective approach to active preference learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 2344–2353, 2017.

Soheil Mohajer, Changho Suh, and Adel Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 2488–2497, 2017.

Ariel D. Procaccia. A note on the query complexity of the Condorcet winner problem. *Information Processing Letters*, 108(6):390–393, 2008.

Siddartha Ramamohan, Arun Rajkumar, and Shivani Agarwal. Dueling bandits: Beyond Condorcet winners to general tournament solutions. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 1253–1261, 2016.

Yanan Sui, Masrour Zoghi, Katja Hofmann, and Yisong Yue. Advancements in dueling bandits. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5502–5510, 2018.

Balázs Szörényi, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Online rank elicitation for Plackett-Luce: A dueling bandits approach. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 604–612, 2015.

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

Huasen Wu and Xin Liu. Double Thompson sampling for dueling bandits. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 649–657, 2016.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1201–1208, 2009.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 241–248, 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, 2012.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 10–18, 2014.

Masrour Zoghi, Zohar Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland dueling bandits. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 307–315, 2015.