mathematics

MDPI

Article

Finding the Best Dueler

Zhengu Zhang † and Sheldon M. Ross *,†

Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA; zhan892@usc.edu

* Correspondence: smross@usc.edu

† These authors contributed equally to this work.

Abstract: Consider a set of n players. We suppose that each game involves two players, that there is some unknown player who wins each game it plays with a probability greater than $1/2$, and that our objective is to determine this best player. Under the requirement that the policy employed guarantees a correct choice with a probability of at least some specified value, we look for a policy that has a relatively small expected number of games played before decision. We consider this problem both under the assumption that the best player wins each game with a probability of at least some specified value $p_0 > 1/2$, and under a Bayesian assumption that the probability that player i wins a game against player j is $\frac{v_i}{v_i+v_j}$, where $v_1, \ldots, v_n$ are the unknown values of n independent and identically distributed exponential random variables. In the former case, we propose a policy where chosen pairs play a match that ends when one of them has had a specified number of wins more than the other; in the latter case, we propose a Thompson sampling type rule.

Keywords: best arm identification; dueling bandit

MSC: 90-10; 62L99

1. Introduction

Consider a set of n players, numbered $1, \ldots, n$. Suppose that each game played involves two players, and that a game between i and j is won by i with some unknown probability $p_{i,j} = 1 - p_{j,i}$. Assuming that there is an unknown player i^* such that $p_{i^*,j} > 1/2$, $j \neq i^*$, our objective is to identify player i^*. To do so, at each stage, we choose two of the players to play a game, with the winner of the game being noted. With a policy being a rule for determining whether to stop and make a choice as to which is the best player (namely, which player is i^*) or to choose a pair to play the next game, we want to find a policy that, with probability at least $1 - \delta$, makes the correct choice, while at the same time minimizing the expected number of games that need be played before a choice is made. We do this both under the Cordorcet assumption that $p_{i^*,j} \geq 0.5 + \epsilon$, $j \neq i^*$, where $\epsilon \in (0, 0.5)$ is a known number, as well as under a Bayesian model that makes the Bradley–Terry–Luce [1,2] assumption that $P_{i,j} = \frac{v_i}{v_i+v_j}$, where $v_1, \ldots, v_n$ are the unknown values of n independent exponential random variables with a mean of 1.

Our problem is closely related to the multi-arm bandit problem, where the objective is to find the best arm. In the conventional stochastic setting, the learner is asked to sample a single arm at each stage and receive a real-valued feedback generated from the unknown distribution associated with the sampled arm. There is a variety of works addressing the identification of the best arm (see, for instance, [3–6]). However, in many scenarios, such as search engine and online recommendation, it is often difficult to obtain explicit and reliable feedback regarding a single arm, as the feedback often shows the preference of the user among a list of options (e.g., 'A looks better than B'). A more appropriate framework, known as dueling bandit, utilizes the pairwise comparison as actions and learns through pairwise preference. Though most dueling bandit algorithms focused on minimizing the

check for updates

Citation: Zhang, Z.; Ross, S.M. Finding the Best Dueler. *Mathematics* **2023**, *11*, 1568. https://doi.org/10.3390/math11071568

Academic Editors: Francisco German Badía and María D. Berrade

Received: 2 February 2023
Revised: 16 March 2023
Accepted: 19 March 2023
Published: 23 March 2023

cumulative regret [7–9], many recent works (such as [10–12]) were developed under various notions of the best arm.

In Section 2, we look at the Condorcet winner setting. We propose two policies that use a knockout tournament structure to successively eliminate players. We suppose that, in each round, players still in contention are randomly paired and play a match, where a round j match ends when one of them has m_j more wins than the other. The match winners move on to the next round and the losers are eliminated from contention. The winner of the final match is then chosen as being the best. We show how to determine the critical numbers m_j so as to guarantee that the probability that i^* is the chosen player is at least $1-\delta$. We also consider a modification of this rule such that if in a round j match there has not been a winner after n_j games, then that match is ended and both of its participants are eliminated. We present upper bounds on the mean number of games needed by these policies as well as numerical evidence that these rules outperform others in the literature.

In Section 3, we turn our attention to the Bradley–Terry–Luce model. We propose a randomized policy whose logic uses a Thompson sampling approach to determine how to choose the next pair. To utilize this policy, we show how to effectively simulate from the posterior joint distribution of the player's values and how to effectively use simulation to determine the posterior probability that a given player has the largest value.

Conclusions are presented in the final section.

2. The Condorcet Winner Model

In this section, we make the Condorcet assumption that there is an unknown player i^* such that $p_{i^*,j} \geq p_0 = 0.5 + \epsilon,\ j \neq i^*$, where $\epsilon \in (0, 0.5)$ is a known number. Let k be the positive integer for which $2^{k-1} < n \leq 2^k$. Our policy utilizes a knockout tournament structure as follows.

