Proceedings of Machine Learning Research 95:740-755, 2018 ACML 2018

Efficient Mechanisms for Peer Grading and Dueling Bandits

Chuang-Chieh Lin JOSEPHCCLIN@GMAIL.COM
Chi-Jen Lu CJLU@IIS.SINICA.EDU.TW
Institute of Information Science, Academia Sinica, Taipei, Taiwan

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Many scenarios in our daily life require us to infer some ranking over items or people based on limited information. In this paper, we consider two such scenarios, one for ranking student papers in massive online open courses and one for identifying the best player (or team) in sports tournaments. For the peer grading problem, we design a mechanism with a new way of matching graders to papers. This allows us to aggregate partial rankings from graders into a global one, with an accuracy rate matching the best in previous works, but with a much simpler analysis. For the winner selection problem in sports tournaments, we cast it as the well-known dueling bandit problem and identify a new measure to minimize: the number of parallel rounds, as one normally would not like a large tournament to last too long. We provide mechanisms which can determine the optimal or an almost optimal player in a small number of parallel rounds and at the same time using a small number of competitions.

Keywords: Ordinal peer-grading; dueling bandits; Borda count

1. Introduction

Massive online open courses (MOOC), such as Coursera, edX, Academic Room, etc., have emerged as popular educational platforms which provide easy and free access to high quality education for students around the world. Yet a big issue in such courses, compared to traditional ones, is that it now becomes extremely costly to grade the papers from homeworks or exams due to the enormous number of students. A compromise solution, known as *peer-grading*, is to outsource the grading task to the students who participate in the course (Aziz et al., 2016; Caragiannis et al., 2015, 2016a; de Alfaro and Shavlovsky, 2014; Kulkarni et al., 2013; Kurokawa et al., 2015; Piech et al., 2013; Raman and Joachims, 2014; Shah et al., 2013; Walsh, 2014), and there have been several tools developed, such as `crowdgrader.org` (de Alfaro and Shavlovsky, 2014), `peergrading.org` (Raman and Joachims, 2014), `co-rank` (Caragiannis et al., 2016a), etc.

While grading by cardinal scores is common in traditional courses, it has some apparent problems in the case of peer grading. First of all, students are not professional graders, and the numerical scores they give may not be accurate. Moreover, they may have incentives to give others low scores as this may make their own scores relatively higher. Therefore, an alternative is for students to provide a ranking of the papers they are asked to grade. This is known as *ordinal peer-grading*. By asking each grader to rank a small number of papers (instead of giving scores to a large number of papers), it may mitigate the problem of graders deliberately misreporting their rankings, as discussed in previous works such

as (Caragiannis et al., 2015, 2016a,b). Ordinal peer-grading has recently drawn attention in the areas of AI and machine learning (Caragiannis et al., 2015, 2016a; Raman and Joachims, 2014; Shah et al., 2013).

Caragiannis et al. (2015) studied the following ordinal peer-grading framework. Suppose there is a total of n students, each having submitted a paper for a homework or exam. All these students are graders themselves, and each student is asked to grade a set, called bundle, of k papers not including her (or his) own. Moreover, each paper is required to be graded by exactly k students. The grading task of each student is to rank (order) the k papers of her bundle. From these n partial rankings, each on k papers, the goal is to produce a global ranking of the n papers which is as accurate as possible. Assuming that there exists some unknown *ground truth* of a global ranking of the n papers and each student can rank her bundle *perfectly* (consistent with the ground truth), Caragiannis et al. (2015) provided a mechanism which can recover an expected fraction of at least $1 - O(1/k)$ of the pairwise relations in the ground truth. Their mechanism is based on a fixed k-regular bipartite graph, one part for bundles and the other for graders, and the assignment of bundles to graders is randomized by randomly placing graders on nodes of the graph. However, this random process does not seem "random" enough, as there are some dependency issues which make their analysis rather complicated. In the following work, Caragiannis et al. (2016b) assumed an *infinite* number of students in order to avoid these dependency issues, which helps them obtain a cleaner analysis. In addition, they consider the imperfect grading scenario in which students can make mistakes when grading, and they conducted extensive simulations to demonstrate the empirical performance of their mechanisms.

Motivated by the interesting works of Caragiannis et al. (2015, 2016b), we would like to investigate further their ordinal peer-grading framework. In particular, we would like to understand if it is possible to have a simple mechanism which has theoretical guarantees with a simple analysis, but without the unrealistic assumption of an infinite number of students. Our first result in this paper is to answer this question affirmatively. We achieve this by modifying the mechanism of Caragiannis et al. (2015) to make the matching of graders to papers done in a more random way, so that the dependency issues due to their mechanism can be avoided. By designing a slightly different mechanism which breaks such dependencies, we are able to have a similar theoretical guarantee of recovering $1 - O(1/k)$ fraction of pairwise relations, but on the other hand using a much simpler analysis.

The peer grading problem can be seen as one type of ranking problem in which we need to produce a total ordering of all participants. On the other hand, there are natural scenarios in which we only need to select one winner. Consider for example a sports tournament in which there are many players (or teams) competing to win the final gold medal. We would like to have a good mechanism for competitions so that the gold medal is likely to be won by the strongest player. This can be seen as the task of the best arm identification for the multi-armed bandit problem, with each player corresponding to an arm and the strength of a player corresponding to the reward of the arm. However, it is not clear how to measure the strength of a player in an absolute scale, and the only way we can learn about a player's strength is through the competitions she (or he) participates in, which only reveals her strength relative to the opponents. There are also many ranking applications with a similar flavor. For example, it is usually much easier for a person to say which of two given movies he or she likes better, compared to giving accurate scores to a list of movies. To model

such a feedback setting, Yue et al. (2012) introduced the so-called *dueling bandit problem*, in which the information one can obtain about an arm is by comparing it with another arm, and the result is a one-bit value indicating which arm is better.

