ESAIM: PS 29 (2025) 184–203
https://doi.org/10.1051/ps/2025002

ESAIM: Probability and Statistics
www.esaim-ps.org

BROADCASTING IN RANDOM RECURSIVE DAGS

Simon Briend[1], Luc Devroye[2] and Gábor Lugosi[3,*]

Abstract. A uniform k-DAG generalizes the uniform random recursive tree by picking k parents uniformly at random from the existing nodes. It starts with k "roots". Each of the k roots is assigned a bit. These bits are propagated by a noisy channel. The parents' bits are flipped with probability p, and a majority vote is taken. When all nodes have received their bits, the k-DAG is shown without identifying the roots. The goal is to estimate the majority bit among the roots. We identify the threshold for p as a function of k below which the majority rule among all nodes yields an error $c + o(1)$ with $c < 1/2$. Above the threshold the majority rule errs with probability $1/2 + o(1)$.

Mathematics Subject Classification. 60C05, 05C80, 60C05.

Received February 24, 2024. Accepted February 3, 2025.

1. Introduction

The interest in network analysis has been growing, in part due to its use in communication technologies, social network studies, and biology, see Coolen *et al.* [1]. The problem we study here is the one of broadcasting on random graphs. We study the setting where a bit propagates with noise and we want to infer the value of the original bit. The question is not if and how the information propagates, but if there is a signal propagating on the graph, or only noise. Variations of this binary classification problem have been studied. For example, in the root-bit estimation problem, the root of a tree has a bit 0 or 1. The value of this bit propagates from the root to the leafs, and at each propagation from a vertex to the next it mutates (flips the bit) with probability p. One can try to infer the root's bit value from observing all the bits of the graph or only the leaf bits. This question was first formulated in Evans *et al.* [2] on general trees, where it was shown that root bit reconstruction is possible depending upon a condition on the branching number. More recently, the case of random recursive trees (Addario-Berry *et al.* [3], Desmarais *et al.* [4]) has been studied. Other variations of these problems on trees include looking at asymmetric flip probabilities (Sly [5]), non-binary vertex values (Mossel [6]) and robustness to perturbation (Janson and Mossel [7]). We refer the reader to Mossel [8] for a survey of reconstruction problems on trees. Many problems are described by more general graphs rather than trees. The original broadcasting question has been studied on deterministic graphs (Harutyunyan and Li [9]) and Harary graphs (Bhabak *et al.* [10], for example). We are interested in the problem of noisy propagation in the spirit of the root-bit reconstruction (Evans *et al.* [2]), but on a class of random graphs that we call k-DAG (for directed acyclic graph). A similar

Keywords and phrases: Broadcasting problem, random recursive dags, network archaeology.

[1] Department of Mathematics, Unidistance Suisse, 3900, Brigue, Switzerland.
[2] School of Computer Science McGill University Montreal, Canada.
[3] Department of Economics and Business, Barcelona School of Economics, Pompeu Fabra University, ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.

* Corresponding author: gabor.lugosi@gmail.com

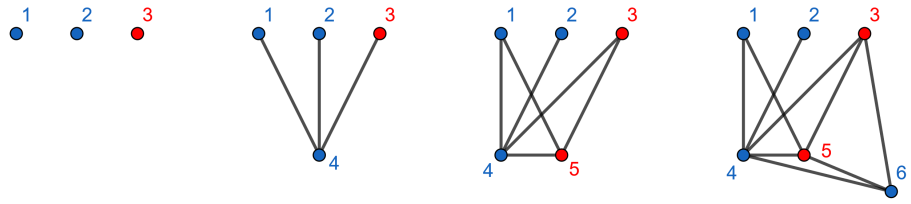

FIGURE 1. A realisation of the process up to time 6, for $k = 3$, starting with $R_3 = 1/3$.

problem – for a different class of random DAGs – has been studied in Makur *et al.* [11]. In a related probelm, Antunović *et al.* [12] studied the case of the preferential attachment model, where initial nodes have a color and the color of the new nodes is a function of the colors of their neighbors.

Since we track the proportion of zero bits in our graph, we cast the process as an urn model. A similar reformulation was already done in Addario-Berry *et al.* [3] to study majority voting properties of broadcasting on random recursive trees. The proportion of zero bits and the bit assignment procedure can be viewed as random processes with reinforcement. A review of results can be found in Pemantle [13] and is extensively used, alongside results of non-convergence found in Pemantle [14]. As in Addario-Berry *et al.* [3], we make ample use of the properties of Pólya urns (Janson [15], Knape and Neininger [16], Wei [17]). Variations of the Pólya urn model that are useful for our analysis include an increase of the number of colors over time (Bertoin [18]), the selection of multiple balls in each draw (Kuba and Mahmoud [19]), and randomization in the color of the new ball (Janson [20], Zhang [21]). We note, in particular, the multi-ball draw with a linear randomized replacement rule of Crimaldi *et al.* [22]. In the present paper, we consider multi-ball draws, but with non-linear randomized replacement.

The paper is organized as follows. After introducing the mathematical model in Section 1.1, in Section 1.2 we present the main result of the paper (Thm. 1.1) that shows that there are three different regimes of the value of the mutation probability that characterize the asymptotic behavior of the majority rule. In Section 2 we discuss the three regimes of p. In Section 3 we establish convergence properties of the global proportion of both bit values assigned to vertices and in Section 4 we finish the proof of Theorem 1.1 by studying the probability of error in all three regimes. Finally, in Section 5 we establish a lower bound for the probability of error that holds uniformly for all mutation probabilities. We conclude the paper by discussing avenues for further research.

1.1. The model

We start by describing the evolution of the uniform random recursive k-DAG and the assigned bit values that we represent by two colors; red and blue.

Let us fix an odd integer $k > 0$. The growth process is initiated at time k. At time k, the graph consists of k isolated vertices. A fraction R_k are red and a fraction $B_k = 1 - R_k$ are blue. We set $R_1 = \cdots = R_k$ and $B_1 = \cdots = B_k$. The network is grown recursively by adding a new colored vertex and at most k edges at each time step. At time n, a new vertex n connects to a sample of k vertices chosen uniformly at random with replacement among the $n-1$ previous vertices. (Possible multiple edges are collapsed into one so that the graph remains simple.) The color of vertex n is determined by the following randomized rule:

- the colors of the k selected parents are observed;
- each of these is independently flipped with probability p (if a parent is selected more than once, its color is flipped independently for each selection);
- the color of vertex n is chosen according to the majority vote of the flipped parent colors (*i.e.*, there are exactly k votes).

If one is only interested in the evolution of the proportion of red and blue vertices (but not the structure of the graph), one may equivalently describe it by an urn model with multiple draws and random (nonlinear)

replacement. The urn process is defined as follows. The urn is initialized with an odd number k of balls, a fraction R_k being red and $B_k = 1 - R_k$ blue. At each time $n \geq k+1$,

- k balls are drawn from the urn, uniformly at random with replacement, and returned to the urn;
- the color of each drawn ball is flipped with probability p (*i.e.*, a drawn ball that is red is observed as blue with probability p);
- a new ball is added to the urn, whose color is chosen as the majority of the k observed colors.

In the root-bit estimation problem considered here, the statistician has access to an unlabelled and undirected version of the graph at time n, along with the vertex colors. The goal of the statistician is to estimate the colors assigned to the k roots. More precisely, based on the observed graph, one would like to guess the majority color at time k.

This problem has been studied in depth by [3] in the case when $k = 1$, that is, when the produced graph is a uniform random recursive tree. Two types of methods for root-bit estimation were studied in [3]. One is based on first trying to localize the root of the tree–disregarding the vertex colors. If one finds a vertex that is close to the root, one may use the color of that vertex as a guess for the root color. Such a vertex is the centroid of the tree. Indeed, it is shown in [3] that the color of the centroid is a nearly optimal estimate of the root color. In the same paper, the majority rule is also studied. This method disregards the structure of the tree and guesses the root color by taking a majority vote among all vertices. It is shown that for small mutation probabilities the majority rule is also nearly optimal.

In the more general problem considered in this paper, one may also try to estimate the colors of the k roots by finding nearby vertices. However, this problem becomes significantly more challenging as the k-DAG does not have a natural centroid. The interested reader is referred to the recent paper of Briend, Calvillo, and Lugosi [23] on root finding in random k-DAGs. Instead of pursuing this direction, we focus on the majority vote. More precisely, we are interested in characterizing the values of the mutation probability p such that the asymptotic probability of error is strictly better than random guessing.

