An Optimal Algorithm for Stopping on the Element Closest to the Center of an Interval

Ewa Kubicka[1], Grzegorz Kubicki[1], Małgorzata Kuchta[2], and Małgorzata Sulkowska[2]

[1] University of Louisville, USA
ewa@louisville.edu, gkubicki@louisville.edu
[2] Wrocław University of Science and Technology, Poland
malgorzata.kuchta@pwr.edu.pl, malgorzata.sulkowska@pwr.edu.pl

Abstract

Real numbers from the interval [0, 1] are randomly selected with uniform distribution. There are n of them and they are revealed one by one. However, we do not know their values but only their relative ranks. We want to stop on recently revealed number maximizing the probability that that number is closest to $\frac{1}{2}$. We design an optimal stopping algorithm achieving our goal and prove that for large n its probability of success is of order $\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}$.

Keywords : *Combinatorial optimization, Optimal stopping, Secretary problem.*

1 Introduction

Consider the following online problem: n numbers randomly selected from the interval [0, 1] are presented to us one number at a time. After revealing t numbers, $1 \leq t \leq n$, we know their ranks but not their values. Our goal is to stop on the presently revealed number x_t hoping that x_t is closest to $\frac{1}{2}$, the center of the interval, among all n numbers. We will construct an optimal stopping algorithm and show that this algorithm, for large values of n, has the probability of success of order $\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}$.

This problem is a new relative of the classical secretary problem. In the classical secretary problem, the goal is to choose the best of n linearly ordered objects. In our model, this corresponds to choosing the object whose value is closest to 1. The classical secretary problem, whose solution was derived by Lindley (1961), has attracted a lot of attention and many generalizations of the classical problem have been studied, for example problems in which linear orders have been replaced by partial orders (Morayne, 1998; Preater, 1999), or by a graph or digraph structure (Kubicki & Morayne, 2005; Benevides & Sulkowska, 2017). Rogerson (1987) derived the probability that the optimal algorithm for choosing the best candidate returns j^{th} candidate. There are still very natural questions referring to the classical secretary problem that remain unanswered. It seems that choosing the middle rank element is the hardest one. The problem we study in this paper is similar to choosing the middle rank element, but it is not exactly the same. In our case, stopping on the element of middle rank does not guarantee that this number would be closest to $\frac{1}{2}$. Also, stopping on elements other than the one of middle rank gives a nonzero probability of success. Another difference is that since we refer to a specific value, namely $\frac{1}{2}$, we have to make some assumption about the distribution of incoming numbers whose ranks we observe. The most natural one seems to be the uniform distribution.

2 Optimal Stopping Algorithm

Assume that n different numbers $x_1, x_2, ..., x_n$ from the interval [0, 1] are randomly selected, with uniform distribution, and presented to us one by one. We know n in advance but after revealing t numbers, $1 \le t \le n$, we know only their relative ranks, not their values. Let's rename them such that, at that moment, we know their order $y_1^{(t)} < y_2^{(t)} < ... < y_t^{(t)}$ and we know that the rank of x_t is r; it means that $x_t = y_r^{(t)}$. Our goal is to stop on the presently revealed number x_t maximizing the probability that $|x_t - \frac{1}{2}| \le |x_i - \frac{1}{2}|$ for all i, $1 \le i \le n$, the probability that x_t will be the closest to the midpoint of the interval; we will call such an event "x_t is the best".

Before constructing optimal stopping algorithm, we need two results providing formulas for the probability that the number of specific rank is the best.

Theorem 1 *If $y_1 < y_2 < ... < y_r < ... < y_n$ are the ranked numbers at time n, then*

$$Pr(y_r \text{ is the best}) = \binom{n-1}{r-1} \cdot \frac{1}{2^{n-1}}.$$

Proof : We have

$$\begin{aligned}\Pr(y_r \text{ is the best}) &= \Pr\Big(\big(y_r < \frac{1}{2} < y_{r+1}\big) \text{ and } \big(|y_r - \frac{1}{2}| \le |y_{r+1} - \frac{1}{2}|\big)\Big) \\ &+ \Pr\Big(\big(y_{r-1} < \frac{1}{2} < y_r\big) \text{ and } \big(|y_{r-1} - \frac{1}{2}| \ge |y_r - \frac{1}{2}|\big)\Big).\end{aligned}$$