More formally, in the dueling bandit problem, there are n arms from which we can choose to play, with each arm a having some hidden strength characterized by some unknown parameter μ_a. The goal is to identify the optimal arm, which has the highest strength, or an ϵ-optimal arm, which has a strength within ϵ of the optimal one. However, we initially have no information at all about these hidden strengths. The only way we can obtain information about the strength of an arm is when we select it to compete with another arm, and the only information we have is which arm wins. We adopt the model that the result of each competition between arms a and b is a random event, with arm a winning this competition independently with probability exactly $\frac{1+(\mu_a-\mu_b)}{2}$. This models the noisy outcome inherent in many scenarios such as sports tournaments that a stronger player does not always win but has a better chance of winning.

While the dueling bandit problem has been studied before, previous works seem to focus on minimizing the total number of competitions (see for example Urvoy et al. (2013); Zoghi et al. (2014)) or minimizing some notion of regret (see for example Ailon et al. (2014); Yue et al. (2012); Zoghi et al. (2014)). However, in scenarios such as sports tournaments, there are other measures which one may also want to minimize. One measure is the number of parallel rounds. We care about this measure because when we have a large tournament with many participants, we naturally would not want it to last too long. Therefore, if possible, it is preferable to have competitions done in a parallel way. Note that a set of competitions can be done in parallel in a round if no arm is involved in more than one competition. Related to this measure is the measure on the number of competitions each arm has to participate. We care about this because in scenarios such as sports competition, playing many more competitions may not be fair for a player as it is likely to wear her down and skew the outcomes. Note that this measure is upper-bounded by the number of rounds, so it suffices to aim for minimizing the latter. As the total number of competitions is also a natural and important measure, we would also like to minimize it in addition to the number of rounds.

Our second contribution is a mechanism for this dueling bandit problem which is efficient in both measures. More precisely, let $\Delta_i = \mu_{i^*} - \mu_i$ denote the gap of strengths between an arm i and the optimal arm i^*, and let $\Delta = \min_{i \neq i^*} \Delta_i$ denote the minimum gap. Then given any $\delta > 0$, our mechanism with probability $1 - \delta$ can identify the optimal arm in $O((1/\Delta^2)\log(n/(\Delta\delta)))$ rounds via a total of at most $O(\sum_i (1/\Delta_i^2)\log(n/(\Delta_i\delta)))$ competitions. Moreover, if we only want to identify an arm in the set of ϵ-optimal arms, denoted as OPT_ϵ, for some $\epsilon > \Delta$, then our mechanism needs at most $O((1/\epsilon^2)\log(n/(\epsilon\delta)))$ rounds and at most $O(\sum_{i \notin \mathrm{OPT}_\epsilon}(1/\Delta_i^2)\log(n/(\Delta_i\delta)) + |\mathrm{OPT}_\epsilon|(1/\epsilon^2)\log(n/(\epsilon\delta)))$ competitions. This will be described in Section 3.

Although the two problems we consider may appear unrelated and have usually been considered separately, we place them in the same framework for study. That is, each can be seen as some ranking problem: to determine a global ranking or a winner from a collection of partial rankings. In fact, our algorithm for the dueling bandit problem is inspired by our algorithm for the peer grading problem, and both algorithms share a similar structure. Thus, we believe that the conceptual contribution of our paper is to provide a connection

between these two seemingly unrelated topics, and we hope that this connection could further enable transfer of ideas between these two topics. In our future work, we plan to consider the more challenging case of imperfect graders. Although this has been considered in (Caragiannis et al., 2015, 2016b), only empirical results were given. We hope to provide theoretical guarantee for this case as well, and we will also conduct some experiments to gain a deeper understanding.

Recall that in the dueling bandit setting discussed above, each arm is assumed to have an underlying strength, and a stronger arm has a higher probability of winning a weaker arm. This gives a natural order among the arms, according to their strengths, so that an obvious winner one would like to identify is the arm with the highest strength. However, this may not always be true. For example, in sports, it is possible that some player i is more likely to win some player j who in turn is more likely to win another player k, while at the same time player k is more likely to win player i. In this case, it is not obvious which player should be considered "optimal" and which player should we try to identify.

Our final contribution is to provide such a mechanism to identify the Borda winner or an ϵ-optimal one. More precisely, now we measure the strength of an arm in terms of its Borda score defined as the probability that it wins a randomly chosen opponent, and an ϵ-optimal Borda winner is an arm with its strength within ϵ of the optimal one. Then given any $\epsilon, \delta > 0$, our mechanism with probability $1 - \delta$ can identify an ϵ-Borda winner in $O((1/\epsilon^2)\log(n/(\epsilon\delta)))$ rounds via $O((n/\epsilon^2)\log(n/(\epsilon\delta)))$ competitions. This also gives us bounds for identifying the Borda winner as it is just an ϵ-optimal Borda winner with ϵ equal to the gap between the two highest strengths. Note that the number of rounds needed here is of the same order as that in our second result, while the total number of competitions now is slightly higher in general. This part of result will be shown in Section 4.

2. Peer Grading

In this section, we consider the ordinal peer-grading problem discussed in the introduction. Formally, there are n papers submitted by n students, respectively, for a homework assignment, and we index these n students as well as their papers by $1, \ldots, n$. The goal is to provide a good ranking (ordering) for these n papers, with better papers ranked before (having smaller ranks than) worse papers. These n students are also graders themselves, and the n papers will be graded by them under the following constraints, for some parameter k:

- Each paper is graded by exactly k graders,
- each grader grades exactly k papers, and
- no grader grades her (or his) own paper.

To simplify our presentation, let us assume that k divides n (it is easy to check that our result can be easily extended to the case that k does not divide n). Following (Caragiannis et al., 2015), we ask each grader to grade the k papers by providing a ranking of them. To make the load of each grader light, we will only consider $k < \sqrt{n}$. As in (Caragiannis et al., 2015), we also make the assumption that there is a true ranking among the n papers, and each grader is *perfect* in the sense that her ranking of the k papers is consistent with the true ranking. For convenience, let us assume without loss of generality that the true rank

of a paper corresponds exactly to the index of the paper, so that paper r is better than (has a smaller rank) paper q if and only if $r < q$. Our task is to design a mechanism which assigns papers to graders under the constraints above, and then takes the partial rankings provided by the graders to produce a global ranking of the n papers, hopefully with a small error. Following (Caragiannis et al., 2015), we measure the error by the expected fraction of pairwise relations between papers incorrectly ranked according to our ranking.