At time n, the majority vote, denoted by b_n^{maj}, is defined as follows:

$$b_n^{maj} = \begin{cases} \text{“R” (red) if } R_n > 1/2 \\ \text{“B”(blue) if } R_n < 1/2 \\ \text{Ber}(1/2) \text{ if } R_n = 1/2 \text{ (a random coin flip).} \end{cases}$$

We define the probability of error by

$$R^{maj}(n,p) = \mathbb{P}\left\{ b_n^{maj} \neq b_k^{maj} \right\}.$$

Note that b_k^{maj} depends on the initial vertex colors that are assumed to be chosen arbitrarily and fixed. Hence, $R^{maj}(n,p)$ is also a function of the initial proportion R_k but to avoid heavy notation, we supress this dependence.

1.2. Related results and our contribution

Our broadcasting model is an extension of the broadcasting on uniform random recursive trees that was extensively studied in Addario-Berry *et al.* [3]. In this problem, $k = 1$ and the only parameter is p, the mutation probability. For the majority voting rule, they prove the following:

(i) There exists a constant $c > 0$ such that

$$\limsup_{n\to\infty} R^{maj}(n,p) \leq cp.$$

(ii) For all $p \in (0, 1/2]$,

$$\lim_{n\to\infty} R_n = \frac{1}{2} \quad \text{with probability one.}$$

(iii) For $p \in [0, 1/4)$

$$\limsup_{n\to\infty} R^{maj}(n, p) < \frac{1}{2}.$$

(iv) For $p \in [1/4, 1/2]$

$$\limsup_{n\to\infty} R^{maj}(n, p) = \frac{1}{2}.$$

In other words, even though the proportion of vertices that have the same color as the root converges to $1/2$, for mutation probabilities smaller than $1/4$, sufficient information is preserved about the root color for the majority vote to work with a nontrivial probability.

We generalize these results to k-DAGS and characterize the values of p for which majority voting outperforms random guessing. In order to state the main result of the paper, we introduce some notation.

For any odd positive integer k, let

$$\alpha_k := \frac{1}{2^{k-2}} \sum_{i>k/2}^{k} \binom{k}{i} (i - k/2) = 4\mathbb{E}\left[\left(\mathrm{Bin}(k, 1/2) - \frac{k}{2}\right)_+\right]. \tag{1.1}$$

For example, $\alpha_1 = 1$, $\alpha_3 = 3/2$, and by a simple application of the central limit theorem, for large k,

$$\alpha_k \sim \sqrt{\frac{2k}{\pi}}. \tag{1.2}$$

In the statement of our main theorem, we assume, without loss of generality, that initially red vertices are in majority, that is, $R_k > 1/2$.

Theorem 1.1. *Let k be an odd positive integer and consider the broadcasting process on a random k-DAG described above. Assume that initially $R_k > 1/2$.*

(i) If $p < \frac{1}{2} - \frac{1}{2\alpha_k}$, then there exist $\beta_1 \in (0, 1/2)$ and $\beta_2 = 1 - \beta_1$ (whose value only depends on k but not on the initial color configuration) such that

$$\mathbb{P}\{R_n \to \beta_1\} + \mathbb{P}\{R_n \to \beta_2\} = 1 \quad \text{and} \quad \mathbb{P}\{R_n \to \beta_1\} < \mathbb{P}\{R_n \to \beta_2\}.$$

In particular, regardless of the initial value of R_k,

$$\limsup_{n\to\infty} R^{maj}(n, p) < \frac{1}{2}.$$

(ii) If $\frac{1}{2} - \frac{1}{2\alpha_k} \le p < \frac{1}{2} - \frac{1}{4\alpha_k}$, then $R_n \to 1/2$ a.s. and

$$\limsup_{n\to\infty} R^{maj}(n, p) < \frac{1}{2}.$$

(iii) If $\frac{1}{2} - \frac{1}{4\alpha_k} \le p \le \frac{1}{2}$ then $R_n \to 1/2$ a.s. and

$$\lim_{n\to\infty} R^{maj}(n, p) = \frac{1}{2}.$$

Theorem 1.1 shows that for all $k \geq 3$, there are three regimes of the value of the mutation probability. In the low-rate-of-mutation regime the proportion of red balls almost surely converges to one of two numbers, both different from $1/2$. Moreover, the limiting proportion is positively correlated with the initial value. In the intermediate phase, the vertex colors are asymptotically balanced, but there is enough signal for the majority vote to perform strictly better than random guessing. Finally, in the high-rate-of-mutation regime, the majority vote is equivalent to a coin toss, at least asymptotically.

Note that for $k = 1$, $\alpha_1 = 1$, so $1/2 - 1/(2\alpha_1) = 0$, and therefore the low-rate-of-mutation regime does not exist. Of course, this is in accordance with the results of [3] cited above.

On the other hand, for $k = 3$ the two thresholds are $1/2 - 1/(2\alpha_3) = 1/6$ and $1/2 - 1/(4\alpha_1) = 1/3$, meaning that from $k = 3$ onward the three different regimes can be observed. For large k, both threshold values are of the order $1/2 - \Theta(1/\sqrt{k})$.

A closely related model has been studied by Makur *et al.* [11]. They study different random DAGS, where important parameters are the number of vertices at distance k from the root and the indegree of vertices. They also suppose that the position of the root vertex is known. Two rules of root bit estimation are studied: a noisy majority rule and the NAND rule. Makur *et al.* [11] show that if the number of vertices of depth k is $\Omega(\log(k))$ then there is a threshold on the mutation probability for which root bit estimation is possible.

As a first step, we study the convergence of the proportion of red balls. To this end, it suffices to study the generalized urn process defined above. We mention here that Crimaldi *et al.* [22] study a somewhat related urn process, though with linear replacement rules.

In order to avoid unessential complications caused by breaking ties, we only consider odd values of k. The same techniques allow one to analyze even values of k. In such cases, in the event of a tie among the k observed colors, one may choose the color of the new vertex at random.

2. Different regimes

We start by studying the evolution of R_n. Let us denote by c_n the color of the nth vertex appearing in the graph. After possible mutation, each edge connecting vertex $n+1$ to an older vertex carries a signal. This signal is red with probability

$$f(R_n) := (1-p)R_n + p(1-R_n) = (1-2p)R_n + p.$$

Because the k parents are chosen independently and that the color is chosen by the majority,

$$\mathbb{P}\{c_{n+1} = R\} \;=\; \mathbb{P}\{\mathrm{Bin}(k, f(R_n)) \geq k/2\}, \tag{2.1}$$

where, conditionally on R_n, $\mathrm{Bin}(k, f(R_n))$ is a binomial random variable. Moreover, we know that the number of red vertices evolves as $(n+1)R_{n+1} = nR_n + \mathbb{1}(c_{n+1} = R)$, where $\mathbb{1}$ is the indicator function. We rewrite this as

$$R_{n+1} = R_n + \frac{\mathbb{1}(c_{n+1} = R) - R_n}{n+1}. \tag{2.2}$$

A key to understanding R_n is then to study the random variable $\mathbb{1}(c_{n+1} = R) - R_n$. We define, for $t \in [0,1]$,

$$g(t) \;:=\; \mathbb{E}\left[\mathbb{1}(c_{n+1} = R) - R_n | R_n = t\right] \;= \mathbb{P}\{\mathrm{Bin}(k, f(t)) > k/2\} - t. \tag{2.3}$$

The evolution of R_n is entirely determined by the function g. Observe first that for any $t \in [0,1]$, $f(1-t) = 1 - f(t)$. Also, since k is odd,

$$\mathbb{P}\{\mathrm{Bin}(k, 1-f(t)) > k/2\} = 1 - \mathbb{P}\{\mathrm{Bin}(k, f(t)) > k/2\},$$

which implies that

$$g(1-t) = -g(t).$$

The extremal values of g are

$$g(0) = \mathbb{P}\{\mathrm{Bin}(k,p) > k/2\} > 0,$$

and

$$g(1) = \mathbb{P}\{\mathrm{Bin}(k,1-p) > k/2\} - 1 < 0.$$

Since g is continuous, the polynomial g has at least one root. From the symmetry property we have $g(1/2) = -g(1-1/2) = -g(1/2)$, so $g(1/2) = 0$. Moreover, we obtain

$$g'(1/2) = \frac{1-2p}{2^{k-2}} \sum_{i>k/2}^{k} \binom{k}{i} (i - k/2) - 1.$$

Recalling the definition of α_k from (1.1), we have $g'(1/2) = (1-2p)\alpha_k - 1$. Since $\alpha_k \geq 1$, we conclude:

$$g'\left(\frac{1}{2}\right) \begin{cases} < 0 & \text{if } p > \frac{1}{2} - \frac{1}{2\alpha_k}, \\ \\ > 0 & \text{if } p < \frac{1}{2} - \frac{1}{2\alpha_k}. \end{cases}$$

To understand the other potential zeros of g, let us study its convexity.