Knockout Tournament Framework

- Initialization: all players are alive
- For round $t = 1, 2, \ldots, k$
  - If the number of alive players is odd, one of the players is randomly selected and given a bye. The others are randomly paired up.
  - If the number of alive players is even, randomly pair up these players.
  - Each pair then plays a match, consisting of a series of games. Depending on the match rules, at some point one of the players is declared the winner of the match.
  - The match winners along with the player given a bye, if there was such a player, remain alive and move on to the next round. The match losers are eliminated.
- Claim the winner of the match in round k as the best dueler.

In the following two sections, we will present two ways of determining the winner for each match. Note that players who receive a bye in some rounds automatically advance to the next round.

2.1. A Gambler's Ruin Rule

Adopting the framework above, we propose a Gambler's Ruin Rule (GRR) to determine the winner of each match. Let $r_0 = \frac{p_0}{1-p_0} = \frac{1+2\epsilon}{1-2\epsilon}$, let k be the positive integer for which $2^{k-1} < n \leq 2^k$, let $m_t^* = \log_{r_0}(2^t/\delta) = \frac{\ln(2^t/\delta)}{\ln(r_0)}$, and let $m_t = \text{ceil}(m_t^*)$, $t \geq 1$, where $\text{ceil}(a)$, called the ceiling of a, is the smallest integer at least as large as a.

Gambler's Ruin Rule

- In round t, each pair plays a sequence of games until one of them has achieved m_t more wins than the other, with the one with more wins being declared the winner.

Lemma 1. *GRR identifies the best dueler i^* with probability at least $1-\delta$.*

Proof. Given that i^* successfully proceeds to round t, the probability that i^* is eliminated in round t, denoted by P_t, can be upper bounded by using the gambler's ruin probability

$$P_t \leq \frac{1 - r_0^{m_t}}{1 - r_0^{2m_t}} = \frac{1}{1 + r_0^{m_t}} < \frac{1}{r_0^{m_t^*}} = \frac{\delta}{2^t}$$

To win the tournament, i^* needs to win all k rounds. Hence,

$$\begin{aligned} P(i^* \text{ is eliminated}) &= P(\cup_{t=1}^{k} \{i^* \text{ is eliminated at round } t\}) \\ &\leq \sum_{t=1}^{k} P(i^* \text{ is eliminated at round } t) \\ &< \sum_{t=1}^{k} P_t \\ &< \delta \end{aligned}$$

which indicates that the probability of finding the best arm is at least $1 - \delta$. □

Next, we show how to upper bound the expected number of games played when using GRR.

Let $N_m(p)$ be the total number of games for a match between players A and B, which ends when one of the players is ahead by m, where p is the probability that player B wins each game. The following Lemma shows that $E[N_m(p)]$ is a unimodal function that is maximized when $p = 0.5$.

Lemma 2. *The expected number of plays until one of the players is ahead by m is a decreasing function of p when* $p \geq 1/2$.

Proof. Suppose that $p \neq 1/2$, and let $r = p/(1-p)$. We first show that $E[N_m(p)]$ is a decreasing function of p for $p > 1/2$. Let, for $i \geq 1$, $X_i = 1$ if player A wins game i and let it be -1 otherwise. Then, Wald's equation gives that

$$\begin{aligned} E[N_m(p)](2p-1) &= E[\sum_{i=1}^{N_m(p)} X_i] \\ &= \frac{m}{1+r^m} - \frac{mr^m}{1+r^m} \end{aligned}$$

where the final equality used the gambler's ruin probability

$$P(\sum_{i=1}^{N_m(p)} X_i = m) = \frac{1 - r^m}{1 - r^{2m}} = \frac{1}{1 + r^m}$$

Because $2p - 1 = \frac{r-1}{r+1}$, the preceding gives

$$E[N_m(p)] = m\frac{r+1}{r-1}\frac{r^m - 1}{r^m + 1}$$

As r is an increasing function of p, it suffices to show that $f(r) \equiv \frac{r+1}{r-1}\,\frac{r^m-1}{r^m+1}$ is a decreasing function of r when $r > 1$. Now,

$$\begin{aligned} f'(r) = & \frac{\left(r^m-1+m(r+1)r^{m-1}\right)(r-1)(r^m+1)}{(r-1)^2(r^m+1)^2} \\ & -\frac{\left(r^m+1+(r-1)mr^{m-1}\right)(r+1)(r^m-1)}{(r-1)^2(r^m+1)^2} \\ = & \frac{2mr^{m+1}-2mr^{m-1}-2r^{2m}+2}{(r-1)^2(r^m+1)^2} \end{aligned}$$