Our mechanism works as follows. It proceeds in k rounds, with the following steps taken in each round t:

1. Select an arbitrary set of n/k students who have not done the grading before as the graders in this round.

2. Randomly partition the n papers into n/k disjoint sets, which we call *bundles*, each having k papers.

3. Match each grader of this round to a bundle to grade such that each grader grades one bundle which does not contain her (or his) own paper and each bundle is graded by exactly one grader. Such a matching can be found efficiently, as guaranteed by Lemma 1 in the following.

4. Each grader is asked to give a ranking of the papers in her bundle, by giving a different score from 0 to $k-1$ to each paper in the bundle, with a higher score to a better paper.

After these k rounds, we compute the total score of each paper by taking the sum of the scores it receives in the k rounds, and we output the ranking of the n papers according to their total scores, in the natural way that papers with higher scores are ranked before (have smaller ranks than) those with lower scores.

The following lemma guarantees that the matching in Step 3 of our mechanism can be found efficiently, which we prove in Subsection 2.1.

Lemma 1 *There is an efficient algorithm for finding the matching in Step 3 of our mechanism when $k^2 < n$.*

Note that the condition $k^2 < n$ naturally holds in the peer-grading scenario as the number n of students is usually very large while we would like to give each of them only a small number k of papers to grade. The performance of our mechanism is guaranteed by the following theorem.

Theorem 2 *The expected fraction of pairwise relations incorrectly ranked by our mechanism is at most $O(1/k)$.*

Note that this $O(1/k)$ upper bound matches the best bound in (Caragiannis et al., 2015) which is achieved when their mechanism is based on a k-regular bipartite graph of girth (the length of a shortest cycle) at least 6. On the other hand, our analysis is simpler as we will see later.

Before proving the theorem, let us introduce some notations first. For a paper r and a round t, we let $S_t(r)$ denote the score the paper receives in round t, which is a random

variable depending on the random bundle it belongs to. As the random partitioning into bundles in each round is done independently from other rounds, the k random variables $S_1(r), \ldots, S_k(r)$ are mutually independent from each other. In fact, all random variables in a round are independent from random variables in all other rounds. This allows us to avoid some dependency issues arising from the mechanism of Caragiannis et al. (2015) which complicate their analysis.

To prove the theorem, let us first consider any two papers r and q, with $r < q$. Let $W_t(r,q) = S_t(r) - S_t(q)$, which is their score difference in round t, and let $W(r,q) = \sum_{t=1}^{k} W_t(r,q)$, which is the difference between their total scores. Recall that we use the convention that paper r is better than paper q, given $r < q$, and hence we would like paper r to receive a higher score such that $W(r,q) > 0$. Our goal then is to show that this holds with high probability, or equivalently, $\Pr[W(r,q) \leq 0]$ is small. Note that $W(r,q)$ is the summation of k i.i.d. random variables, so we would like to apply some tail bound for such random variables. However, as each random variable $W_t(r,q)$ may take a value in a large range, from $-(k-1)$ to $k-1$, Hoeffding's inequality does not seem to give us a good enough bound. Therefore, we turn to the following tail bound, known as Bernstein's inequality (Bernstein, 1946).

Lemma 3 *Let $X_1, \ldots, X_k$ be a sequence of k independent random variables such that each variable X_t has mean 0, variance σ^2, and $|X_t| \leq b$ almost surely. Then there exists some constant $c_0 > 0$ such that for any $V > 0$,*

$$\Pr\left[\sum_{t=1}^{k} X_t \geq V\right] \leq \exp\left(-\frac{c_0 V^2}{k\sigma^2 + bV}\right).$$

Before being able to apply the lemma, we need to bound the expectation and the variance of our random variable $W_t(r,q)$. This is provided by the following lemma, which we will prove in Subsection 2.2.

Lemma 4 *Fix any $r < q$ and any t. The random variable $W_t(r,q)$ has expected value $\mu = (k-1)(q-r)/(n-1) = \Theta(k(q-r)/n)$ and variance $\sigma^2 \leq O(k+\mu^2)$.*

Then to apply Lemma 3 to upper-bound $\Pr[W(r,q) \leq 0]$, we let $X_t = \mu - W_t(r,q)$, with $\mu = \mathrm{E}[W_t(r,q)]$, and $V = k\mu$. Note that $|X_t| \leq k$ and the variance of X_t is $\sigma^2 = \mathrm{E}[(\mu - W_t(r,q))^2]$, which equals the variance of $W_t(r,q)$ and is at most $O(k+\mu^2)$. As a result, for any $r < q$, we have

$$\begin{aligned}
\Pr\left[W(r,q) \leq 0\right] &= \Pr\left[\sum_{t=1}^{k} (\mu - W_t(r,q)) \geq k\mu\right] \\
&\leq \exp\left(-\frac{c_0 k^2 \mu^2}{k^2 + k\mu^2 + k^2\mu}\right) \\
&\leq \exp\left(-c_0\mu^2/2\right) + \exp\left(-c_0\mu/4\right),
\end{aligned}$$

using the fact that $\mu \leq k$ and $\exp(-\alpha/(\beta+\gamma)) \leq \max\{\exp(-\beta/(2\beta)), \exp(-\alpha/(2\gamma))\}$ for positive α, β, γ. Since $\mu = \Theta(k(q-r)/n)$ by Lemma 4, the expected number of pairwise

relations incorrectly ranked by our mechanism, which is $\sum_{r=1}^{n}\sum_{q=r+1}^{n}\Pr\left[W(r,q)\le 0\right]$, can be upper-bounded by