Lemma 2.1. *The function g is strictly convex on $(0,1/2)$ and strictly concave on $(1/2,1)$.*

Proof. We may use the elementary identities

$$\mathbb{P}\left\{\mathrm{Bin}(k,x) \geq \frac{k+1}{2}\right\} = \mathbb{P}\left\{\mathrm{Beta}\left(\frac{k+1}{2}, \frac{k+1}{2}\right) < x\right\}, \tag{2.4}$$

where $\mathrm{Beta}(a,b)$ is a $\mathrm{beta}(a,b)$ random variable. Hence,

$$g(t) = \int_0^{f(t)} (x(1-x))^{\frac{k-1}{2}} \frac{\Gamma(k+1)}{\Gamma^2\left(\frac{k+1}{2}\right)} \mathrm{d}x \quad - t,$$

and therefore

$$g'(t) = (1-2p)\left(f(t)(1-f(t))\right)^{\frac{k-1}{2}} \frac{\Gamma(k+1)}{\Gamma^2\left(\frac{k+1}{2}\right)} - 1. \tag{2.5}$$

Since $f(t)(1-f(t)) = -(1-2p)t(t-1) + p(1-p)$ is increasing for $t \in (0,1/2)$ and decreasing for $t \in (1/2,1)$, g is strictly convex on $(0,1/2)$ and strictly concave on $(1/2,1)$. □

In summary, if $p > \frac{1}{2} - \frac{1}{2\alpha_k}$, then $g'(1/2) < 0$, and thus g is monotonically decreasing on $[0,1]$ and has only one zero in $[0,1]$. If $g'(1/2) = 0$, then there is only one zero (at $1/2$) and g exhibits an inflection point at $1/2$. If $p < \frac{1}{2} - \frac{1}{2\alpha_k}$, then $g'(1/2) > 0$ and thus g has exactly one zero in $(0,1/2)$ and by symmetry, it also has one zero on $(1/2,1)$. We denote these zeros by β_1 and β_2, respectively.

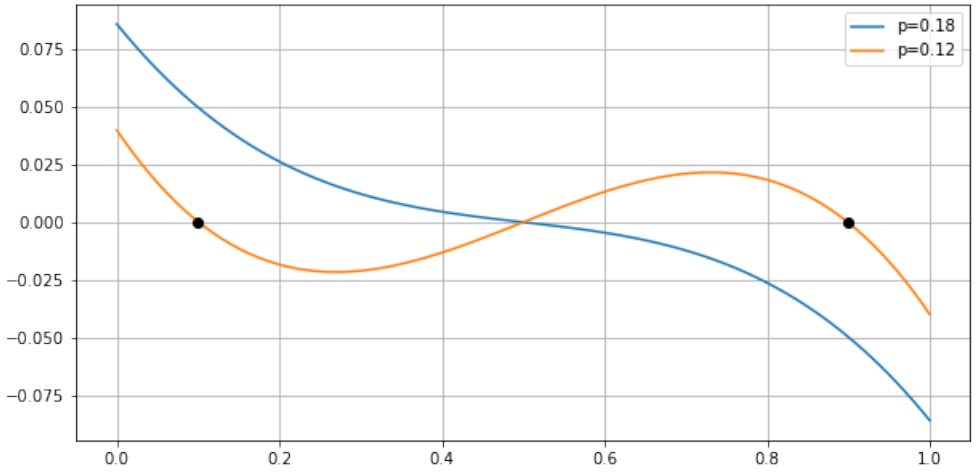

FIGURE 2. g as a function of $t \in [0,1]$, for $k = 3$, with the choices $p = 0.18 > 1/6$ and $p = 0.12 < 1/6$.

Figure 2 shows two examples of the graph of the function g.

It is also interesting to know the position of β_1 (recall that $\beta_2 = 1 - \beta_1$). First, we note that for fixed k, if p tends to the threshold $1 - 1/(2\alpha_k)$, then β_1 tends to $1/2$. In the following lemma we study the case when p is far enough from the threshold, that is, when $p \leq \frac{1}{2} - \frac{C}{2\alpha_k}$, for a sufficiently large constant C.

Lemma 2.2. *Let* $p \leq \frac{1}{2} - \frac{C}{2\alpha_k}$ *for* $C \geq \sqrt{\frac{8\log(2)}{\pi}}$. *Then*

$$\beta_1 \leq \exp\left(-\frac{k(1-2p)^2}{8}\right).$$

Proof. β_1 is the smallest root of $g(t)$ and since $g(0) > 0$, its smallest root is smaller than the smallest root of any upper bound of g. On the other hand,

$$g(t) = \mathbb{P}\left\{\mathrm{Bin}(k, f(t)) \geq \frac{k}{2}\right\} - t \leq \exp\left(-2k\left(\frac{1}{2} - f(t)\right)^2\right) = \exp\left(-2k(1-2p)^2\left(\frac{1}{2} - t\right)^2\right) - t.$$

Thus, β_1 is at most the first zero of $b(t) := \exp\left(c_1\left(\frac{1}{2} - t\right)^2\right) - t$, for $c_1 = 2k(1-2p)^2$. Since $b(0) > 0$, if for some t^*, $b(t^*) \leq 0$ then the first zero of b and therefore β_1 is at most t^*. Taking $t^* = \mathrm{e}^{-c_1/16}$, we have

$$b(t^*) \leq 0 \iff \left(\frac{1}{2} - \mathrm{e}^{-c_1/16}\right)^2 \geq 1/16 \iff c_1 \geq 32\log(2).$$

From (1.2) and the expression of c_1, we have that by taking $C \geq \sqrt{\frac{8\log(2)}{\pi}}$,

$$2k(1-2p)^2 \geq 32\log(2).$$

This shows that for $p \leq \frac{1}{2} - \frac{C}{2\alpha_k}$, we have

$$\beta_1 \leq \exp\left(-\frac{k(1-2p)^2}{8}\right).$$

□

3. Convergence of the proportion of red balls

In order to analyze the probability of error of the majority vote, first we establish convergence properties of R_n. The two possible regimes of g suggest that there are two distinct regimes of the evolution of R_n. From (2.2) we note that R_n has a positive drift if $g(R_n)$ is positive, and a negative drift otherwise. This suggests that in the high-rate-of-mutation regime, R_n converges to $1/2$ and in the low-rate-of-mutation regime it converges to either β_1 or β_2. The following section investigates this intuition, using Lemma 2.6 and Corollary 2.7 from Pemantle [13] about the convergence of reinforced random processes. We state them here.

Lemma 3.1 (Pemantle [13]). *Let $\{X_n;\ n \geq 0\}$ be a stochastic process in $\mathbb{R}$ adapted to a filtration $\{\mathcal{F}_n\}$. Suppose that X_n satisfies*

$$X_{n+1} - X_n = \frac{1}{n}\left(F(X_n) + \xi_{n+1} + E_n\right),$$

where F is a function on $\mathbb{R}$, $\mathbb{E}\left[\xi_{n+1} \mid \mathcal{F}_n\right] = 0$ and the remainder term E_n goes to 0 and satisfies $\sum_{n=1}^{\infty} n^{-1}|E_n| < \infty$ almost surely. Suppose that F is bounded and that $\mathbb{E}\left[\xi_{n+1}^2 \mid \mathcal{F}_n\right] < K$ for some finite constant K. If for $a_0 < x < b_0$, $F(x) \geq \delta$ for some $\delta > 0$, then for any $[a,b] \subset (a_0, b_0)$ the process $\{X_n\}$ visits $[a,b]$ finitely many times almost surely. The same result holds if $F(x) \leq -\delta$.