Let $g(r) = mr^{m+1} - mr^{m-1} - r^{2m} + 1$. It suffices to show that $g(r) < 0$ for all $r > 1$. Now,

$$\begin{aligned} g(r) &= (r^2-1)mr^{m-1} - r^{2m} + 1 \\ &= (r^2-1)mr^{m-1} - (r^2-1)\Big(\sum_{i=0}^{m-1} r^{2i}\Big) \\ &= (r^2-1)\Big(mr^{m-1} - \sum_{i=0}^{m-1} r^{2i}\Big) \end{aligned}$$

By the arithmetic and geometric means' inequality,

$$\frac{\sum_{i=0}^{m-1} r^{2i}}{m} \geq \sqrt[m]{\prod_{i=0}^{m-1} r^{2i}} = r^{m-1}$$

Thus,

$$g(r) \leq (r^2-1)(mr^{m-1} - mr^{m-1}) = 0$$

Hence, $E[N_m(p)]$ decreases in p when $p > 1/2$. Because $E[N_m(p)]$ is a continuous function of p that is symmetric about $1/2$, it follows that its maximal value occurs when $p = 1/2$, which completes the proof. □

Corollary 1. $E[N_m(p) \leq m^2$.

Proof. This follows as it is well known that $E[N_m(1/2)] = m^2$. □

Now, let G_t be the number of games played in round t, and let $G = \sum_{j=1}^{k} G_t$ be the total number of games played. As Lemma 2 implies that $E[X_m] \leq m^2$, we see that $E[G] \leq \sum_{t=1}^{k} 2^{k-t} m_t^2$. This upper bound can be improved by using that the m^2 upper bound can be decreased if the best player is involved in the match. Indeed, it follows from Lemma 2 that the mean number of games in a match involving the best player ,which ends when one of the players is ahead by m, is upper bounded by

$$b(m) = m\,\frac{r_0+1}{r_0-1}\,\frac{r_0^m-1}{r_0^m+1}.$$

Proposition 1.

$$E[\textit{number of plays}] \leq \sum_{t=1}^{k} 2^{k-t} m_t^2 - \sum_{t=1}^{k} \left(m_t^2 - b(m_t)\right) \prod_{s=1}^{j-1} \frac{r_0^{m_t}}{1+r_0^{m_t}}$$

Proof. Let R be the number of rounds played by the best player. Conditioning on whether the best player plays in round t yields that

$$\begin{aligned} E[G_t] &\leq (2^{k-t}-1)m_t^2 + P(R \geq t)b(m_t) + P(R < t)m_t^2 \\ &= 2^{k-t}m_t^2 - P(R \geq t)(m_t^2 - b(m_t)) \end{aligned}$$

and the result follows because the proof of Lemma 1 implies that $P(R \geq t) \geq \prod_{s=1}^{t-1} \frac{r_0^{m_s}}{1+r_0^{m_s}}$.

□

Remark 1. *The upper bound of Proposition 1 is attained when* $n = 2^k$, $p_{i^*,j} = p_0$, $j \neq i$, *and* $p_{i,j} = 0.5, i, j \neq i^*$.

Of other methods considered in the literature, the closest to ours is the rule proposed in [13]. (Other rules, such as those of [12,14], deal with more specific models that typically assume, among other things, that there is a ranking of the players such that the probability that a higher ranked player will win a game against a lower ranked one is at least 0.5. In addition, numerical results cited in [13] indicate that its rule tends to outperform the others.)

Although the rule of [13], like GRR, uses a knockout tournament structure that eliminates half the remaining players in each round, it differs in two ways from GRR. The first is in how a match is decided, with the rule in [13] having a match consisting of a fixed odd number g of games and then letting the winner of the match be the one with more wins. The second way is that g is fixed and does not depend on the round. We now argue that the GRR way of deciding the winner of a match is superior.

Let the m-rule be the rule where each match, in any round, is decided when one of the players has m more wins than the other, and let the g-rule be one where each match consists of g games. To compare these, let $L_1(m, p)$ and $L_2(g, p)$ be the probabilities that the better player would lose a match when using an m-rule and when using a g-rule, when the better player wins each game with probability p. (Thus, $L_2(g, p) = P(\text{Bin}(g, p) < (g + 1)/2)$, where $\text{Bin}(g, p)$ is a binomial random variable with parameters (g, p).) The following table gives some values for these quantities when $p = 0.6$.

Thus, for instance, if $p_0 = 0.6$, then the use of the g-rule with $g = 77$ would result in each match being 77 games and have a resulting success probability of about $1 - k \times 0.0376$. On the other hand, use of the m-rule with $m = 8$ would lead to the same success probability, with the mean number of games in a match between i and j having a value that ranges between 8 and $m^2 = 64$ as $|P_{i,j} - 0.5|$ ranges from 0.5 to 0. On the other hand, if one wanted a larger success probability, then a g-rule with $g = 93$ and the m-rule with $m = 9$ both would result in a success probability of approximately $1 - k \times 0.02536$, with the g-rule requiring 93 games per match, and the m-rule requiring a mean number of games per match ranging from 9 to a maximum of 81.