$$\sum_{r=1}^{n}\sum_{q=r+1}^{n}\left(\exp\left(-\frac{ck^2(q-r)^2}{n^2}\right)+\exp\left(-\frac{ck(q-r)}{n}\right)\right),$$

for some constant $c>0$, which is at most

$$n\cdot\sum_{d=1}^{n-1}\left(e^{-ck^2\left(\frac{d}{n}\right)^2}+e^{-ck\left(\frac{d}{n}\right)}\right)\le n^2\cdot\int_0^1\left(e^{-ck^2x^2}+e^{-ckx}\right)dx$$

with $d=q-r$ and $x=d/n$. Finally, using the well-known bound on the Gaussian integral that $\frac{2}{\sqrt{\pi}}\int_0^\infty e^{-t^2}dt\le 1$, we can upper-bound the righthand side above by

$$n^2\cdot\left(\frac{1}{\sqrt{ck}}\cdot\frac{\sqrt{\pi}}{2}-\frac{1}{ck}\cdot\left(1-e^{-ck}\right)\right)\le\binom{n}{2}\cdot O\left(\frac{1}{k}\right).$$

This shows that the expected fraction of incorrect pairwise relations ranked by our mechanism is at most $O(1/k)$, which proves Theorem 2.

2.1. Proof of Lemma 1

Let us consider the graders one by one and match each grader to a remaining bundle that does not contain her own paper. The only possible grader that we may fail to find a matched bundle this way is the last one because each grader is allowed to grade all but one bundle and thus can find a matched bundle as long as there are at least two remaining. In case the last grader, denoted as g, indeed cannot grade the remaining bundle, denoted as b, we can simply choose a previous grader, denoted as h, who can grade b, take h's bundle away and have it rematched to g, and then rematch h to the bundle b. Note that such a grader h can always be found, with $n>k^2$, because the bundle b has size k while there are $n/k>k$ graders in each round.

2.2. Proof of Lemma 4

Fix any $q>r$ and any t. Recall that $W_t(r,q)=S_t(r)-S_t(q)$, where $S_t(r)$ and $S_t(q)$ are the scores paper r and q receive respectively in round t.

Let us first bound $\mathrm{E}[W_t(r,q)]=\mathrm{E}[S_t(r)]-\mathrm{E}[S_t(q)]$. For $\mathrm{E}[S_t(r)]$, note that the random process of forming the bundle containing r is equivalent to the random process of sampling the other $k-1$ papers one by one without replacement. From this equivalent random process, let us define the binary random variables $V_1,\ldots,V_{k-1}$ such that $V_i=1$ if the i'th sampled paper has its true rank worse (larger) than paper r's and $V_i=0$ otherwise. Then we have

$$\mathrm{E}\left[S_t(r)\right]=\mathrm{E}\left[\sum_{i=1}^{k-1}V_i\right]=\sum_{i=1}^{k-1}\mathrm{E}\left[V_i\right]=(k-1)\cdot\frac{n-r}{n-1}.$$

Similarly, one can also show that $\mathrm{E}[S_t(q)]=(k-1)\cdot\frac{n-q}{n-1}$, and as a result, we have

$$\mathrm{E}\left[W_t(r,q)\right]=(k-1)\cdot\frac{(n-r)-(n-q)}{n-1}=(k-1)\cdot\frac{q-r}{n-1}.$$

Next, let us bound the variance of $W_t(r,q)$. To simplify our notation, let us write W for $W_t(r,q)$. From the definition, the variance of W is $\mathrm{E}\left[W^2\right] - (\mathrm{E}\left[W\right])^2 \leq \mathrm{E}\left[W^2\right]$. To bound $\mathrm{E}\left[W^2\right]$, we consider two cases, according to whether or not the following event happens:

- Event A: papers r and q belong to the same bundle.

For the first case that event A happens, let us consider the equivalent random process of sequentially sampling the other $k-2$ papers without replacement to form the bundle with r and q. From this random process, let us define the binary random variables $D_1, \ldots, D_{k-2}$, with $D_i = 1$ if the i'th sampled paper has its true rank fall between r and q while $D_i = 0$ otherwise. Then we have

$$\mathrm{E}\left[W^2 \mid A\right] = \mathrm{E}\left[\left(\sum_{i=1}^{k-2} D_i\right)^2\right],$$

which can be expanded into

$$\sum_{i=1}^{k-2} \mathrm{E}\left[D_i^2\right] + 2\sum_{i=1}^{k-2}\sum_{j=i+1}^{k-2} \mathrm{E}\left[D_i D_j\right].$$

We simply upper-bound the first sum by $k-2$ as $D_i^2 \leq 1$ for each i. To bound the second sum above, note that for any $i \neq j$,

$$\mathrm{E}\left[D_i D_j\right] = \frac{q-r}{n-2} \cdot \frac{q-r-1}{n-3} \leq O\left(\left(\frac{q-r}{n}\right)^2\right).$$

As a result, we have

$$\mathrm{E}\left[W^2 \mid A\right] \leq O\left(k + \left(\frac{k(q-r)}{n}\right)^2\right). \tag{1}$$

For the second case that event A does not happen, there are two different bundles involved for determining the random variable W, one for paper r and one for paper q. Thus, let us consider the following equivalent random process: sequentially sampling the $k-1$ papers without replacement to form the bundle with r and then sequentially sampling other $k-1$ papers without replacement to form the bundle with q. From this, let us define the binary random variables $Y_1, \ldots, Y_{k-1}$, with $Y_i = 1$ if and only if the i'th sampled paper in the bundle of r has a rank worse (higher) than r's, as well as the binary random variables $Z_1, \ldots, Z_{k-1}$, with $Z_i = 1$ if and only if the i'th sampled paper in the bundle with q has a rank worse than q's. Then we have

$$\mathrm{E}\left[W^2 \mid \neg A\right] = \mathrm{E}\left[\left(\sum_{i=1}^{k-1} (Y_i - Z_i)\right)^2\right],$$

which can be expanded into

$$\sum_{i=1}^{k-1} \mathrm{E}\left[(Y_i - Z_i)^2\right] + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k-1} \mathrm{E}\left[(Y_i - Z_i)(Y_j - Z_j)\right].$$