Corollary 3.2 (Pemantle [13]). *If F is continuous on $\mathbb{R}$, then X_n converges almost surely to the zero set of F.*

3.1. The high-rate-of-mutation regime $\left(\frac{1}{2} - \frac{1}{2\alpha_k} \leq p \leq \frac{1}{2}\right)$

Rewrite (2.2) as

$$\begin{aligned} R_{n+1} - R_n =& \frac{1}{n+1}\left(\mathbb{P}\left\{\mathrm{Bin}(k, f(R_n)) \geq \frac{k}{2}\right\} - R_n\right) \\ &+ \frac{1}{n+1}\left(\mathbb{1}(c_{n+1} = R) - \mathbb{P}\left\{\mathrm{Bin}(k, f(R_n)) \geq \frac{k}{2}\right\}\right). \end{aligned}$$

Since $g(R_n) = \mathbb{P}\left\{\mathrm{Bin}(k, f(R_n)) \geq k/2\right\} - R_n$, we see that

$$R_{n+1} - R_n = \frac{g(R_n) + \xi_{n+1}}{n+1}, \tag{3.1}$$

where $\xi_{n+1} = \mathbb{1}(c_{n+1} = R) - \mathbb{P}\left\{\mathrm{Bin}(k, f(R_n)) \geq k/2\right\}$. Because g is continuous and $\mathbb{E}\left[\xi_{n+1}|R_n\right] = 0$, our process satisfies all the requirements for Corollary 3.2. It states that R_n converges almost surely to the set of zeros of g. In this regime, this implies that R_n converges to $1/2$ almost surely.

3.2. The low-rate-of-mutation regime $\left(0 < p < \frac{1}{2} - \frac{1}{2\alpha_k}\right)$

In this regime, the requirements of Corollary 3.2 are still met. So R_n converges almost surely to the set of zeros of g, which is $\{\beta_1, 1/2, \beta_2\}$. We first show that R_n does not converge to $1/2$: $1/2$ seems to be an unstable

equilibrium point, since the drift in the process has a tendency to pull R_n away from 1/2. We state Theorem 2.9 from Pemantle [13] here:

Theorem 3.3 (Pemantle [13])**.** *Suppose* $\{X_n\}$ *satisfies the conditions of Lemma 3.1 and that for some* $w \in (0,1)$ *and* $\epsilon > 0$, *signF*$(x) = sign(x - w)$ *for all* $x \in (w - \epsilon, w + \epsilon)$. *For* $\xi^+_{n+1} = \max(\xi_{n+1}, 0)$ *and* $\xi^-_{n+1} = \max(-\xi_{n+1}, 0)$, *suppose that* $\mathbb{E}[\xi^+_{n+1} \mid \mathcal{F}_n]$ *and* $\mathbb{E}[\xi^-_{n+1} \mid \mathcal{F}_n]$ *are bounded above and below by positive numbers when* $X_n \in (w - \epsilon, w + \epsilon)$. *Then* $\mathbb{P}\{X_n \to w\} = 0$.

Corollary 3.4. *In the low-rate-of-mutation regime, almost surely the process* R_n *does not converge to* $\frac{1}{2}$.

Proof. Since the conditional distribution of ξ_{n+1}, given $R_n = 1/2$ does not depend on n, it is immediate that

$$c < \mathbb{E}[\xi^+_{n+1} | R_n = 1/2] < 1,$$

and

$$c < \mathbb{E}[\xi^-_{n+1} | R_n = 1/2] < 1,$$

for some $c > 0$ that does not depend on n. Since $t \mapsto \mathbb{E}[\xi^{\pm}_{n+1} | R_n = t]$ is continuous and does not depend on n, there exists $\epsilon > 0$ such that for all $t \in (1/2 - \epsilon, 1/2 + \epsilon)$,

$$\frac{c}{2} < \mathbb{E}[\xi^{\pm}_{n+1} | R_n = t] < 2.$$

Moreover, g is negative on $(1/2 - \epsilon, 1/2)$ and positive on $(1/2, 1/2 + \epsilon)$. So, by Theorem 3.3,

$$\mathbb{P}\left\{R_n \mapsto \frac{1}{2}\right\} = 0.$$

□

Corollary 3.5. *In the low-rate-of-mutation regime, the process* R_n *converges almost surely, either to* β_1 *or to* β_2, *that is,*

$$\mathbb{P}\{R_n \to \beta_1\} + \mathbb{P}\{R_n \to \beta_2\} = 1.$$

Proof. It suffices to check that R_n converges to β_1 or β_2 and does not oscillate between them. Between 1/2 and β_2 the function g is positive, so there exists $1/2 < a_0 < a_1 < \beta_2$ and $\delta > 0$ such that for all $t \in (a_0, a_1)$, $g(t) > \delta$.

Lemma 3.1 shows that R_n visits any set $[a, b] \subset (a_0, a_1)$ finitely often almost surely. Because the step sizes of R_n are of order $1/n$, if R_n visits $[a, b]$ finitely many times, it crosses it finitely many times. Indeed, for n large enough it cannot cross $[a, b]$ without visiting $[a, b]$. Since R_n converges almost surely to the set $\{\beta_1, \beta_2\}$, but R_n crosses the set (a_0, a_1) finitely many times, we see that R_n converges almost surely either to β_1 or β_2, as claimed. □

4. Is majority voting better than random guessing?

As a first step of understanding if majority voting is better than random guessing, we prove the following lemma. It gives an equivalent condition to the success of majority voting in terms of the first time the majority flips.

Lemma 4.1. *Let* T *denote the random time at which the majority flips for the first time, that is,*

$$T = \min\left\{n \in \mathbb{N} : \ b_n^{maj} \neq b_k^{maj}\right\}.$$

Then $\limsup_{n\to\infty} R^{maj}(n,p) < 1/2$ *if and only if* $\mathbb{P}\{T=+\infty\} > 0$.

Proof. From the definition of $R^{maj}(n,p)$,

$$\limsup_{n\to\infty} R^{maj}(n,p) = 1 - \liminf_{n\to\infty} \mathbb{P}\left\{b_n^{maj} = b_k^{maj}\right\}.$$

Fix a positive ϵ. Since the sequence of events $\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\}$ is decreasing, and $\{T=+\infty\} = \{\forall i > k;\ b_i^{maj} = b_k^{maj}\}$, by continuity of measure we can choose n such that

$$\mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\} \le \mathbb{P}\{T=+\infty\} + \epsilon.$$

For $N \ge n+1$, we have

$$\begin{aligned}\mathbb{P}\left\{b_N^{maj} = b_k^{maj}\right\} = \mathbb{P}\left\{b_N^{maj} = b_k^{maj} \text{ and } \forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\} \\ + \mathbb{P}\left\{b_N^{maj} = b_k^{maj} \text{ and } \exists i \in [n] :\ b_i^{maj} \neq b_k^{maj}\right\}.\end{aligned} \tag{4.1}$$

The second term on the right-hand side decomposes as

$$\begin{aligned}&\mathbb{P}\left\{b_N^{maj} = b_k^{maj} \text{ and } \exists i \in [n] :\ b_i^{maj} \neq b_k^{maj}\right\} \\ &\quad = \left(1 - \mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\}\right) \mathbb{P}\left\{b_N^{maj} = b_k^{maj} \;\middle|\; \exists i \in [n] :\ b_i^{maj} \neq b_k^{maj}\right\}.\end{aligned}$$

From the definition of our process, if $R_i = 1/2$, then, conditionally on this event, the distribution of R_N for $N > i$ is symmetric. Therefore

$$\mathbb{P}\left\{b_N^{maj} = b_k^{maj} \;\middle|\; \exists i \in [n] :\ b_i^{maj} \neq b_k^{maj}\right\} = \frac{1}{2}. \tag{4.2}$$

Plugging this into (4.1) yields

$$\begin{aligned}\mathbb{P}\left\{b_N^{maj} = b_k^{maj}\right\} = \mathbb{P}\left\{b_N^{maj} = b_k^{maj} \ \cap\ \forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\} \\ + \frac{1}{2}\left(1 - \mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\}\right),\end{aligned} \tag{4.3}$$

The first term of the right-hand side is bounded from below by $\mathbb{P}\{T=+\infty\}$, which transforms (4.3) into

$$\mathbb{P}\left\{b_N^{maj} = b_k^{maj}\right\} \ge \frac{1}{2} + \mathbb{P}\{T=+\infty\} - \frac{1}{2}\mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\}.$$

Taking the limit on N and recalling the choice of n gives

$$\liminf_{N\to\infty} \mathbb{P}\left\{R_N > \frac{1}{2}\right\} \ge \frac{1}{2} + \frac{1}{2}\mathbb{P}\{T=+\infty\} - \frac{\epsilon}{2}.$$

Since the above holds for any ϵ, if $\mathbb{P}\{T=+\infty\} > 0$ then $\liminf_{N\to\infty} \mathbb{P}\left\{b_N^{maj} = b_k^{maj}\right\} > 1/2$. This proves the "if" direction of the statement.