The GRR rule modifies the m-rule by allowing a different value of m in each round. Because the number of matches in each round decreases exponentially, it seems intuitive to have shorter matches in earlier rounds, which is what GRR does. For instance, in the case where $k = 5$ and $P_{i^*,j} = 0.6$, $j \neq i$ and $P_{i,j} = 0.5, i, j \neq i^*$, Table 1 indicates that if $m_t = 11, t \leq k$, then the probability of an incorrect choice is approximately 0.057, with the mean number of games needed being 3422.31. On the other hand, the mean number of games needed in this case by the GRR rule with $\delta = 0.057$ is 3093.72 (The means are computed by using Proposition 1).

The next section considers a modification of the GRR rule.

Table 1. Comparison of match win probabilities for g- and m-rules when $p = 0.6$.

m	$L_1(m, 0.6)$	g	$L_2(g, 0.6)$
8	0.0376	77	0.0376
9	0.02535	93	0.02537
10	0.01704	109	0.01724
11	0.0114	125	0.0118
15	0.00228	197	0.00226

2.2. Modified Gambler's Ruin Rule

One underlying drawback of GRR is that it may play too many games between two suboptimal arms to determine which seems better. In such cases, one might consider eliminating both arms as none of them show the potential to be best. Therefore, we can often improve GRR by limiting the number of games in each match, and drop both arms if none of them can win the match by the end. The resulting rule, called the Modified Gambler's Ruin Rule (MGRR), is as follows.

Modified Gambler's Ruin Rule

- Let $w_t^* = \frac{1}{4\epsilon} \ln(2^t/\delta)$, let $w_t = \text{ceil}(w_t^*)$, and let $n_t = \text{ceil}(3w_t^*/\epsilon)$, $t \geq 1$. In round t, play each pair until either one is ahead by w_t, with the leader being the winner, or until the total number of games reaches n_t, in which case both arms are eliminated.

  As a preparation of showing the strength of MGRR, we need the following Lemma.

Lemma 3. *For* $0 \leq x \leq 1$

$$\frac{1-x}{1+x} \leq e^{-2x}.$$

Proof. Let $f(x) = (1-x)e^{2x} - (1+x)$. It suffices to show that $f(x) \leq 0$ for $0 \leq x \leq 1$. Now,

$$\begin{aligned} f'(x) &= e^{2x} - 2xe^{2x} - 1 \\ f''(x) &= -4xe^{2x} \end{aligned}$$

Since $f''(x) \leq 0$, it follows that $f'(x)$ is decreasing, which, since $f'(0) = 0$, shows that $f(x)$ is decreasing. Hence, $f(x) \leq f(0) = 0$. □

Lemma 4. *MGRR identifies the best arm* i^* *with probability at least* $1 - \delta$.

Proof. Given that the best player successfully advances to round t and that she wins each game played in round t with probability a, let $P_t(a)$ denote the conditional probability that the best player is eliminated in round t. Let $X_i, i \geq 1$ be independent Bernoulli random variables such that

$$X_i = \begin{cases} 1 & \text{with probability } a \\ -1 & \text{with probability } 1-a \end{cases}$$

and let $S_r(a) = \sum_{i=1}^{r} X_i,\ r \geq 1$. Then,

$$\begin{aligned} P_t(a) &= P(S_r(a) \text{ hits } -w_t \text{ before } w_t \cup S_r(a) \text{ does not hit } w_t \text{ within } n_t \text{ steps}) \\ &\leq P(S_r(a) \text{ hits } -w_t \text{ before } w_t) + P(S_r(a) \text{ does not hit } w_t \text{ within } n_t \text{ steps}) \\ &\leq P(S_r(a) \text{ hits } -w_t \text{ before } w_t) + P(S_{n_t}(a) < w_t) \end{aligned}$$

Because $a \geq p_0 = 1/2 + \epsilon$ and both terms on the right side of the preceding inequality are decreasing in a, we have that

$$\begin{aligned} P(S_r(a) \text{ hits} - w_t \text{ before } w_t) &\leq (1/r_0)^{w_t} \\ &\leq (\frac{1-2\epsilon}{1+2\epsilon})^{\frac{1}{4\epsilon}\ln(2^t/\delta)} \\ &\leq e^{-\ln(2^{t+1}/\delta)} \\ &= \frac{\delta}{2^{t+1}} \end{aligned}$$

where the second inequality follows by Lemma 4. In addition,

$$\begin{aligned} P(S_{n_t}(a) < w_t) &\leq P(S_{n_t}(p_0) < w_t) \\ &= P(S_{n_t}(p_0)) - 2n_t\epsilon < w_t - 2n_t\epsilon) \\ &\leq \exp(-\frac{(w_t - 2n_t\epsilon)^2}{2n_t}) \\ &\leq \exp(-\frac{25}{24}\ln(2^{t+1}/\delta)) \\ &< \exp(-\ln(2^{t+1}/\delta)) \\ &= \frac{\delta}{2^{t+1}} \end{aligned}$$

where the third inequality uses Azuma inequality (see [15]). Hence, $P_t(a) \leq \frac{\delta}{2^t}$, which shows that the conditional probability that the best player is eliminated in round t given that she advances to that round is at most $\frac{\delta}{2^t}$. However, by the same argument as in Lemma 1, this shows that the probability that the best arm is identified is at least $1 - \delta$. □

Remark 2.

