Dueling Bandits with Delayed Feedback

Jasmin Brandt JASMIN.BRANDT@UNI-PADERBORN.DE
University of Paderborn, Germany

Björn Haddenhorst
University of Paderborn, Germany

Viktor Bengs
LMU Munich, Germany,
Munich Center for Machine Learning, Germany

Eyke Hüllermeier
LMU Munich, Germany,
Munich Center for Machine Learning, Germany

Abstract

Dueling Bandits is a well-studied extension of the Multi-Armed Bandits problem, in which the learner must select two arms in each time step and receives a binary feedback as an outcome of the chosen duel. However, all of the existing best arm identification algorithms for the Dueling Bandits setting assume that the feedback can be observed immediately after selecting the two arms. If this is not the case, the algorithms simply do nothing and wait until the feedback of the recent duel can be observed, which is a waste of runtime. We propose an algorithm that can already start a new duel even if the previous one is not finished and thus is much more time efficient. Our arm selection strategy balances the expected information gain of the chosen duel and the expected delay until we observe the feedback. By theoretically grounded confidence bounds we can ensure that the arms we discard are not the best arms with high probability.

Keywords: Multi-Armed Bandits, Dueling Bandits, Online Learning

1. Introduction and Related Work

The Dueling Bandits or also called Preference-based Multi-Armed Bandits is a variant of the standard Multi-Armed Bandits (MAB) problem in which a learner can compare a pair of choice alternatives that are called arms in the following in a sequential manner (see [1] or [2]). Instead of a numerical reward the learner can only observe a winner information in form of a binary feedback for each selected duel. As in the standard MAB problem, we assume this feedback to be stochastic. The goal of the learner is to identify the "best" arm as fast as possible.

Dueling Bandits algorithms were sucessfully applied to other settings in which the goal is to find the best among different choice alternatives by sequential comparisons. A classical example is the Algorithm Configuration (AC) setting in which we want to find the best parameter configuration for a given target algorithm by repeatedly racing two of them against each other. However, these target algorithms are often complex and need a long runtime until we can observe which parameter configuration performs best. All of the existing Dueling Bandits algorithms wait paitiently until the winner feedback is observed and do nothing in between. By this, a huge part of the overall runtime of the algorithm is wasted by just waiting for the observation. While there are some existing MAB methods that can deal with such a delayed feedback like [3], this setting is not studied until now for the Dueling Bandits case.

2. Problem Formulation

Assume that we have given a finite set of m choice alternatives that we denote by their indices $[m] = \{1, \dots, m\}$ and a finite number of allowed parallel duels $K \in \mathbb{N}_{>0}$. In each time step $t \in \mathbb{N}$ the learner has to choose a new duel $S_t = (i, j)$ for $i, j \in [m]$ if the number of active duels has not reached the number of allowed duels K yet. For each duel, the environment creates a pair $(\tau_t, \omega_t) \sim \mathbb{P}_{\tau,\omega}(\cdot|S_t)$ of a delay τ_t and winner ω_t of the selected duel. The learner can observe the winner $\omega_t = 1_{\{i \succ j\}}$ at time step $s = t + \tau_t$.

The goal is to identify the "best" arm as fast as possible, where fast as possible means the least wall-clock time here. The best arm ist defined as follows.

Definition 1 (Condorcet Winner) *We call arm $i^* \in [m]$ the Condorcet Winner (CW) iff $\mathbb{P}(i^* \succ j) > \mathbb{P}(j \succ i^*)$ for all $j \in [m]$ with $j \neq i^*$.*

3. Algorithm

To identify the CW in the Dueling Bandits problem with delayed feedback, we propose an algorithm that chooses the duels in each time step according to a trade-off of the expected information gain and the expected delay to observe the feedback. This is done by choosing the duel with the minimal ratio between the average observed delay for this duel and the gap between the winning probabilities of the arms. If this gap is huge, the chance is high that we can eliminate the worse arm soon. Let $\hat{\mathbb{P}}$ denote in the following the empirical probability, then we can solve the problem as given in the pseudo-code in algorithm 1.