On the other hand, from (4.3),

$$\mathbb{P}\left\{b_N^{maj} = b_k^{maj}\right\} \leq \mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\} + \frac{1}{2}\left(1 - \mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\}\right).$$

Taking the limit on N and recalling the choice of n yields

$$\begin{aligned}\liminf_{N\to\infty}\ \mathbb{P}\left\{b_N^{maj} = b_k^{maj}\right\} &\leq \frac{1}{2} + \frac{1}{2}\mathbb{P}\left\{\forall i \in [n] :\ b_i^{maj} = b_k^{maj}\right\} \\ &\leq \frac{1}{2} + \frac{1}{2}\mathbb{P}\left\{T = +\infty\right\} + \frac{\epsilon}{2}.\end{aligned}$$

As this holds for any positive ϵ, if $\liminf_{N\to\infty} \mathbb{P}\left\{b_N^{maj} = b_k^{maj}\right\} > 1/2$, then $\mathbb{P}\left\{T = +\infty\right\} > 0$. This concludes the proof. □

Lemma 4.2. *If*

$$\limsup_{n\to\infty} R^{maj}(n,p) \geq \frac{1}{2},$$

then

$$\lim_{n\to\infty} R^{maj}(n,p) = \frac{1}{2}.$$

Proof. If $\limsup_{n\to\infty} R^{maj}(n,p) \geq \frac{1}{2}$ then Lemma 4.1 shows that T is almost surely finite. But since

$$\mathbb{P}\left\{b_n^{maj} \neq b_k^{maj} \mid T \leq n\right\} = \frac{1}{2},$$

this implies

$$\mathbb{P}\left\{b_n^{maj} \neq b_k^{maj},\ T \leq n\right\} = \frac{1}{2}\mathbb{P}\left\{T \leq n\right\}.$$

Moreover, since T is finite almost surely, $\lim_{n\to\infty} \mathbb{P}\left\{T \leq n\right\} = 1$ and by the continuity of measure,

$$\lim_n \mathbb{P}\left\{b_n^{maj} \neq b_k^{maj},\ T \leq n\right\} = \mathbb{P}\left\{b_n^{maj} \neq b_k^{maj}\right\}.$$

This concludes the proof of the the lemma. □

4.1. The low-rate-of-mutation regime $\left(0 < p < \frac{1}{2} - \frac{1}{2\alpha_k}\right)$

As explained in Section 3.2, if $p < \frac{1}{2} - \frac{1}{2\alpha_k}$, then R_n converges to either β_1 or β_2. Next, we show that if $R_1 > 1/2$, then R_n is more likely to converge to β_2 than to β_1. To do so, recall (2.2) and write it as

$$R_{n+1} = \frac{n}{n+1}R_n + \frac{1}{n+1}B_n(g(R_n) + R_n),$$

where the B_n are independent Bernoulli random variables. We fix $\tau \in (1/2, \beta_2)$. From the analysis of g we know that $g(\tau) > 0$. Since $g(t) + t = \mathbb{P}\{\mathrm{Bin}(k, f(t)) \geq k/2\}$ and f is increasing, for all $t \geq \tau$,

$$g(t) + t \geq g(\tau) + \tau.$$

Fix a positive integer N and introduce the mapping

$$t \mapsto h(t) : \begin{cases} h(t) = 1/2 \text{ if } t < \tau \\ \\ h(t) = g(\tau) + \tau \text{ otherwise.} \end{cases}$$

Then define $D_k = 1$. For $n \geq k$, let

$$D_{n+1} = \frac{n}{n+1} D_n + \frac{1}{n+1} B'_n \left(h(D_n)\right),$$

where B'_n are independent Bernoulli random variables. From the definition of the process (D_n), on the event $\{D_n \geq \tau,\ \forall n \geq 1\}$

$$nD_n \geq D_k + \mathrm{Bin}(n-k, g(\tau) + \tau).$$

Hence, by the union bound and Hoeffding's inequality,

$$\mathbb{P}\{\exists i \geq N :\ D_i \leq \tau \mid \forall n \in [k, N] :\ D_n \geq \tau\} \leq \sum_{i \geq N} \mathbb{P}\{\mathrm{Bin}(i-k, g(\tau)+\tau) \leq i\tau\} \leq \frac{2\mathrm{e}^{-(N-k)g(\tau)^2}}{1 - \mathrm{e}^{-2g(\tau)^2}}.$$

Choosing N such that the last term above is less than one yields

$$\mathbb{P}\{\forall i \geq N :\ D_i \geq \tau \mid \forall n \in [k, N] :\ D_n \geq \tau\} > 0.$$

Since

$$\mathbb{P}\{\forall i \geq k :\ D_i \geq \tau\} = \mathbb{P}\{\forall i \in [k, N] :\ D_i \geq \tau\} \times \mathbb{P}\{\forall i \geq N :\ D_i \geq \tau \mid \forall n \in [k, N];\ D_n \geq \tau\},$$

we just proved that

$$\mathbb{P}\{\forall i \geq k :\ D_i \geq \tau\} > 0. \tag{4.4}$$

Define the stopping time $T' = \min\{n \geq k;\ D_n \leq \tau\}$. Since for all $t \geq \tau$, $g(t) + t \geq g(\tau) + \tau$, on the event $\{R_k \geq D_k \geq \tau\}$, there exists a coupling of the Bernoulli random variables B and B' such that

$$\forall n \in [k, T'] :\ B_n \geq B'_n,$$

and thus a coupling of the random variables R_n and D_n such that

$$\forall n \in [k, T'] :\ R_n \geq D_n.$$

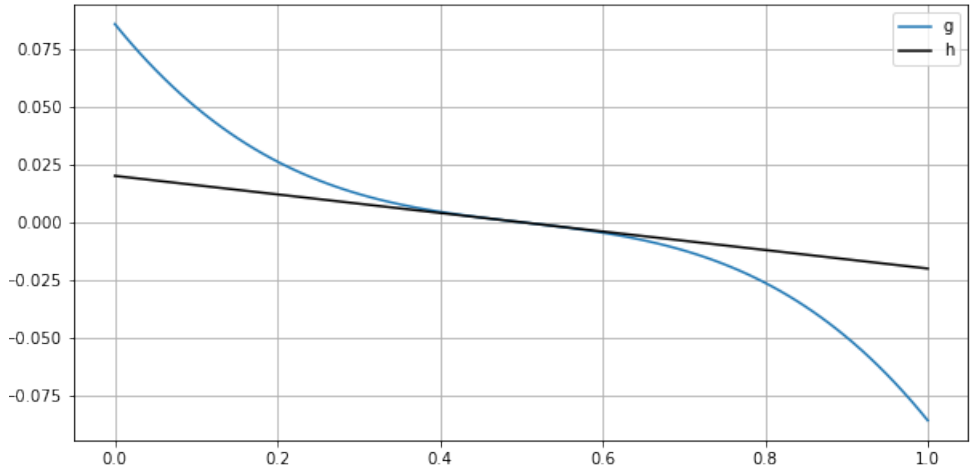

FIGURE 3. A linear lower bound for $|g|$, $k = 3$ and $p = 0.18$.

From this coupling and (4.4) we have

$$\mathbb{P}\left\{\forall n \geq k : \ R_n > \frac{1}{2}\right\} > 0,$$

which, thanks to Lemma 4.1, proves that in the regime $p < 1/2 - 1/(2\alpha_k)$,

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2},$$

proving the first statement of Theorem 1.1.

4.2. The high-rate-of-mutation regime $\left(\frac{1}{2} - \frac{1}{2\alpha_k} \leq p \leq \frac{1}{2}\right)$

In the range $p > 1/2 - 1/(2\alpha_k)$ the proportion of red balls converges to $1/2$. It does not mean that majority voting cannot be better than random guessing. Indeed, the proportion can converge to $1/2$ from above. This is this possibility that will now be investigated.