- *Since the number of games is upper bounded in each match, we are able to derive the upper bound of the total number of games when using MGRR:*

$$\begin{aligned} \textit{number of game} &\leq \sum_{t=1}^{k} 2^{k-t} X_t \\ &= \frac{3}{4\epsilon^2}\sum_{t=1}^{k} 2^{k-t}(\ln 2^{t+1} + \ln\frac{1}{\delta}) \\ &= \frac{3n}{4\epsilon^2}\sum_{t=1}^{k}\frac{\ln 2^{t+1} + \ln\frac{1}{\delta}}{2^t} \\ &< \frac{3n}{4\epsilon^2}(4 + \ln\frac{1}{\delta}) \\ &= O(\frac{n\ln\frac{1}{\delta}}{\epsilon^2}) \end{aligned}$$

- *There is basically no downside in using MGRR as opposed to GRR. Although $w_t^* > m_t^*$, the difference is usually small and often $w_t = m_t$. To see this, note that*

$$\frac{w_t^*}{m_t^*} = \frac{\ln(\frac{1+2\epsilon}{1-2\epsilon})}{4\epsilon} \tag{1}$$

Since $\frac{1+2\epsilon}{1-2\epsilon} - 1 = \frac{4\epsilon}{1-2\epsilon}$, *the Taylor series expansion of* $f(x) = \ln(x)$ *about* 1 *gives that*

$$\ln(\frac{1+2\epsilon}{1-2\epsilon}) \approx \frac{4\epsilon}{1-2\epsilon} - (\frac{4\epsilon}{1-2\epsilon})^2/2 + (\frac{4\epsilon}{1-2\epsilon})^3/3$$

For an illustration, suppose $\epsilon = 0.05$, $\delta = 0.01$. *Then,* $w_3^* = 33.42, m_3^* = 33.31, n_3 = 2006$, *so* $w_3 = m_3 = 34$. *Now, if* $P_{i,j} = 1/2$, *then the mean and variance of the number of games needed between players i and j until one is up by m is* m^2 *and* $2m^2(m^2-1)/3$ *(see [16] for the variance formula). Letting* N_{GRR} *and* N_{MGRR} *be the number of round* 3 *games such a match would take when using GRR and when using MGRR, it follows that the mean and standard deviation of* N_{GRR} *are* 1156 *and* 943.46. *Hence, as* $N_{MGRR} = \min(N_{GRR}, 2006)$, *it follows that MGRR stops the match when the number of games played is roughly one standard deviation above the mean of* N_{GRR}, *which should result in a reasonable decrease in the mean number of games needed. (For instance, if X is exponential with mean* 1, *then* $E[\min(X,2)] = 1 - e{-}2 = 0.865$.*)*

- *The validity of* $w_t^* > m_t^*$ *follows from (1) upon using Lemma 3.*

The following Table 2 compares the performances of GRR and MGRR when $p_{i^*,j} = p_0$, $p_{i,j} = 0.5, i^* \neq i \neq j$, and $n = 2^k$.

Table 2. Mean number of games needed by GRR and MGRR.

$\epsilon = 0.1, \delta = 0.05$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
GRR	203.18	591.48	1482.14	3415.14
MGRR	196.82	565.21	1378.99	3112.67
$k = 3, \delta = 0.05$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
GRR	2240.49	591.48	153.51	61.40
MGRR	2156.00	563.03	148.96	75.65

3. The Bradley–Terry–Luce Bayesian Model

Suppose now that player i has an unknown associated value v_i, and that a game between players i and j is won by i with probability $\frac{v_i}{v_i+v_j}$. Furthermore, suppose that $v_1, \ldots, v_n$ are the values of n independent exponential random variables $V_1, \ldots, V_n$ having a mean of 1. As before, our objective is to identify player i^*, where $i^* = \text{argmax}\, v_i$. However, because we are assuming a prior distribution on the values, we now require that the posterior probability that our decision is correct is at least $1 - \delta$. That is, if C is the event that we made the correct choice, then we require that our rule is such that $P(C|\text{all data}) \geq 1 - \delta$. Subject to this constraint, we want the expected number of games played to be relatively small. Because we want to finish as soon as possible and we require that the posterior probability that we have made the correct decision is at least $1 - \delta$, it is clearly optimal to stop as soon as there is some r for which $P(V_r = \max_j V_j|\text{all data}) \geq 1 - \delta$. More precisely, if $w_{i,j}$ is the number of times that i has beaten j, then we should stop and declare for r if $P(V_r = \max_j V_j|w_{i,j}, i \neq j) \geq 1 - \delta$.

