POLITECNICO DI TORINO

Master Degree in Mathematical Engineering

Master Thesis

A mathematical model for the emergence of innovations

Advisor:
Prof. Giacomo Como

Candidate:
Alessandro Mastrototaro

October 2018

Summary

In this thesis, a recently proposed urn-based model with triggering, describing how novelties and innovations emerge in real systems, is studied. Through the urn process a sequence of elements is generated. Relevant statistics can be studied such as how many distinct elements appear in the sequence until a certain time or the frequency with which each of the elements has been observed. These statistics turn out to follow the Heaps' and Zipf's law, respectively. Some heuristic arguments have been proposed in the literature to justify the emergence of these laws.

One of the main contributions of this thesis consists in providing rigorous proofs for these results. This is achieved using stochastic approximation techniques, that allow one to approximate certain classes of stochastic processes through ordinary differential equations.

A second contribution consists in the extension of these results, including the analysis of the number of elements that have appeared at least k times in the sequence, for arbitrary k. Such extensions allow one to consider design problems where the two main parameters of the model are chosen in such a way to optimize meaningful objectives.

These analytical results are then supported and validated by numerical simulations, which also allow one to observe other possibly interesting emerging behaviors and patterns. Finally, some modifications of the model are proposed in order to address specific issues that could better match certain phenomena observed in real systems.

Contents

Chapter 1

Introduction

In the world every day, every hour and even every second novelties emerge. Some of these novelties are only new in the perspective of the observer, while others are real innovations, which appear in the world for the first time. The processes leading to these discoveries are complex and heterogeneous among the different aspects of life where novelties and innovations are observed. However there are some general patterns that have been observed in nature and two of the most important ones are described by Heaps' and Zipf's laws. The first law states that the number of distinct elements in a time sequence grows at sub-linear rate. It was intended at first to describe the number of distinct words appearing in a text, where "time", in the specific case, represented the order at which each word appears in the text. The second law states that the tail of the rank-size distribution, which describes how many times an element is observed as a function of its rank, is a power law of order α, i.e., the i-th most observed element, when i is sufficiently large, has frequency $f(i) \asymp i^{-\alpha}$, for some $\alpha \in \mathbb{R}^+$.

In the literature, some models have been proposed to justify the emergence of these empirically observed laws from first modeling principles or mechanisms, see, e.g., MIT Technology Review [1]. However, most of them failed to replicate the concept of *adjacent possible*, theorized for the first time by Kauffman ([2],[3]). By *adjacent possible* it is meant the idea that new innovations open the way to the discovery other ones. The crucial point is the difficulty of creating a model for the unknown, or better, to describe something that has not appeared yet in the real world.

An important class of models used for the emergence of these laws are *urn-based models*. These are able to reproduce *richer-gets-richer* mechanisms, which means that most frequent items are more likely to be observed again. Some models have been proposed in literature: the latest ones are based on urns that contain colored balls, each color representing a different item, and then the focus is to study the sequence of colors drawn.

The most common and generic model is Polya's urn: the basic rule is that inside there are some initial balls of different colors and, when one of them is drawn, it is replaced with some copies of the same color. In the model introduced by Hoppe there is also a heavier ball that is drawn with probability proportional to its weight and if it is drawn it is replaced together with a brand new color. This procedure allows to explore the space but it does not really take into account the concept of *adjacent possible*. This can be achieved through the Polya urn with triggering model, in which every time a ball is drawn for the first time a defined number of balls of brand new colors are inserted into the urn, i.e. the draw of new colors triggers the possible draw of further new colors.

The thesis starts from the work of Loreto et al. (2016) [4], which is a review on the more or less recent models proposed in literature for modeling the emergence of innovations. In particular the model based on a specific Polya urn is highlighted and modified in order to give the concept of *adjacent possible* a practical representation.

Some of the main works that focused on studying models reproducing Heaps' and Zipf's laws are described by Loreto et al. [4]; the common goal is to find a way to generate a sequence of elements $S(t)$, $t \in \mathbb{N}$, represented by positive integer numbers, with the aim of calculating $D(t)$ and $f(i)$ as described before, with t representing the index of the element in the sequence. They include:

- Simon-like models, in which at each time a new element appears with probability p, or one among the already appeared ones is drawn again with probability proportional to its frequency;
- the sample-space reducing model, in which the number of distinct elements has an upper bound, we call it N, and each element is sampled with uniform probability among the integer numbers between 1 and the preceding number in the sequence;
- the Hoppe urn;
- the Polya urn with triggering, main focus of this thesis.