4.2.1. Extreme rate

First, we examine the "extreme" case when the rate of mutation is near $1/2$, more precisely when $p > 1/2 - 1/(4\alpha_k)$. Define the linear function h by $h(t) := g'(1/2)\,(t - 1/2)$. Then

$$g(t) \begin{cases} \geq h(t), \text{ if } t \in [0, 1/2], \\ \leq h(t), \text{ if } t \in [1/2, 1]. \end{cases}$$

In Figure 3 we plot h and g.
Let us define an auxiliary process R^*_n by the stochastic recursion $R^*_k = 1$ and for $n \geq k$

$$R^*_{n+1} = R^*_n + \frac{B_n\,(h(R^*_n) + R^*_n) - R^*_n}{n+1},$$

where $B_n\left(h(R_n^*)+R_n^*\right)$ is a Bernoulli random variable with parameter $h(R_n^*)+R_n^*$, conditionally independent of R_n^*. In particular,

$$\mathbb{E}\left[B_n\left(h(R_n^*)+R_n^*\right)-R_n^*|R_n^*=t\right]=h(t).$$

Since the value of g (for (R_n)) and h (for (R_n^*)) represents a drift in the processes R_n and R_n^* we expect that the process (R_n^*) is further away from $1/2$. Indeed, we may introduce a coupling as follows. Define the stopping time T^* as the first time R^* reaches $1/2$:

$$T^*:=\min\left\{n\geq k:\ R_n^*\leq\frac{1}{2}\right\}.$$

Since for the times $n\in[k,T^*]$, $h(R_n^*)\geq g(R_n)$, we may use a similar coupling argument as in Section 4.1. Thus, there is a coupling of R^* and R such that

$$\forall n\in[k,T^*];\ R_n\leq R_n^*.$$

From this coupling, for T defined in Lemma 4.1 we have

$$\mathbb{P}\left\{T=+\infty\right\}\leq\mathbb{P}\left\{T^*=+\infty\right\}. \tag{4.5}$$

Observe that in the case of $k=1$, g is linear and the two processes R_n and R_n^* coincide. The linear case was analyzed in Addario-Berry *et al.* [3] and we may use their results to understand the behavior of R_n^*. Indeed, the process defined in Addario-Berry *et al.* [3] is the same as R^* if one sets the flip probability of Addario-Berry *et al.* [3] equal to $-g'(1/2)/2$ and starts at time k. They prove that if $p\geq 1/4$, then, for the process starting at time 1, majority voting has an error probability of $1/2+o(1/2)$. Lemma 4.1 implies that this process reaches $1/2$ in finite time almost surely. So even conditioned on its value being 1 at time k it will reach $1/2$ in finite time almost surely. This proves that even for R_n^* starting at time k its error probability is $1/2+o(1)$. According to Lemma 4.1 this implies that for this range of p, $\mathbb{P}\left\{T^*=+\infty\right\}=0$. Hence, using Lemma 4.1 and (4.5), shows that if $g'(1/2)\leq-\frac{1}{2}$, then

$$\limsup_{n\to\infty}R^{maj}(n,p)=\frac{1}{2}.$$

Lemma 4.2 shows that $\lim_{n\to\infty}R^{maj}(n,p)=1/2$. Because $g'(1/2)=(1-2p)\alpha_k-1$, we just proved that if $p\geq 1/2-1/4\alpha_k$, then

$$\lim_{n\to\infty}R^{maj}(n,p)=\frac{1}{2},$$

completing the proof of the third statement of Theorem 1.1.

4.2.2. Intermediate rate

It remains to study the "intermediate" case $p\in[1/2-1/(2\alpha_k),1/2-1/(4\alpha_k))$. To this end, we may couple R_n to a process for which majority voting outperforms random guessing. Let us fix $p\in[1/2-1/(2\alpha_k),1/2-1/(4\alpha_k))$, which implies that $g'(1/2)/2>-1/4$. Then choose $q=-g'(1/2)/2+\epsilon$ with $\epsilon>0$ small enough so that $q<1/4$ and $g(0)>h(0)$. We define the linear function $h(t):=-2q(t-1/2)$, and as illustrated in Figure 4, we denote by a and b the intersection points between h and g (apart from $1/2$). More precisely a and b are defined as the the roots of $g-h$ distinct from 0. Since $g-h$ is strictly convex on $(0,1/2)$ and $(g-h)(0)>0$, $(g-h)'(1/2)<0$, a and b are well defined and sit respectively in $(0,1/2)$ and $(1/2,1)$.

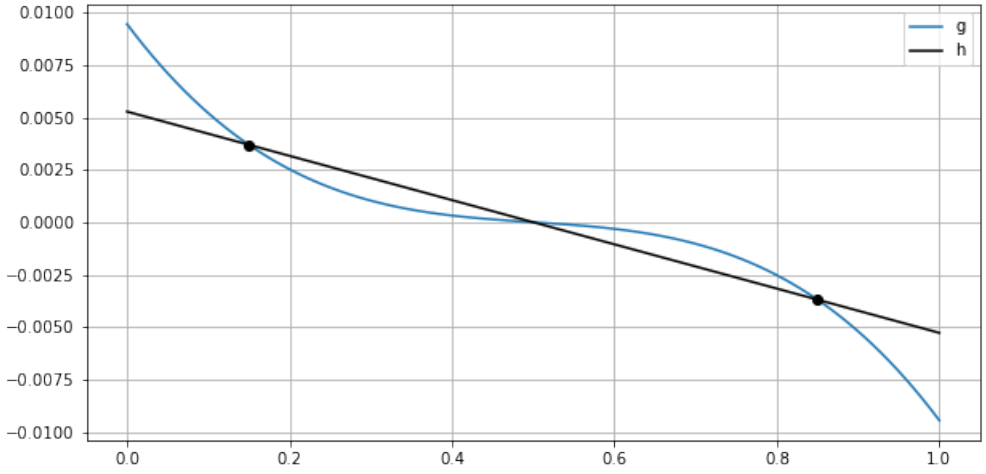

FIGURE 4. Comparison of h and g, for $k = 3$ and $p = 0.34$ (rescaled for clarity).

We define R_n^* similarly as in the previous section but now with $h(t) = -2q(t - 1/2)$, that is $R_k^* = 1$ and

$$R_{n+1}^* = R_n^* + \frac{B_n \left(h(R_n^*) + R_n^*\right) - R_n^*}{n+1},$$

where the B_n are conditionally independent Bernoulli random variables. In particular,

$$\mathbb{E}\left[B_n \left(h(R_n^*) + R_n^*\right) - R_n^* | R_n^* = t\right] = -2q\left(t - \frac{1}{2}\right).$$

Just as in the previous section, we may use the analysis of Addario-Berry *et al.* [3] for the case $k = 1$ with mutation probability of q. Addario-Berry *et al.* [3] state that for the process R_n^* started at time 1 and for $q < 1/4$ majority voting is better than random guessing. As it was pointed out to us by a referee, the proof in Addario-Berry *et al.* [3] that uses general limit theorems for Pólya urns with randomized replacements due to Janson [20] is incorrect. In the Appendix we give a self-contained proof of this statement.

To use the result for trees, we need to make sure that it holds when R_n^* is defined as above, started at time k. Let $R_n'^*$ be the process started at time 1, for which the majority voting is known to outperform random guessing. Let T'^* and T^* be the first time indices at which $R_n'^*$ and R_n^* reach 1/2, respectively. Finally, let $(U_n)_{n\in\mathbb{N}}$ be a collection of independent uniform random variables. For $n \in [k, T'^*)$, we couple $R_n'^*$ and R_n^* as follows

$$R_{n+1}'^* = R_n'^* + \frac{\mathbb{1}\left(U_n \le h(R_n'^*) + R_n'^*\right) - R_n'^*}{n+1},$$

and

$$R_{n+1}^* = R_n^* + \frac{\mathbb{1}\left(U_n \le h(R_n^*) + R_n^*\right) - R_n^*}{n+1}.$$

With this coupling, a recursion proves that for all $n \in [k, T'^*)$, $R_n^* \ge R_n'^*$. Because majority voting is known to outperform random guessing for R'^*, Lemma 4.1 proves that $\mathbb{P}\{T'^* = +\infty\} > 0$. The coupling directly implies that $\mathbb{P}\{T^* = +\infty\} > 0$. So majority voting outperforms random guessing for the process R^*. Thus,

from Lemma 4.1 it follows that

$$\mathbb{P}\left\{\forall n \geq k : \ R_n^* > \frac{1}{2}\right\} > 0.$$

Now, from Lemma 3.1 we deduce that both processes R_n and R_n^* converge almost surely to $1/2$ and exceed b only finitely many times. Thus, there exists an almost surely finite random time T' such that and $\forall n \geq T'$; $R_n \leq b$ and $R_n^* \leq b$. We use similar coupling arguments as in Section 4.1. So, on the event that R^* does not reach $1/2$ we can couple R_n and R_n^* from T' onwards such that $R_n \geq R_n^*$. This proves that

$$\mathbb{P}\left\{\forall n \geq T' : \ R_n > \frac{1}{2} \mid T'\right\} > 0.$$

Using that T' is finite almost surely and Lemma 4.1 we conclude that majority voting is better than random guessing in this regime. More precisely, if $1/2 - 1/2\alpha_k \leq p < 1/2 - 1/4\alpha_k$, then

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2}.$$

This completes the proof of Theorem 1.1.