We simply upper-bound the first sum by $k-1$ as $(Y_i - Z_i)^2 \leq 1$ for each i. To bound the second sum above, note that for any $i \neq j$, $\mathrm{E}\left[(Y_i - Z_i)(Y_j - Z_j)\right] = \mathrm{E}\left[Y_i Y_j\right] + \mathrm{E}\left[Z_i Z_j\right] - \mathrm{E}\left[Y_i Z_j\right] - \mathrm{E}\left[Y_j Z_i\right]$, which equals

$$\frac{n-r-1}{n-2} \cdot \frac{n-r-2}{n-3} + \frac{n-q}{n-2} \cdot \frac{n-q-1}{n-3} - \frac{n-q}{n-2} \cdot \frac{n-r-2}{n-3} \cdot 2,$$

and a simple calculation shows that it is at most $O((\frac{q-r}{n})^2)$. As a result, we have

$$\mathrm{E}\left[W^2 \mid \neg A\right] \leq O\left(k + \left(\frac{k(q-r)}{n}\right)^2\right). \tag{2}$$

Finally, since

$$\mathrm{E}\left[W^2\right] = \Pr[A] \cdot \mathrm{E}\left[W^2 \mid A\right] + \Pr[\neg A] \cdot \mathrm{E}\left[W^2 \mid \neg A\right],$$

Lemma 4 follows from the bounds in (1) and (2).

3. Round-Efficient Dueling Bandits

In this section, we consider the task of determining the best player (team) in a large tournament via pairwise competitions, discussed in the introduction.

We model this task as the following dueling bandit problem, with each arm corresponding to a player. Formally, there are n arms, numbered from 1 to n, with each arm a characterized by some unknown strength parameter $\mu_a \in [0,1]$. The goal is to identify the optimal arm, denoted as i^*, which has the highest strength parameter $i^* = \arg\max_a \mu_a$, or a relaxed ϵ-optimal arm in the set $\mathrm{OPT}_\epsilon = \{a : \mu_a \geq \mu_{i^*} - \epsilon\}$. However, the only information one can learn about these unknown strength parameters is through pairwise competitions between arms. The information is limited in the way that each time when arms a and b compete, we only know a binary outcome of which arm wins. Moreover, the information is noisy in the sense that each time when arms a and b compete, arm a wins independently with probability $\frac{1+\mu_a-\mu_b}{2}$.

While the dueling bandit problem has been studied before, previous works mostly focused on minimizing the total number of pairwise competitions (or on a different objective of minimizing regret). As we are interested in scenarios such as sports tournaments, there are additional objectives which we would like to optimize. First of all, for a large tournament with many players, one naturally would not like the tournament to last too long. Therefore, a measure we would like to minimize is the number of rounds, where a set of pairwise competitions can be carried out concurrently in a round if each arm is involved in at most one competition in the set. A related objective is to minimize the number of competitions any arm has to participate, as each competition may induce some cost to (e.g. take some energy from) a participant, or to make sure no arm competes in many more competitions than others as it may be unfair to the exhausted arm (player or team). In the following, we will focus on minimizing the number of rounds, in addition to minimizing the total number of pairwise competitions.

Note that when not concerned about the number of rounds as in previous works, one can reduce the dueling bandit problem to the traditional multi-armed bandit by letting each

arm compete with a fixed benchmark arm b, and use the winning probability of an arm as the mean reward in the multi-armed bandit setting. In this way, better arms in the dueling bandit setting remain better in the multi-armed bandit setting, as the winning probability of a stronger arm a, which is $\frac{1+\mu_a-\mu_b}{2}$, is higher than that of a weaker arm a', which is $\frac{1+\mu_{a'}-\mu_b}{2}$, with $\mu_a > \mu_{a'}$. However, in doing so, that arm b has to participate in many more competitions than others and this also makes the competitions highly sequential requiring many rounds.

Our mechanism instead is based on our mechanism in the previous section, using bundles of size two. Note that now there are no graders involved so that we no longer have those grading constraints to satisfy. On the other hand, as now we also want to minimize the total number of competitions, we would like to eliminate arms from future competitions when we are confident that they are suboptimal, following previous works on (dueling) bandits. However, this makes things complicated as in each round, the expected score of a remaining arm now becomes dependent on the randomness of previous outcomes, and moreover, it actually changes with rounds.

Our mechanism works as follows. Given a parameter $\delta > 0$, which is the allowed error probability, it again proceeds in rounds, but now it keeps eliminating arms which are unlikely to be the optimal one, until there is only one arm left. In each round t, it maintains and updates a set A_t of plausible arms, with $n_t = |A_t|$, and takes the following steps:

1. Randomly partition the arms in A_t into disjoint pairs and let each pair of arms compete with each other, except for one unpaired arm when n_t is odd.

2. For any arm $a \in A_t$, let $S_t(a) = 1$ if it is paired in a competition and wins it, and let $S_t(a) = 0$ otherwise. From each $S_t(a)$, we compute the scaled score

$$X_t(a) = \gamma_t \cdot S_t(a), \text{ with } \gamma_t = \begin{cases} 2\frac{n_t-1}{n_t} & \text{if } n_t \text{ is even,} \\ 2 & \text{otherwise,} \end{cases}$$

   as well as the average score $\bar{X}_t(a) = \frac{1}{t}\sum_{\tau=1}^{t} X_\tau(a)$.

3. Let $i_t^* = \arg\max_{a \in A_t} \bar{X}_t(a)$ be the empirically best arm so far, and use it to update A_{t+1} as

$$A_{t+1} = A_t \setminus \left\{ a \in A_t : \bar{X}_t(a) < \bar{X}_t(i_t^*) - C_t(\delta) \right\}$$

   which excludes any implausible arm with an empirical average less than that of i_t^* by the amount

$$C_t(\delta) = \sqrt{(18/t)\ln(2nt^2/\delta)}.$$

The performance of our mechanism is guaranteed by the following theorem, which bounds the number of rounds and pairwise competitions in terms of the gaps of arms, defined as $\Delta_a = \mu_{i^*} - \mu_a$ for each arm a, and $\Delta = \min_{a \neq i^*} \Delta_a$.