Note that

$$\begin{aligned}&\Pr\Big(\big(|y_r - \frac{1}{2}| \le |y_{r+1} - \frac{1}{2}|\big)\Big|\big(y_r < \frac{1}{2} < y_{r+1}\big)\Big) \\ &\qquad = \Pr\Big(\min\{Z_1, Z_2, \ldots, Z_r\} < \min\{Z_{r+1}, Z_{r+2}, \ldots, Z_n\}\Big) = \frac{r}{n},\end{aligned}$$

where $Z_1, Z_2, \ldots, Z_n$ are independent random variables drawn from the uniform distribution on the interval $[0, 1/2]$. Analogously, we obtain

$$\Pr\Big(\big(|y_{r-1} - \frac{1}{2}| \ge |y_r - \frac{1}{2}|\big)\Big|\big(y_{r-1} < \frac{1}{2} < y_r\big)\Big) = \frac{n-r+1}{n}$$

and, finally,

$$\begin{aligned}\Pr(y_r \text{ is the best}) &= \binom{n}{r} \cdot \frac{1}{2^n} \cdot \frac{r}{n} + \binom{n}{r-1} \cdot \frac{1}{2^n} \cdot \frac{n-r+1}{n} \\ &= \frac{1}{2^n}\left[\frac{(n-1)!}{(r-1)!(n-r)!} + \frac{(n-1)!}{(r-1)!(n-r)!}\right] = \binom{n-1}{r-1} \cdot \frac{1}{2^{n-1}}.\end{aligned}$$

$\square$

Theorem 2 *If $y_1^{(t)} < y_2^{(t)} < ... < y_r^{(t)} < ... < y_t^{(t)}$ are the ranked numbers at time t, then*

$$Pr(y_r^{(t)} \text{ will be the best}) = \frac{1}{2^{n-1}} \sum_{j=0}^{n-t} \binom{n-1}{r-1+j} \binom{n-t}{j} \frac{r^j (t+1-r)^{n-t-j}}{(t+1)^{n-t}}. \qquad (1)$$

Proof : Since $n - t$ additional numbers will appear, the rank r of the number $y_r^{(t)}$ will increase by some j, where $0 \le j \le n - t$. Each number following $y_r^{(t)}$ will independently fall into one of the intervals $(0,\ y_1^{(t)})$, $(y_1^{(t)},\ y_2^{(t)})$, ..., $(y_t^{(t)},\ 1)$ with the same probability, $\frac{1}{t+1}$. Every time a number falls into one of the first r intervals, the rank of $y_r^{(t)}$ increases by 1. Therefore, the probability that after the appearance of all n numbers, the rank of $y_r^{(t)}$ will be $r+j$ is $\binom{n-t}{j}\frac{r^j(t+1-r)^{n-t-j}}{(t+1)^{n-t}}$. Thus, from Theorem 1,

$$\Pr\left(y_r^{(t)} \text{ will be the best} \mid \text{its rank is } r+j\right) = \binom{n-1}{r-1+j}\frac{1}{2^{n-1}}$$

and formula (1) follows from the law of total probability. $\square$

From now on, $\Pr(y_r^{(t)}$ will be the best) will be abbreviated to $\mathrm{P}_r^{(t)}$. Also, we denote the optimal algorithm from the set of algorithms that stop only in rounds $t, t+1, ..., n-1$, or n by $\mathcal{A}_n^{(t)}$ (i.e. such algorithms never stop before time t). We now construct an optimal stopping algorithm $\mathcal{A}_n$ using recursion. Note that $\mathcal{A}_n = \mathcal{A}_n^{(1)}$.

$\mathcal{A}_n^{(n)}$ is the algorithm that stops only in the last round, thus the stopping interval for n is $[1, n]$ and $\Pr(\mathcal{A}_n^{(n)}$ succeeds$) = \frac{1}{n}$.

Recursively, assume that for $k = t+1, t+2, \ldots, n$ we know the probabilities $\Pr(\mathcal{A}_n^{(k)}$ succeeds) and the stopping interval $[r_{n-k},\ k+1-r_{n-k}]$ in round k. The optimal algorithm $\mathcal{A}_n^{(t)}$ stops on the number $y_r^{(t)}$ in round t if and only if its rank r satisfies the inequality

$$\mathrm{P}_r^{(t)} \geq \Pr(\mathcal{A}_n^{(t+1)} \text{ succeeds}). \tag{2}$$

If inequality (2) has a solution, then the solution set, which is a symmetric interval $[r_{n-t},\ t+1-r_{n-t}]$, is the stopping interval for $\mathcal{A}_n^{(t)}$ in round t and

$$\Pr(\mathcal{A}_n^{(t)} \text{ succeeds}) = \sum_{r=r_{n-t}}^{t+1-r_{n-t}} \frac{1}{t}\mathrm{P}_r^{(t)} + \frac{2(r_{n-t}-1)}{t}\Pr(\mathcal{A}_n^{(t+1)} \text{ succeeds}). \tag{3}$$

If there is no r satisfying inequality (2), then the algorithm $\mathcal{A}_n^{(t)}$ never stops in round t and $\Pr(\mathcal{A}_n^{(t)}$ succeeds$) = \Pr(\mathcal{A}_n^{(t+1)}$ succeeds).