5. A general lower bound

In this final section, we derive a lower bound for the probability of error that holds for all mutation probabilities. In particular we show the following.

Proposition 5.1. *Let k be a positive odd integer and let $k/2 < \ell < k$. Assume that initially there are ℓ red vertices, that is $R_k = \ell/k$. Letting*

$$h_k := \mathbb{P}\left\{Beta\left(\frac{k+1}{2}, \frac{k+1}{2}\right) \geq 1 - \frac{1}{k}\right\},$$

the probability of error of the majority rule satisfies

$$\inf_{\substack{0\leq p\leq 1\\ n\geq 2\ell}} \mathbb{P}\left\{b_n^{maj} \neq b_k^{maj}\right\} \geq \frac{1}{2} h_k^{2\ell-k}.$$

Proof. The proposition follows by simply considering the event that the first $2\ell - k$ new vertices are all blue. In that case, at time 2ℓ the number of red and blue vertices are equal. We may write, for any $n \geq 2\ell$,

$$\mathbb{P}\left\{b_n^{maj} \neq b_k^{maj}\right\} \geq \mathbb{P}\left\{b_n^{maj} \neq b_k^{maj} \mid c_{k+1} = \cdots = c_{2\ell} = B\right\} \times \mathbb{P}\left\{c_{k+1} = \cdots = c_{2\ell} = B\right\}.$$

From the symmetry of our model, $\mathbb{P}\left\{b_n^{maj} \neq b_k^{maj} \mid c_{k+1} = \cdots = c_{2\ell} = B\right\} = 1/2$. Thus

$$\mathbb{P}\left\{b_n^{maj} \neq b_k^{maj}\right\} \geq \frac{\mathbb{P}\left\{c_{k+1} = \cdots c_{2\ell} = B\right\}}{2}.$$

To estimate the probability on the right-hand side, we use (2.4), which implies

$$\mathbb{P}\{c_i = B\} = \int_{f(R_i)}^{1} (x(1-x))^{\frac{k-1}{2}} \frac{\Gamma(k+1)}{\Gamma^2\left(\frac{k+1}{2}\right)} \mathrm{d}x.$$

If $R_k = \ell/k$ and $c_{k+1} = \cdots = c_{i-1} = B$, where $k < i \le 2k$, then $R_{i-1} = \ell/i$. Since $0 \le p \le 1/2$,

$$f(R_{i-1}) = (1-2p)\frac{\ell}{i} + p \le \max\left(\frac{1}{2}, \frac{\ell}{i}\right) \le \frac{k-1}{k} = 1 - \frac{1}{k}.$$

Therefore,

$$\min_{k<i\le 2\ell} \mathbb{P}\{c_i = B \mid c_{k+1} = \cdots = c_{i-1} = B\} \ge h_k,$$

as claimed. □

6. Concluding remarks

In this paper we study the majority rule for guessing the initial bit values at the roots of a random recursive k-DAG in a broadcasting model. The main result of the paper characterizes the values of the mutation probability for which the majority rule performs strictly better than random guessing. Even in this exact model, many interesting questions remain open. For example, we do not have sharp bounds for the probability of error. It would also be interesting to study other, more sophisticated, classification rules that take the structure of the observed k-DAG into account. In particular, the optimal probability of error (as a function of k and the mutation probability p) is far from being well understood. For an initial study of localizing the root vertices, we refer the interested reader to Briend *et al.* [23].

A natural extension of the model is obtained by considering $q > 2$ colors. In this model, one aims at guessing the most common color of the initial configuration. In order to extend our results, instead of a single number, one needs to consider a vector of dimension $q-1$ to track the proportion of each color. For example, one may consider the following rule to assign a color to a new vertex. At each step, among the k observed colors, pick the most common (break ties uniformly at random). However, the analysis becomes most complex since instead of comparing one random variable to $1/2$, one needs to compare a random variable to $q-2$ others to determine which one is the most common. If one manages to write this recursion in a tractable manner, we believe that a similar approach as the one of this paper may be used to understand the evolution of the proportion of each color. Depending on the convergence regime, an important part of our proof relies on the comparison to the tree case, that is, the case $k = 1$. In a tree, one way to study the multi-color problem is to group $q-1$ of the colors together. By doing so, the multi-color problem is simplified to a two-color problem in the tree with a non-symmetric flip probability. However, the details may be nontrivial and are left to future research.

Acknowledgments

We thank the referees for their interesting comments. We are especially grateful to a referee for pointing out a mistake in [3] which is fixed in the Appendix below.

Simon Briend acknowledges the support of Région Ile de France. Gábor Lugosi acknowledges the support of Ayudas Fundación BBVA a Proyectos de Investigación Científica 2021 and the Spanish Ministry of Economy and Competitiveness grant PID2022-138268NB-I00, financed by MCIN/AEI/10.13039/501100011033, FSE+MTM2015-67304-P, and FEDER, EU.

Data availability statement

The research data associated with this article are included in the article.

References

[1] T. Coolen, A. Annibale and E. Roberts, Generating Random Networks and Graphs. Oxford University Press (2017). ISBN 9780198709893.

[2] W. Evans, C. Kenyon, Y. Peres and L.J. Schulman, Broadcasting on trees and the Ising model. *Ann. Appl. Probab.* **10** (2000) 410–433.

[3] L. Addario-Berry, L. Devroye, G. Lugosi and V. Velona, Broadcasting on random recursive trees. *Ann. Appl. Probab.* **32** (2022) 497–528.

[4] C. Desmarais, C. Holmgren and S. Wagner, Broadcasting-induced colorings of preferential attachment trees. Random Structures & Algorithms, **63**, (2023) 364–405.

[5] A. Sly, Reconstruction for the Potts model. *Ann. Probab.* **39** (2011) 1365–1406.

[6] E. Mossel, Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.* **11** (2001) 285–300.

[7] S. Janson and E. Mossel, Robust reconstruction on trees is determined by the second eigenvalue. *Ann. Probab.* **32** (2004) 2630–2649.

[8] E. Mossel, Survey: information flow on trees, in Graphs, morphisms and statistical physics. Vol. 63 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* American Mathematical Society, Providence, RI (2004) 155–170.

[9] H.A. Harutyunyan and Z. Li, A new construction of broadcast graphs. *Discrete Appl. Math.* **280** (2020) 144–155.

[10] P. Bhabak, H.A. Harutyunyan and S. Tanna, Broadcasting in Harary-like graphs, in *2014 IEEE 17th International Conference on Computational Science and Engineering* (2014) 1269–1276.

[11] A. Makur, E. Mossel and Y. Polyanskiy, Broadcasting on random directed acyclic graphs. *IEEE Trans. Inform. Theory* **66** (2020) 780–812.

[12] T. Antunović, E. Mossel and M.Z. Rácz, Coexistence in preferential attachment networks. *Combinatorics Probab. Comput.* **25** (2016) 797–822.

[13] R. Pemantle, A survey of random processes with reinforcement. *Probab. Surv.* **4** (2007) 9–12.

[14] R. Pemantle, Nonconvergence to unstable points in urn models and stochastic approximations. *Ann. Probab.* **18** (1990) 698–712.

[15] S. Janson, Functional limit theorems for multitype branching processes and generalized pólya urns. *Stoch. Processes Applic.* **110** (2004) 177–245.