Theorem 5 *Given any $\delta > 0$, our mechanism with probability at least $1 - \delta$ can identify the optimal arm in*

$$O\left(\frac{1}{\Delta^2}\log\frac{n}{\Delta\delta}\right)$$

rounds, with the total number of competitions at most

$$O\left(\sum_a \frac{1}{\Delta_a^2}\log\frac{n}{\Delta_a\delta}\right).$$

To prove the theorem, we start by showing that the scaled score $X_t(a)$ in each round t provides a meaningful measure for each arm a. Formally, we have the following.

Lemma 6 *For any round t, any history before round t, and any two surviving arms $a, b \in A_t$, we have $\mathrm{E}_t[X_t(a) - X_t(b)] = \mu_a - \mu_b$, where $\mathrm{E}_t[\cdot]$ denotes the expectation conditioned on the history.*

This shows that better remaining arms indeed have higher expected scores. We will prove the lemma in Subsection 3.1.

Next, we show that with high probability, the best arm i^* is never eliminated. For this, we would also like to apply some standard tail bound to show that for any round t, it is unlikely to have $\bar{X}_t(i^*)$ significantly smaller than $\bar{X}_t(i)$ of some arm i (as i_t^*). However, although for any arm a and round t, $\bar{X}_t(a)$ is the average of t random variables $X_1(a), \ldots, X_t(a)$, these random variables are not mutually independent from each other. In fact, for any $\tau > 1$, the random variable $X_\tau(a)$ depends on previous random variables $X_1(a), \ldots, X_{\tau-1}(a)$. Moreover, the conditional mean $\mathrm{E}_t[X_\tau(a)]$ actually changes with round τ and depends also on previous random variables.

To handles these issues, we introduce new random variables $Y_t(a)$ and $Z_t(a)$, for any round t and given $a, i^* \in A_t$, defined as $Y_t(a) = X_t(a) - X_t(i^*) + \Delta_a$ and $Z_t(a) = \sum_{\tau=1}^{t} Y_\tau(a)$. Since for any τ and a, $|Y_\tau(a)| \le 2 + 1 = 3$ and by Lemma 6, $\mathrm{E}_t[Y_\tau(a)] = \mu_a - \mu_{i^*} + \Delta_a = 0$, the sequence $Z_1(a), \ldots, Z_t(a)$ of random variables form a martingale, given $i^* \in A_t$, which allow us to apply Azuma's inequality. Formally, we have the following lemma, which we will prove in Subsection 3.2.

Lemma 7 *The probability that the optimal arm is ever eliminated by our mechanism is at most $\delta/2$.*

Now we bound the number of rounds a suboptimal arm $a \neq i^*$ with gap Δ_a can survive. Note that for any t,

$$\begin{aligned}\Pr\left[a, i^* \in A_{t+1}\right] &\le \Pr\left[\bar{X}_t(a) > \bar{X}_t(i_t^*) - C_t(\delta) \text{ and } i^* \in A_t\right] \\ &\le \Pr\left[\bar{X}_t(a) - \bar{X}_t(i^*) > -C_t(\delta) \text{ and } i^* \in A_t\right],\end{aligned}$$

as $\bar{X}_t(i_t^*) \ge \bar{X}_t(i^*)$ by the definition of i_t^*. The last probability above by definition equals

$$\begin{aligned}&\Pr\left[\frac{1}{t}\sum_{\tau=1}^{t}\left(Y_\tau(a) - \Delta_a\right) > -C_t(\delta) \text{ and } i^* \in A_t\right] \\ =\ &\Pr\left[Z_t(a) > t\cdot\left(\Delta_a - C_t(\delta)\right) \text{ and } i^* \in A_t\right] \\ \le\ &\exp\left(-\frac{\left(t\left(\Delta_a - C_t(\delta)\right)\right)^2}{2\cdot t\cdot 3^2}\right),\end{aligned} \tag{3}$$

by Azuma's inequality. Thus, when $t \geq (72/\Delta_a^2)\ln(2nt^2/\delta)$ so that $C_t(\delta) \leq \Delta_a/2$, we can upper-bound (3) by

$$\exp\left(-\frac{t\Delta_a^2}{72}\right) \leq \frac{\delta}{2nt^2} \leq \frac{\delta}{2n}.$$

This implies that whenever $t \geq t_a$ for some $t_a = O((1/\Delta_a^2)\log(n/(\Delta_a\delta))$, the probability above is at most $\delta/(2n)$. Then by a union bound, we have

$$\Pr\left[\exists t : i^* \notin A_t \text{ or } \exists a \neq i^* : a \in A_{t_a+1}\right] \quad \leq \quad \Pr\left[\exists t : i^* \notin A_t\right] + \sum_{a \neq i^*} \Pr\left[a, i^* \in A_{t_a+1}\right]$$

which is at most $\frac{\delta}{2} + n \cdot \frac{\delta}{2n} = \delta$. This proves Theorem 5.

From the analysis above, it is easy to see that given any $\epsilon > \Delta$, if we only want to identify an ϵ-optimal arm $a \in \mathrm{OPT}_\epsilon$, we can stop our mechanism earlier, in $O(\frac{1}{\epsilon^2}\log\frac{n}{\epsilon\delta})$ rounds and return any remaining arm. Thus, we have the following.