The article from Tria et al. [5] introduced for the first time the last model in the list, with further modifications in which they take into account the correlations between elements that are semantically similar; Monechi et al. [6] also model how elements that appear later in time can be successful as the already present in the sequence. Another completely different work, based on edge-reinforced random walks, was proposed by Iacopini et al. [7].

The objectives of the thesis are the following.

- To review the state of the art with particular focus on the before stated Polya urn with triggering.

- To formalize some heuristic arguments introduced in the article: the stochastic approximation will give a tool to study the model with formal arguments and extend the results for a more general analysis. The main points of interest are two: how the number of distinct drawn elements increases in time and, given the rank based on the number of draws for each distinct element (color), how that number depends on the rank and on the time of first appearance.
- To extend the model with some modifications to the original one, with the specific aim of changing some characteristics in order to give a better model of the real systems.

In Chapter 2 the model is presented in details, together with state of the arts and some applications that motivated it. In Chapter 3 the stochastic approximation, whose fundamentals are presented in Appendix A, is applied to the model; as a result of that the ODE that should formally predict the behavior of some statistics of the model is studied. In Chapter 4 the previous analysis of the ODE is used in order to retrieve expressions for Heaps' and Zipf's laws. It is also shown that it is possible to find the fraction of balls of colors with at least k draws, $k \geq 1$, and give an optimum value to maximize it. In Chapter 5 the analytical results are verified with numerical simulations of the model and some data and plots are shown, in order to give a better understanding on how the model behaves even in a finite, but large enough, time. In Chapter 6 some changes are applied to the model, with specific motivations, and it is observed how some patterns change or remain the same: in one case a different number of balls is reintroduced depending on the color, while in the other one the balls in urn can disappear, since each of them has geometric distributed life. In Chapter 7 the conclusions of the work are presented together with some suggestions for future work.

Chapter 2

Mathematical models for the emergence of innovations

In this chapter, the motivation and state of the art regarding the modeling of emergence of innovations is presented, together with the important definitions of Heaps' and Zipf's laws. After that, the main focus will be the description of Polya's urn with triggering and the heuristic results in literature.

2.1 Motivation

In the world, it continuously happens to observe novelties in different aspects of life: new songs, new scientific publications, new start-ups and many others. Some of them are just new for the one who observes them, some others are real innovations never seen before by anyone. Generally processes of innovations are observed in biological systems, human society, and technology but it is quite difficult to understand the mechanisms through which some novelties get more success than others, since they are quite complex and heterogeneous among different system. The randomness plays a big role in the innovation, even though there are some global statistics of many considered system that present general patterns. The statistics of interest are mainly two:

1. How many different innovations (novelties) from the beginning of the process have been observed until a specific time? At what rate do they appear?

2. Given a measure of popularity of innovations (novelties), and a ranking based on this measure, what is the dependence of the popularity of an innovation on its rank?

Regarding the first question it is necessary to clarify what is considered as *time*, while on the second one we need to define a measure of popularity: the answer is that they

depend on the system considered. For example if we consider a user perspective in a music streaming service the time can be represented by a listening of any song, ordered in time, while the measure of popularity can be the number of plays of each distinct song; considering instead a global perspective the time can be identified with the time of release of each song and create a time sequence, with the popularity based on total plays of a song. A similar consideration can be done for scientific articles, using their time of publication to create a time sequence and using the number of citations received as a measure of popularity. Regarding start-ups instead we can consider a regional ecosystem since about the date of its foundation e define time-steps: at each of them we can count the number of different start-ups that are present and give a measure of popularity based on the total funding received in its history. The examples could be many more, and in most of them the two statistics previously presented generally assume the form power laws: specifically the answer to the former question is the Heaps law, while to latter is the Zipf law.

Definition 1. *The Heaps law, introduced by Heaps in 1978 [8], is an empirical law that describes the number of different elements d_n occurring in a sequence of length n when n is large, with the following rule:*

$$d_n \asymp n^{\beta}, \qquad \beta \in (0,1).$$

Definition 2. *The Zipf law is an empirical law, described by Zipf in 1949 [9], which describes the frequency-rank distribution $f(j)$ of some quantity: the element at position j in the rank has frequency that follows a power law:*

$$f(j) \propto j^{-\alpha}, \qquad \alpha > 0$$

and clearly this is decreasing in j. If $\alpha = 1$ we call it "exact" Zipf's law, while if $\alpha \neq 1$ it is called "generalized".