The example below illustrates how the optimal stopping strategy looks like for $n = 10$.

t	$10-t$	r_{10-t}	stopping interval	$Pr(\mathcal{A}_{10}^{(t)}$ succeeds)
1	9			**0.1893**
2	8			0.1893
3	7	2	$\{2\}$	0.1893
4	6			0.1858
5	5	3	$\{3\}$	0.1858
6	4	3	$[3,4]$	0.1798
7	3	3	$[3,5]$	0.1701
8	2	4	$[4,5]$	0.1585
9	1	4	$[4,6]$	0.1378
10	0	1	$[1,10]$	0.1

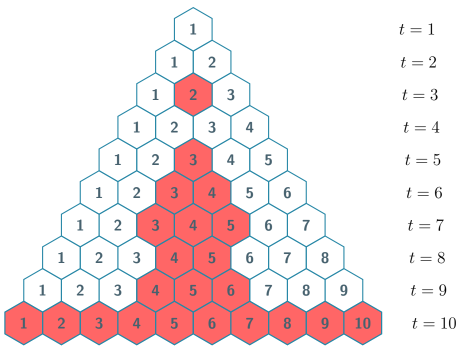

Figure 1. Stopping intervals at time t and probabilities that the algorithm $\mathcal{A}_n^{(t)}$ succeeds for $n = 10$. Red cells denote the stopping region, which is the union of the stopping intervals over all values of t.

As can be seen from the example, the stopping region for our algorithm $\mathcal{A}_n$ is rather irregular and the recursive formulas used to calculate $\Pr(\mathcal{A}_n$ succeeds) give little hope of finding a closed formula for this probability. Despite these difficulties, in the next section we will derive the asymptotic performance of the optimal algorithm $\mathcal{A}_n$.

3 Asymptotics

First, we define the algorithm $\mathcal{A}(h_n, w_n)$, which is not optimal, but has a more regular stopping region than the optimal algorithm $\mathcal{A}_n$. This will be helpful in finding a reasonable lower bound for the performance of $\mathcal{A}_n$. The stopping region of the algorithm $\mathcal{A}(h_n, w_n)$ is defined by two natural numbers h_n and w_n. This algorithm never stops before time h_n. For $t \geq h_n$, it stops on x_t if and only if x_t falls between $y^{(t-1)}_{\lceil \frac{t}{2} \rceil - w_n}$ and $y^{(t-1)}_{\lfloor \frac{t}{2} \rfloor + w_n}$, where $y^{(t-1)}_1 < y^{(t-1)}_2 < \ldots < y^{(t-1)}_{t-1}$ are the ordered numbers at time $t-1$. If this never happens, $\mathcal{A}(h_n, w_n)$ stops at x_n. Figure 2 illustrates the rectangular stopping region for the algorithm $\mathcal{A}(h_n, w_n)$. Note that $n - h_n + 1$ and $2w_n$ can be interpreted as the height and width of this stopping region, respectively.

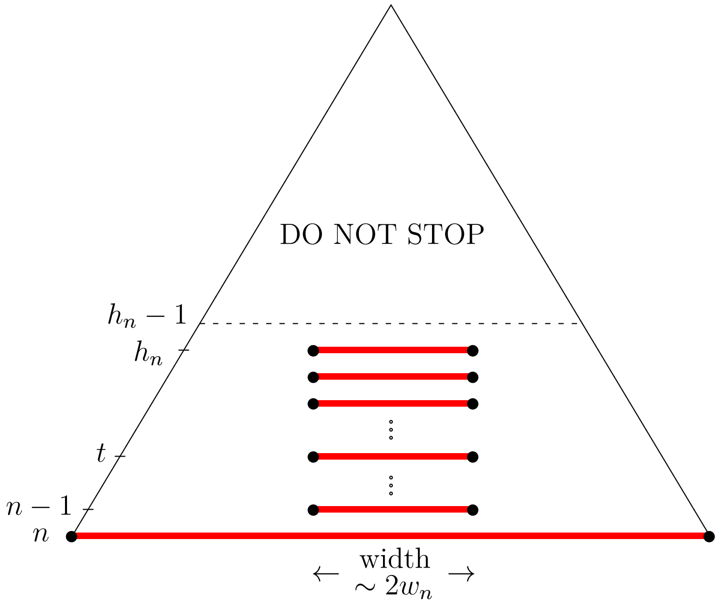

Figure 2. The stopping region for the algorithm $\mathcal{A}(h_n, w_n)$.

Using several technical lemmas, we arrive at the following result.