Corollary 8 *Given any $\epsilon, \delta > 0$, our mechanism with probability at least $1-\delta$ can identify an ϵ-optimal arm in*

$$O\left(\frac{1}{\epsilon^2}\log\frac{n}{\epsilon\delta}\right)$$

rounds, and the total number of competitions is at most

$$O\left(\sum_{a \notin \mathrm{OPT}_\epsilon} \frac{1}{\Delta_a^2}\log\frac{n}{\Delta_a\delta} + \frac{|\mathrm{OPT}_\epsilon|}{\epsilon^2}\log\frac{n}{\epsilon\delta}\right).$$

3.1. Proof of Lemma 6

Fix any t and any previous history before round t. Recall that $n_t = |A_t|$. By definition, we have

$$\mathrm{E}_t[X_t(a) - X_t(b)] = \gamma_t \cdot (\mathrm{E}_t[S_t(a) - S_t(b)]), \tag{4}$$

where $\gamma_t = 2\frac{n_t-1}{n_t}$ if n_t is even and $\gamma_t = 2$ if n_t is odd. We claim that (4) equals $\frac{2}{n_t} \cdot \alpha_t$ with

$$\alpha_t \equiv \sum_{i \in A_t \setminus \{a\}} \frac{1+\mu_a-\mu_i}{2} - \sum_{i \in A_t \setminus \{b\}} \frac{1+\mu_b-\mu_i}{2}.$$

To see this, note that when n_t is even, (4) equals

$$\frac{2(n_t-1)}{n_t} \cdot \frac{1}{n_t-1} \cdot \alpha_t = \frac{2}{n_t} \cdot \alpha_t,$$

and when n_t is odd, (4) equals

$$2 \cdot \left(1 - \frac{1}{n_t}\right) \cdot \frac{1}{n_t-1} \cdot \alpha_t = \frac{2}{n_t} \cdot \alpha_t,$$

with the $(1 - \frac{1}{n_t})$ factor is the probability that a given arm is paired with a component to compete. As a result, we have $\mathrm{E}_t[X_t(a) - X_t(b)]$ equal to

$$\begin{aligned}\frac{2}{n_t} \cdot \alpha_t &= \frac{2}{n_t} \cdot \left(\sum_{i \in A_t \setminus \{a\}} \frac{\mu_a - \mu_i}{2} - \sum_{i \in A_t \setminus \{b\}} \frac{\mu_b - \mu_i}{2} \right) \\ &= \frac{1}{n_t} \cdot \left(\sum_{i \in A_t} (\mu_a - \mu_i) - \sum_{i \in A_t} (\mu_b - \mu_i) \right) \\ &= \mu_a - \mu_b.\end{aligned}$$

3.2. Proof of Lemma 7

Recall that for any round t and given $i^*, a \in A_t$, we have the random variables

$$Y_t(a) = X_t(a) - X_t(i^*) + \Delta_a \text{ and } Z_t(a) = \sum_{\tau=1}^{t} Y_\tau(a),$$

and observe that

$$\bar{X}_t(a) - \bar{X}_t(i^*) = \frac{1}{t} \sum_{\tau=1}^{t} (X_\tau(a) - X_t(i^*)) = \frac{1}{t} Z_t(a) - \Delta_a.$$

Moreover, as discussed before, the sequence of random variables $Z_1(a), \ldots, Z_t(a)$ form a martingale. Therefore, with the choice of $C_t(\delta) = \sqrt{(18/t) \ln(2nt^2/\delta)}$, we can bound the probability that arm i^* is ever eliminated by

$$\begin{aligned}&\Pr\left[\exists t \exists a \in A_t : \bar{X}_t(i^*) < \bar{X}_t(a) - C_t(\delta) \text{ and } i^* \in A_t\right] \\ &= \Pr\left[\exists t \exists a \in A_t : \bar{X}_t(a) - \bar{X}_t(i^*) > C_t(\delta) \text{ and } i^* \in A_t\right] \\ &= \Pr\left[\exists t \exists a \in A_t : \frac{1}{t} Z_t(a) - \Delta_a > C_t(\delta) \text{ and } i^* \in A_t\right] \\ &\leq \sum_t \sum_{a \in A_t} \Pr\left[Z_t(a) > t \cdot C_t(\delta) \text{ and } i^* \in A_t\right],\end{aligned}$$

which by Azuma's inequality is at most

$$\sum_t n \cdot \exp\left(-\frac{(t \cdot C_t(\delta))^2}{2 \cdot t \cdot 3^2} \right) \leq \sum_t n \cdot \frac{\delta}{2nt^2} \leq \frac{\delta}{2}.$$

This proves the lemma.

4. Borda Winner Identification

Recall that in the previous section, we consider the setting in which each arm a has an underlying strength μ_a. This means that the arms can be arranged in a total order, according to their strengths, so that an arm with a higher strength has a higher probability of winning

an arm with a lower strength. In this section, we consider a relaxed setting in which no such total ordering exists.

Formally, there are again n arms, and they are associated with some $n \times n$ matrix P. The (i, j)-entry of the matrix P, denoted as $P(i, j)$, indicates the probability that arm i wins arm j each time they compete with each other. We do not place any assumption on these n^2 entries of P, except that they take values from the interval $[0, 1]$. Therefore, we allow the possibility that arm i is more likely to win arm j and arm j is more likely to win arm k, while at the same time arm k is more likely to win arm i. In this case, it is not obvious which arm should be considered "optimal" and which arm should we try to identify. Nevertheless, a reasonable choice is the so-called Borda winner (Jamieson et al., 2015), which is the arm with the highest probability of winning another randomly chosen arm.