For the theoretical derivation of the confidence bounds used in algorithm 2, we need the following assumptions.

```
Algorithm 1: Best arm identification in Dueling Bandits with Delay
Input: confidence $1 - \gamma$, set of arms $[m]$,
       number of parallel slots $K$
Output: the CW of $[m]$
Initialize remaining winner candidates $R_1 = [m]$,
  epoch $e = 1$
while $|R_e| > 1$ do
    Divide $R_e$ in duels $\{S_1, \dots, S_{|R_e|/2}\}$
    $E_e \leftarrow$ ELIMINATE($\{S_1, \dots, S_{|R_e|/2}\}$)
    $R_{e+1} \leftarrow R_e \backslash E_e$
    $e \leftarrow e + 1$
end
Return remaining arm in $R_e$
```

Assumption 2 *(1) A CW exists.*
(2) No ties in duels are allowed and we have for each duel set $\{i, j\} \subset [m]$ that $|\mathbb{P}(i \succ j) - \mathbb{P}(j \succ i)| \geq h$.
(3) The delays are upper bounded by $\tau \leq b$.

With this, we can derive the following confidence bounds which proofs are beyond the scope of this extended abstract.

```
Algorithm 2: Eliminate
Input: set of duels $\mathcal{S} = \{S_1, \dots, S_n\}$
Output: set $E$ of arms to eliminate
Initialize arms to eliminate $E = \emptyset$, active duels
  $A = \emptyset$, time step $t = 0$
while $E = \emptyset$ do
    for each set $S = \{i, j\} \in \mathcal{S}$ do
        estimate winning probability gap
          $\hat{\Delta}_t(S) = |\hat{\mathbb{P}}_t(i \succ j) - \hat{\mathbb{P}}_t(j \succ i)|$
        estimate average observed delay $\hat{\tau}_t(S)$
        estimate confidence bound $c_t(S)$
    end
    if number of active duels $|A| < K$ then
        Play duel $S_t = \operatorname{argmin}_{S \in \mathcal{S}} \frac{\hat{\tau}_t(S)}{\hat{\Delta}_t^2(S)} + c_t(S)$
        add $S_t$ to set of active duels $A = A \cup S_t$
    end
    for each set in active duels $S \in A$ do
        Possibly observe feedback $(\tau_S, w_S)$
        if feedback for $S$ is observed then
            remove from active duels $A = A \backslash S$
            Update arms to eliminate
              $E = E \cup \{i \in S | \hat{\Delta}_t(S) \geq c_t^\delta(S)\}$
        end
    end
end
```

Theorem 3 (Confidence bound) *For the confidence bound $c_t(S) = \sqrt{\frac{2b^2}{h^2 t} ln\left(\frac{t}{4}\right)}$, we have $\mathbb{P}\left(\left|\frac{\tau(S)}{\Delta^2(S)} - \frac{\hat{\tau}_t(S)}{\hat{\Delta}_t^2(S)}\right| \geq c_t(S)\right) \leq t^{-1}$.*

Theorem 4 (Eliminated arms) *For $c_t^\delta(S) = \max\left\{3h, \sqrt{-\frac{9t}{2} ln\left(\frac{\delta}{4}\right)}\right\}$, we eliminate a wrong arm only with probability δ in algorithm 2.*

4. Conclusion

We introduced the dueling bandits with delayed feedback problem and to the best of our knowledge are the first ones who study this problem. Our proposed algorithm for the best arm identification is guaranteed to only discard good arms with low probability.

Acknowledgments

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless

Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] 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.

[2] Róbert Busa-Fekete, Eyke Hüllermeier, and Adil El Mesaoudi-Paul. Preference-based online learning with dueling bandits: A survey. *CoRR*, abs/1807.11398, 2018.

[3] Tal Lancewicki, Shahar Segal, Tomer Koren, and Y. Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, 2021.