The rule we suggest for determining the pair to play the next game is a randomized policy that relates to the Thompson sampling approach used in bandit problems (see [17,18]). Letting $V_{(1)} > V_{(2)} > \ldots > V_{(n)}$ be the ordered values of $V_1, \ldots, V_n$, and $P_{i,j}, i \neq j$, be the posterior probability that $V_{(1)} = V_i, V_{(2)} = V_j$, then i and j are chosen to be the next pair with probability $P_{i,j} + P_{j,i}$. We can implement this rule by simulating a random vector $V_1^*, \ldots, V_n^*$ having the conditional distribution (given all data) of $V_1, \ldots, V_n$. If V_i^* and V_j^* are the two largest of $V_1^*, \ldots, V_n^*$ then i and j are chosen to play the next game. Because it is difficult to directly simulate from the posterior distribution of $V_1, \ldots, V_n$, we next develop a Markov chain Monte Carlo approach for doing so.

3.1. The Sampling Approach: MCMC

With $w_{i,j}$ denoting the current number of times player i has beaten j, the conditional (e.g., posterior) density of $\mathbf{V} = (V_1, \ldots, V_n)$ is

$$f(x_1, \ldots, x_n) = Ce^{-\sum_i x_i} \prod_{i \neq j} \left(\frac{x_i}{x_i + x_j} \right)^{w_{i,j}} \tag{2}$$

for a normalization factor C.

As noted previously, we now want to simulate from the preceding distribution and let the next game be between the two indices whose simulated values are largest. However, because directly simulating $\mathbf{V}$ from (2) does not seem computationally feasible (for one thing, C is difficult to compute), we utilize the Hasting–Metropolis algorithm (see [19]) to generate a Markov chain whose limiting distribution is given by (2). The Markov chain is defined as follows. When its current state is $\mathbf{x} = (x_1, \ldots, x_n)$, a coordinate that is equally like to be any of $1, \ldots, n$ is selected. If i is selected, a random variable Y is generated from an exponential distribution with mean x_i, and if $Y = y$, then $\mathbf{y} = (x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$ is considered as the candidate next state. In other words, if we let $\mathbf{y} = (x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$, the density function for the candidate next state is

$$q(\mathbf{y}|\mathbf{x}) = \frac{1}{n} \frac{1}{x_i} e^{-y/x_i}$$

The next state of the Markov chain, call it $\mathbf{x}^*$, is such that

$$\mathbf{x}^* = \begin{cases} \mathbf{y} & \text{with probability } \alpha(\mathbf{x}, \mathbf{y}) \\ \mathbf{x} & \text{with probability } 1 - \alpha(\mathbf{x}, \mathbf{y}) \end{cases}$$

where

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ \frac{f(\mathbf{y})}{f(\mathbf{x})} \frac{q(\mathbf{x}|\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}, 1 \right\}$$

The limiting distribution of this Markov chain is the posterior distribution of $V_1, \ldots, V_n$. Consequently, we can approximately simulate from the posterior by generating a large number of states of the chain and then choosing the two largest indices of the final state to play the next game. However, as it probably makes little difference if we choose i and j to play the next game not with the exact posterior probability that these are the two arms with largest values but with a probability close to the exact one, in practice, we do not need to determine many states of the Markov chain. Indeed, it is not clear that using the exact probabilities would lead to improved results. (In practice, for $n \leq 10$, 100 states of the Markov chain should suffice.) Moreover, after choosing a pair and observing the result of their game, then because of the new posterior distribution, which given the result of the last game should not be much different from the previous one, the initial state of the Markov chain used to determine the next pair should be chosen to be the final state of the previous chain.

Whereas the preceding simulations can be used to estimate the probability that a given player is best, we do not recommend using it to determine when to stop. Indeed, if a player's probability of being best appears to have a reasonable chance of being as large as $1 - \delta$, we propose to use the method in the next subsection to estimate $P(V_r = \max_j V_j | \text{all data})$.

3.2. The Stopping Criteria: A Simulation Approach

In this subsection, we will present a simulation approach to estimate $P(V_r = \max_j V_j|$ all data$)$. It follows from (2) that for $r = 1, \ldots, n$

$$P(V_r = \max_i V_i | w_{i,j}, i \neq j) = \frac{E[I\{V_r = \max_i V_i\} \prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}}]}{E[\prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}}]} \quad (3)$$

$$= \frac{E[\prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}} | V_r = \max_i V_i]}{nE[\prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}}]}$$

$$= K\, E[\prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}} | V_r = \max_i V_i], \quad (4)$$

where $V_1, \ldots, V_n$ are iid exponentials with rate 1.