Theorem 3 *For any sequences h_n and w_n of natural numbers such that $h_n \leq n$ and $4w_n^2 < n$, we have*

$$Pr(\mathcal{A}(h_n, w_n) \text{ succeeds}) \geq v(h_n, w_n),$$

where $v(h_n, w_n)$ is a function such that for $w_n \xrightarrow[n\to\infty]{} \infty$

$$v(h_n, w_n) \sim \frac{h_n}{n 2^{n-1}} \binom{n-1}{\frac{n-1}{2} - w_n} \cdot \left(1 - \left(1 - \frac{2w_n}{h_n}\right)^{n-h_n}\right).$$

There exist choices of the sequences h_n and w_n, for example $w_n = \lceil n^{1/3} \rceil$ and $h_n = \lceil n(1 - \frac{\sqrt{\ln n}}{n^{1/3}}) \rceil$, for which the probability of success under the algorithm $\mathcal{A}(h_n, w_n)$ is bounded from below by a function which asymptotically behaves like $\frac{1}{\sqrt{n}}\sqrt{\frac{2}{\pi}}$. Since the optimal algorithm $\mathcal{A}_n$ is not worse, this lower bound also applies to $\mathcal{A}_n$. It remains to show that this function of n is also asymptotically an upper bound for the performance of $\mathcal{A}_n$. It can be done by considering an easier online decision problem. Suppose that a decision maker observes the ranks of n numbers which are independent realizations from the uniform distribution on the interval $[0, 1]$. The decision maker must choose one of these numbers with the aim of maximizing the probability of choosing the element which is closest to $\frac{1}{2}$. However, after all n numbers are revealed, any number can be selected, not necessarily the last one. From Theorem 1, we know that the optimal strategy is to select the number of rank r such that the binomial coefficient

$\binom{n-1}{r-1}$ takes its maximum value. This happens if $r-1 = \lfloor \frac{n-1}{2} \rfloor$ or $r-1 = \lceil \frac{n-1}{2} \rceil$. Thus

$$\Pr(x_r \text{ is the best }) = \binom{n-1}{\lfloor \frac{n-1}{2} \rfloor} \cdot \frac{1}{2^{n-1}}$$

giving the asymptotic performance of order $\frac{1}{\sqrt{n}}\sqrt{\frac{2}{\pi}}$.

This leads to the following theorem.

Theorem 4 *For the online decision problem considered here, the optimal stopping algorithm $\mathcal{A}_n$ has asymptotic performance*

$$Pr(\mathcal{A}_n \text{ succeeds}) \sim \frac{1}{\sqrt{n}}\sqrt{\frac{2}{\pi}}.$$

4 Final remarks

If the interval $[0,1]$ is replaced by the interval $[a,b]$, where $a < b$, and the goal is to stop on the element closest to the interval's midpoint, then the optimal stopping algorithm is identical to our algorithm $\mathcal{A}_n$.

How does the situation change if we sequentially observe n numbers from the interval $[0,1]$, but we are informed about the value of each number drawn? Since we now know whether the revealed number is greater or smaller than $\frac{1}{2}$, by replacing each x_k greater than $\frac{1}{2}$ by $1-x_k$, we obtain a problem equivalent to finding the maximum element of a sequence of n numbers. This problem was solved by Gilbert & Mosteller (1966) and the optimal strategy in what they called 'the full-information game' has an asymptotic probability of success approximately equal to 0.580164. On the other hand, if our aim is to minimize the expected difference between the number selected and $\frac{1}{2}$, then we should adopt another stopping algorithm from (Gilbert & Mosteller, 1966) which gives an expected difference of order $\frac{1}{n}$.

References

[1] Benevides, F. S. & Sulkowska, M. (2017). Percolation and best choice problem for powers of paths, *Journal of Applied Probabability, 54*, no. 2, 343–362.

[2] Gilbert, J. P. & Mosteller, F. (1966). Recognizing the maximum of a sequence, *Journal of the American Statistical Association, 61*, 35–73.

[3] Kubicki, G. & Morayne, M. (2005). Graph-theoretic generalization of the secretary problem: the directed path case, *SIAM Journal on Discrete Mathematics, 19*, no. 3, 622–632.

[4] Lindley, D. V. (1961). Dynamic programming and decision theory, *Applied Statistics, 10*, 39–51.

[5] Morayne, M. (1998). Partial-order analogue of the secretary problem; the binary tree case, *Discrete Mathematics, 184*, 165–181.

[6] Preater, J. (1999). The best-choice problem for partially ordered objects, *Operations Research Letters, 25*, 187–190.

[7] Rogerson, P. A. (1987). Probabilities of choosing applicants of arbitrary rank in the secretary problem, *Journal of Applied Probabability, 224*, no. 2, 527–533.