[16] M. Knape and R. Neininger, Pólya urns *via* the contraction method. *Combinatorics Probab. Comput.* **23** (2014) 1148–1186.

[17] L.J. Wei, The generalized Pólya's urn design for sequential medical trials. *Ann. Statist.* **7** (1979) 291–296.

[18] J. Bertoin, Limits of Pólya urns with innovations (2022). URL https://arxiv.org/abs/2204.03470.

[19] M. Kuba and H.M. Mahmoud, Two-color balanced affine urn models with multiple drawings. *Adv. Appl. Math.* **90** (2017) 1–26.

[20] S. Janson, Random replacements in Pólya urns with infinitely many colours. *Electron. Commun. Probab.* **24** (2019) 1–11.

[21] L.-X. Zhang, Convergence of randomized urn models with irreducible and reducible replacement policy. arXiv preprint arXiv:2204.04810 (2022).

[22] I. Crimaldi, P.-Y. Louis and I.G. Minelli, An urn model with random multiple drawing and random addition. *Stoch. Processes Applic.* **147** (2022) 270–299.

[23] S. Briend, F. Calvillo and G. Lugosi, Archaeology of random recursive dags and cooper-frieze random networks. *Combinatorics, Probab. Comput.* **32** (2023) 859–873.

Appendix A.

A.1 In trees majority is better than random guessing for $p < 1/4$

In Section 2.3 of [3] it is stated that the majority vote is asymptotically better than random guessing when $k = 1$ (*i.e.*, the graph is a uniform random recursive tree) and the mutation probability p is less than $1/4$. This result is used in Section 4.2.2 above.

In [3] it is claimed that this follows simply from a general limit theorem for Pólya urns with randomized replacements due to Janson [20]. However, the claimed symmetry in Janson's limit distribution was not checked in [3]. In order to remedy this, in the next Proposition we give a self-contained proof of the statement.

Suppose, without loss of generality, that the root vertex is red, that is, $R_1 = 1$. Define the difference between the number of red and blue balls at time n by $\Delta_n = n(R_n - B_n)$.

Proposition A.1. *If $p < 1/4$, then*

$$\liminf_{n\to\infty} \mathbb{P}\{\Delta_n > 0\} > \frac{1}{2}.$$

Proof. We use the representation defined in Section 2.1 of [3] for the difference Δ_n, which we recall now.

The URRT is generated in the standard way, without attached colors, with 0 being the root vertex, and for $i \in \{1, \dots, n\}$, $p_i \in \{0, \dots, i-1\}$ is the uniform random index of the parent of vertex i. Coloring the vertices may be equivalently done as follows:

- let $M_1, M_2, \dots, M_n \in \{0, 1\}$ be independent Bernoulli$(2p)$ random variables. When $M_i = 1$, vertex i is *marked*. Then there is an independent coin flip ξ_i that takes values uniformly at random in $\{-1, 1\}$ and determines if a marked node takes the same color as its parent or it flips.
- when $M_i = 0$, vertex i is *not marked*. These nodes have the same color as their parent.

The root and marked nodes become roots of subtrees that are disjoint and partition the uniform recursive tree into many pieces. Each of the subtrees consists of nodes of the same color, and the roots have the color of their original parent if $\xi = 1$ and different otherwise (if $\xi_i = -1$). More precisely, if $B_i \in \{-1, 1\}$ is the color of vertex i (with $+1$ interpreted as "red" and -1 as "blue"), then

$$B_i = \begin{cases} B_{p_i} & \text{if } M_i = 0 \text{ (no marking) or if } M_i = 1,\ \xi_i = +1 \text{ (no flipping);} \\ -B_{p_i} & \text{if } M_i = 1,\ \xi_i = -1. \end{cases}$$

Let N_i denote the size of the maximal subtree rooted at vertex i such that all its vertices apart from i are unmarked (and therefore monochromatic).

With this notation, Δ_n may be written as $\Delta_n = N_0 + W_n$, where

$$W_n = \sum_{i=1}^{n} N_i B_{p_i} \xi_i M_i.$$

Since the Rademacher random variables ξ_i are independent of all other random variables, W_n has a symmetric distribution about 0. In particular, by conditioning on all other random variables, we get that

$$\mathbb{P}\{\Delta_n \le 0\} = \frac{1}{2}\mathbb{P}\{N_0 \le |W_n|\}.$$

Hence, it suffices to show that $\limsup_{n\to\infty} \mathbb{P}\{N_0 \le |W_n|\} < 1$. To this end, for a positive integer k, let $\mathcal{E}_k$ be the event that the first k vertices are unmarked, that is, $M_1 = \cdots = M_k = 0$. Clearly, $\mathbb{P}\{\mathcal{E}_k\} = (1-2p)^k$. Then

$$\begin{aligned} \mathbb{P}\{|W_n| \ge N_0\} &\le \mathbb{P}\{\mathcal{E}_k^c\} + \mathbb{P}\{|W_n| \ge N_0, \mathcal{E}_k\} \\ &= 1 - \mathbb{P}\{\mathcal{E}_k\} + \mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| > N_0, \mathcal{E}_k\right\} \\ &= 1 - \mathbb{P}\{\mathcal{E}_k\}\left(1 - \mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| > N_0\right\}\right), \end{aligned}$$

where we used the fact that the events $\mathcal{E}_k$ and $\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| > N_0$ are independent. Thus, it suffices to prove that there exists an integer $k > 0$ such that $\limsup_{n\to\infty} \mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| > N_0\right\} < 1$. To this end, we may write

$$\mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| > N_0\right\} \le \mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| > \frac{\mathbb{E}N_0}{2}\right\} + \mathbb{P}\left\{N_0 < \frac{\mathbb{E}N_0}{2}\right\}. \tag{A.1}$$

We show that the second term on the right-hand side is bounded away from one, while the first term can be made arbitrarily small by choosing k sufficiently large. In order to bound the second term on the right-hand side of (A.1), we use the fact that by [3], Lemmas 4 and 6, $\mathbb{E}N_0 \ge \mathrm{e}^{-1}(n+1)^{1-2p}$ and $\mathrm{Var}(N_0) \le c(p)(n+1)^{2-4p} + O(n\log n)$ for a constant $c(p) > 0$ depending on p only. Hence, using the Chebyshev-Cantelli inequality,

$$\mathbb{P}\left\{N_0 \ge \frac{\mathbb{E}N_0}{2}\right\} \ge \frac{(\mathbb{E}N_0)^2}{(\mathbb{E}N_0)^2 + 4\mathrm{Var}(N_0)} \ge \frac{\mathrm{e}^{-2}}{\mathrm{e}^{-2} + c(p) + o_n(1)},$$

which is clearly bounded away from 0.

Using once again the bound $\mathbb{E}N_0 \ge \mathrm{e}^{-1}(n+1)^{1-2p}$, the first probability on the right-hand side on (A.1) may be upper bounded, using Markov's inequality, by

$$\mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| \ge \frac{(n+1)^{1-2p}}{2e}\right\} \le \frac{\left(\mathbb{E}\left(\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right)^2\right)^{1/2}}{(n+1)^{1-2p}/2e}$$

Since the ξ_i are independent of all other random variables,

$$\mathbb{E}\left(\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right)^2 = \sum_{i=k+1}^{n} \mathbb{E}N_i^2 \le \sum_{i=k+1}^{n} \left(\left(\frac{n+1}{i+1}\right)^{2-4p} \mathrm{e}^4(4+e) + e\right),$$

where the upper bound for $\mathbb{E}N_i^2$ follows from [3], Lemma 6. Thus,

$$\begin{aligned}\mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| \ge \frac{(n+1)^{1-2p}}{2e}\right\} &\le \left(2\mathrm{e}^3\sqrt{4+e}\right)\left(\sum_{i=k+1}^{n} \frac{1}{(i+1)^2}\right)^{1/2} + 2\mathrm{e}^{3/2} n^{-1/2+2p} \\ &\le \frac{2\mathrm{e}^3\sqrt{4+e}}{\sqrt{k}} + o_n(1).\end{aligned}$$

Thus, by choosing k sufficiently large, we clearly have $\limsup_{n\to\infty} \mathbb{P}\left\{\left|\sum_{i=k+1}^{n} N_i B_{p_i} \xi_i M_i\right| > N_0\right\} < 1$ as desired. □