Thus, we can use simulation to estimate $P_r \equiv P(V_r = \max_i V_i | w_{i,j}, i \neq j)$, $r = 1, \ldots, n$ as follows. In the tth simulation run, generate n independent exponentials with rate 1, $V_1, \ldots, V_n$ and let i^* be such that $V_{i^*} = \max_i V_i$. To estimate $E[\prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}} | V_r = \max_i V_i]$, let

$$X_j(r) = \begin{cases} V_j, & \text{if } j \neq i^*, j \neq r \\ V_{i^*}, & \text{if } j = r \\ V_r, & \text{if } j = i^* \end{cases}$$

and let $b_r^{(t)} = \prod_{i \neq j} (\frac{X_i(r)}{X_i(r)+X_j(r)})^{w_{i,j}}$. Perform the preceding for each $r = 1, \ldots, n$. If we conduct m simulation runs, then the estimator of $P(V_r = \max_i V_i | w_{i,j}, i \neq j)$ is $\frac{\sum_{t=1}^m b_r^{(t)}}{\sum_{r=1}^n \sum_{t=1}^m b_r^{(t)}}$.

In practice, it turns out that the variance of $\prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}}$ is very large. While this might not make much difference when using the proposed policy, it makes simulation studies of the effectiveness of the procedure difficult. To ameliorate this difficulty, we suggest using the following importance sampling estimator, which in our numerical experiments tended to reduce the variance by over 30%.

An Importance Sampling Estimator

Suppose we are at a stage where every player has at least one win. Let $w_i = \sum_{j \neq i} w_{i,j}$ be the total number of wins of player i, and let $w = \sum_{i=1}^n w_i$ be the total number of games played. Further, let $Y_1, \ldots, Y_n$ be independent, with Y_i being exponential with rate $\frac{w}{nw_i}$, $i = 1, \ldots, n$. Then, the importance sampling identity (see [19]) gives

$$E[I\{V_r = \max_i V_i\} \prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}}]$$

$$= (\prod_{i=1}^n \frac{nw_i}{w})\, E[I\{Y_r = \max_i Y_i\} \prod_{i \neq j} (\frac{Y_i}{Y_i+Y_j})^{w_{i,j}} \prod_{i=1}^n \exp\big((\frac{w}{nw_i} - 1)Y_i\big)] \quad (5)$$

Thus, each simulation run generates $Y_1, \ldots, Y_n$ and, for each $r = 1, \ldots, n$, yields an unbiased estimator of $E[I\{V_r = \max_i V_i\} \prod_{i \neq j} (\frac{V_i}{V_i+V_j})^{w_{i,j}}]$. In each run, all but one of these n estimators will equal 0.

We now give numerical examples comparing the Thompson sampling rule with the MGRR rule. It is worth noting that the implementation of Thompson sampling rule does not require knowledge of ϵ, which specifies the least gap between the best player and an arbitrary player. We consider two examples with 5 players, where in the first example we use fixed strength $v = (0.3, 0.5, 0.7, 0.9, 1.5)$ and in the second example we randomly generate strengths from exponential (1) for each replication—that is, all replications in the

first example use the same strength vector, whereas in the second example each replication starts by simulating player strengths from an exponential with rate 1.

In all cases, when using the MGRR rule, we take $\epsilon = \frac{V_{(1)}}{V_{(1)}+V_{(2)}} - 1/2$. We run 100 iterations of MCMC to determine the next pair and we utilize importance sampling in estimating the probabilities that check for stopping. The results, using $\delta = 0.05$, are summarized in Tables 3 and 4. The standard deviation columns refers to the standard deviation of the estimator of the expected number of games until stopping.

Table 3. Numerical example of Thompson sampling rule where strengths v = (0.3, 0.5, 0.7, 0.9, 1.5). Replication = 5000.

Method	Percentage of Correct	Mean Number of Games	Standard Deviation
MGRR	0.99	115.612	1.39
Thompson Sampling	0.9886	98.9916	0.8671405

Table 4. Numerical example of Thompson sampling rule where strengths are randomly generated from exponential (1). Replication = 3000.

Method	Percentage of Correct	Mean Number of Games	Standard Deviation
MGRR	0.99	8520	354
Thompson Sampling	0.953	248.3	13.5

4. Conclusions

We have considered the problem of finding the best among a set of n players when we learn about the player's skills by successively choosing a pair of players and having them play a game. Our objective is to find a policy that minimizes the expected number of games to find the best player, subject to the condition that the probability of a correct choice is at least some specified value.