Formally, we define the *Borda score* of an arm a as

$$B(a) = \frac{1}{n-1} \sum_{i \neq a} P(a, i).$$

Then the *Borda winner* is the arm $a^* = \arg\max_a B(a)$, which has the highest Borda score, and we now consider gaps of arms according to the Borda scores, defined as

$$\Delta_a = B(a^*) - B(a) \text{ for any arm } a, \text{ and } \Delta = \min_{a \in i^*} \Delta_a.$$

As before, we are also interested in finding an ϵ-optimal Borda winner in the set

$$\mathrm{OPT}_\epsilon = \{a : B(a) \geq B(i^*) - \epsilon\}.$$

For the task of identifying the Borda winner or an ϵ-optimal Borda winner, in the dueling bandit setting, we can take our mechanism in the previous section, but make the change of keeping all the arms for competition in every round and using the score

$$X_t(a) = \begin{cases} S_t(a) & \text{if } n \text{ is even,} \\ \frac{n}{n-1} \cdot S_t(a) & \text{otherwise.} \end{cases}$$

Then we claim that for any t and a, $\mathrm{E}[X_t(a)] = B(a)$. To see this, note that when n is even,

$$\mathrm{E}[X_t(a)] = \mathrm{E}[S_t(a)] = \frac{1}{n-1} \sum_{i \neq a} P(a, i) = B(a),$$

and when n is odd, $\mathrm{E}[X_t(a)]$ equals

$$\frac{n}{n-1} \cdot \mathrm{E}[S_t(a)] = \frac{n}{n-1} \cdot \left(1 - \frac{1}{n}\right) \frac{1}{n-1} \sum_{i \neq a} P(a, i) = B(a).$$

Moreover, since we now do not eliminate arms, the random variables $X_1(a), \ldots, X_t(a)$ become mutually independent. Now as each variable has $|X_\tau(a)| \leq 1$, we can simply apply Hoeffding's inequality, with $C_t(\delta) = \sqrt{(1/t)\ln(2nt^2/\delta)}$, and have

$$\Pr\left[\left|\bar{X}_t(a) - B(a)\right| > C_t(\delta)\right] = \Pr\left[\left|\frac{1}{t}\sum_{\tau=1}^{t}\left(X_\tau(a) - \mathrm{E}[X_\tau(a)]\right)\right| > C_t(\delta)\right] \leq \frac{\delta}{nt^2},$$

for any t and a. This implies that given any $\epsilon > 0$, there exists some $t_\epsilon \leq O(\frac{1}{\epsilon^2} \log \frac{n}{\epsilon\delta})$ such that when $t \geq t_\epsilon$, we have $C_t(\delta) \leq \frac{\epsilon}{2}$ and

$$\Pr\left[\exists a \notin \mathrm{OPT}_\epsilon : \bar{X}_t(a) > \bar{X}_t(i^*)\right] \leq \Pr\left[\exists a : \left|\bar{X}_t(a) - B(a)\right| > \frac{\epsilon}{2}\right] \leq n \cdot \frac{\delta}{nt^2},$$

by a union bound. From this, we can conclude that

$$\Pr\left[\exists t \geq t_\epsilon, \exists a \notin \mathrm{OPT}_\epsilon : \bar{X}_t(a) > \bar{X}_t(i^*)\right] \leq \sum_{t \geq t_\epsilon} \frac{\delta}{t^2} \leq \delta.$$

This means that for any round $t \geq t_\epsilon$, any arm a with the highest average score $\bar{X}_t(a)$ must be an ϵ-optimal Borda winner. As a result, we have the following.

Theorem 9 *Given parameters $\epsilon, \delta > 0$, we can identify with probability at least $1 - \delta$ an ϵ-optimal Borda winner in $O\left(\frac{1}{\epsilon^2} \log \frac{n}{\epsilon\delta}\right)$ rounds via at most $O\left(\frac{n}{\epsilon^2} \log \frac{n}{\epsilon\delta}\right)$ competitions.*

Note that the theorem also applies to the task of identifying the Borda winner as it is just an ϵ-optimal Borda winner with $\epsilon = \Delta$. Furthermore, let us remark that the total number of competitions needed here is higher than that of Theorem 5 in the previous section. This is because here we keep all the arms for competitions in every round, in order to estimate the Borda score. In summary, here we consider a more general setting and solve the more difficult task of identifying an ϵ-optimal Borda winner using more competitions, while on the other hand, the mechanism and its analysis are simpler.

Acknowledgments

The work was supported by the Ministry of Science and Technology of Taiwan (MOST 106-2221-E-001-005-MY3).

References

N. Ailon, T. Joachims, and Z. Karnin. Reducing dueling bandits to cardinal bandit. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*, pages 856–864, 2014.

H. Aziz, O. Lev ad N. Mattei, J. S. Rosenschein, and T. Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*, pages 390–396, 2016.

S. N. Bernstein. The theory of probabilities. *Gastehizdat Publishing House, Moscow*, 1946.

I. Caragiannis, G. A. Krimpas, and A. A. Voudouris. Aggregating partial rankings with applications to peer grading in massive online open courses. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15)*, pages 675–683, 2015.

I. Caragiannis, G. A. Krimpas, M. Panteli, and A. A. Voudouris. Co-rank: an online tool for collectively deciding efficient ranking among peers. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*, pages 4351–4352, 2016a.

I. Caragiannis, G. A. Krimpas, and A. A. Voudouris. How effective can simple ordinal peer grading be? In *Proceedings of the 17th ACM Conference on Economics and Computation (EC'16)*, pages 323–340, 2016b.

L. de Alfaro and M. Shavlovsky. Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the Symposium on Special Interest Group on Computer Science Education (SIGCSE'14)*, pages 415–420, 2014.

K. Jamieson, S. Katariya, A. Deshpande, and R. Nowak. Sparse dueling bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS'15)*, pages 416–424, 2015.

C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, 20(6):33:1–33:31, 2013.

D. Kurokawa, O. Lev, J. Morgenstern, and A. Procaccia. Impartial peer review. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI'15)*, pages 582–588, 2015.

C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM'13)*, pages 153–160, 2013.

K. Raman and T. Joachims. Methods for ordinal peer grading. In *Proceedings of the 20th ACM Conference on Knowledge Discovery and Data Mining*, pages 1037–1046, 2014.

N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran. A case for ordinal peer-evaluation in moocs. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS'13)*, 2013.

T. Urvoy, F. Clerot, R. Féraud, and S. Naamane. Generic exploration and k-armed voting bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, pages 91–99, 2013.

T. Walsh. The peerrank method for peer assessment. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI'14)*, pages 909–914, 2014.

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78:1538–1556, 2012.

M. Zoghi, S. Whiteson, R. Munos, and M. deRijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*, pages 10–18, 2014.