In our first model, we suppose that it is known that one of the players, called the best, will win each game it plays with a probability of at least $1/2 + \epsilon$, where ϵ is a known positive value. The policy we suggest is based on a knockout tournament structure, where we have pairs play a match, with the winner of the match remaining in contention and the loser being eliminated. Whereas other policies in the literature using a knockout tournament structure let a match consist of a fixed number of odd games, with the winner being the one with more wins, we let a match end when one of the players has won a fixed number of games more than the other. We argue that our sequential-type matches lead to superior results. We also show how to improve this policy by letting the number of games one must be ahead to win the match depend on the number of remaining players, and by allowing for the stopping of a match after a fixed number of games if neither player has won by then, with both players being eliminated in this case.

Our second model supposes that each player has an unknown value, and that a game between two players with values v and w is won by the player with value v with probability $\frac{v}{v+w}$. Supposing that these values have a known exponential prior distribution, the objective is to minimize the expected number of games needed to identify the player with the largest value, subject to the condition that the posterior probability that our decision is correct is at least some specified value. We present a Thompson sampling type policy and give a simulation approach to estimate its resulting expected number of games needed. The simulation results give evidence of the strength of this policy. Additional numerical work is planned for future research.

Author Contributions: Investigation, Z.Z. and S.M.R.; Writing—review & editing, Z.Z. and S.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: The second author's work was supported by, or in part, by the National Science Foundation under contract/grant CMMI2132759.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Braley, R.A.; Terry, M.E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **1952**, *39*, 324–345.
2. Luce, R.D. *Individual Choice Behavior: A Theoretical Analysis*; Courier Corporation: North Chelmsford, MA, USA, 2012.
3. Audebert, J.Y.; Bubeck, S.; Munos, R. Best arm identification in multi-armed bandits. In Proceedings of the 23rd Annual Conference on Learning Theory (COLT 2010), Haifa, Israel, 27–29 June 2010; pp. 41–53.
4. Azizi, M.J.; Ross, S.M.; Zhang, Z. Choosing the Best Arm with Guaranteed Confidence. *J. Stat. Theory Pract.* **2022**, *16*, 71. [CrossRef]
5. Even-Dar, E.; Mannor, S.; Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.* **2006**, *7*, 1079–1105.
6. Jamieson, K.; Katariya, S.; Despande, A.; Novak, R. Sparse dueling bandits. In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015), San Diego, CA, USA, 9–12 May 2015; pp. 416–424.
7. Komiyama, J.; Honda, J.; Kashima, H.; Nakagawa, H. Regret lower bound and optimal algorithm in dueling bandit problem. In Proceedings of the 28th Conference on Learning Theory (COLT 2015), Paris, France, 3–6 July 2015; pp. 1141–1154.
8. Peköz, E.; Ross, S.M.; Zhang, Z. Dueling Bandits. *Prob. Eng. Inf. Sci.* **2022**, *36*, 264–275. [CrossRef]
9. Zoghi, M.; Whiteson, S.; Munos, R.; Rijke, M. Relative upper confidence bound for the k-armed bandit dueling problem. In Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML'14), Beijing, China, 21–26 June 2014; pp. 10–18.
10. Jamieson, K.; Malloy, M.; Novak, R.; Bubeck, S. lilucb: An optimal exploration algorithm for multi-armed bandits. In Proceedings of the 27th Conference on Learning Theory (COLT 2014), Barcelona, Spain, 13–15 June 2014; pp. 423–439.
11. Szorenyi, B.; Busa-Fekete, R.; Paul, A.; Hullermeier, E. Online rank elicitation for Plackett-Luce: A dueling bandits approach. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 604–612.
12. Yue, Y.; Joachims, T. Beat the mean bandit. In Proceedings of the 28th International Conference on Machine Learning, ICML-11, Bellevue, WA, USA, 28 June–2 July 2011; pp. 241–248.
13. Mohajer, S.; Suh, C.; Elmahdy, A. Active learning for top-k rank aggregation from noisy comparisons. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2488–2497.
14. Falahatgar, M.; Orlitski, A.; Pichapati, V.; Suresh, A.T. Maximum selection and ranking under noisy comparisosns. *arXiv* **2017**, arXiv:1705.05388.
15. Ross, S.M. *Stochastic Processes*, 2nd ed.; John Wiley: Hoboken, NJ, USA, 1996.
16. Andel, J.; Hudecova, S. Variance of the game duration in the gambler's ruin problem. *Stat. Probab. Lett.* **2012**, *82*, 1750–1754. [CrossRef]
17. Agrawal, S.; Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In Proceedings of the 25th Annual Conference on Learning Theory, Edinburgh, UK, 25–27 June 2012; pp. 1–39.
18. Russo, D.; Van Roy, B.; Kazerouni, A.; Osband, I.; Wen, A. A tutorial on Thompson sampling. *arXiv* **2017**, arXiv:1707.02038.
19. Ross, S.M. *Simulation*, 6th ed.; Academic Press: Cambridge, MA, USA, 2023.