There is also a relation between the two laws: in general, if we estimate α in the tail of the distribution, i.e. for low ranks, it should hold $\alpha = \frac{1}{\beta}$, where β is the estimated coefficient for Heaps' law.
Heaps' law was first introduced to describe the rate at which distinct words appear in a text: in this case the novelties are from the text perspective, the time is the order of appearance of a word and the popularity is clearly based on the number of times it appears in the text. The case of texts is a bit different, since we have a predefined vocabulary and, even though there may be neologisms, the number of distinct elements has an upper bound. Finding a model that could reproduce Heaps' and Zipf's laws has been the objective of many studies, however the problem of modelling some future event whose probability is not zero but it has not been observed yet was an issue, as Zabell stated [10]:

> *This is not the problem of observing the 'impossible', that is, an event whose possibility we have considered but whose probability we judge to be 0. Rather,*

the problem arises when we observe an event whose existence we did not even previously suspect; this is the so-called problem of 'unanticipated knowledge'.

2.1.1 Review of models for the emergence of innovations

Reviewing the models proposed in literature, already presented in the introduction of this thesis, some of them tried to model the *unanticipated knowledge* introducing a probability of observing a never-seen-before element. The purpose of each model is to create a sequence in which the position of the element represents the time and then observe the rate at which the number of distinct elements in the sequence increase in time and, at the end, studying the frequency-rank distribution of the elements. The common pattern of all the models is the one called *richer-gets-richer*, which determines a major popularity for elements already among the most popular, through different reinforcement mechanisms. The following models have different ways to reinforce the most common elements in the sequence and to insert new ones.

- The Simon-like models [11] were among the first ones proposed to study frequency-rank distribution in texts: the sequence starts with a single element and at each time a new element is recorded with probability p, otherwise an element among the ones already drawn is randomly chosen, proportional to its frequency, such that the *richer-gets-richer* mechanism is reproduced. In the original version p was a constant probability, while following upgrades made it decreasing in time or introduced factors that increased the probability of choosing most recent elements in the sequence than older ones.

- The Hoppe urn model [12], based on generic Polya urn ([13],[14],[15]) and already used in genetics, is based on a urn that, at the beginning, has some colored balls of weight 1 and a black ball of weight θ. At each time n a ball is drawn from the urn proportional to its weight: if it is a regular ball it is replaced with a copy of the same color (*richer-gets-richer* since each color has a probability of being drawn based on its frequency in the urn); if it is the black one it is replaced in the urn together with a ball of brand new color. Instead of recording a sequence we are here interested in determining how the number of distinct colors in urn increases with time and the frequency-rank distribution in it.

- The Polya urn with triggering, fully described in the next section.

- The sample-space reducing model [16] is quite different from other models: it has instead an upper bound N to the number of maximum distinct elements and does not directly reproduce the *richer-gets-richer* mechanism, however it has been used to represent power law frequency distributions. It starts sampling an integer number from 1 to N and, after that, the following recursion is used: the n-th element in

the sequence is uniformly sampled among the integers between 1 and the preceding number in the sequence. Once number 1 appears, the process starts again from N. There is also a case in which we can consider at each step a probability λ of sampling again from 1 to N instead of looking at the previous number.

Another concept developed in literature is the already cited *adjacent possible*: it makes a step beyond the dichotomy between the actual and the possible [17], defining an abstract space where we find innovations that are just one step from appearing in the reality. The ideal model is to represent the world of innovations as a graph in which each node represents a single element: imagining a random walker on a graph, the nodes are divided in two different categories, visited and unvisited ones (figure 2.1). Once one of the second

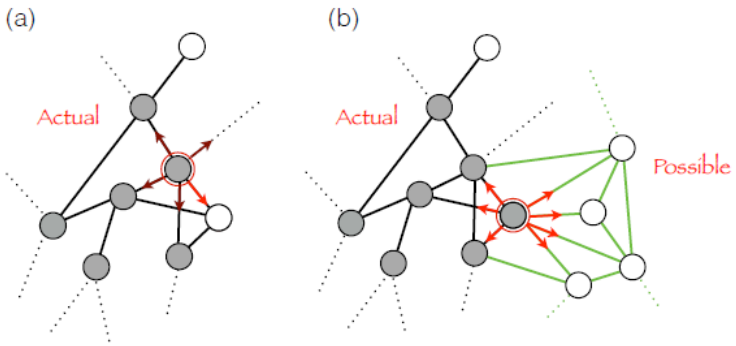


Figure 2.1: Visual representation of *adjacent possible*: grey nodes are the already visited ones, while the white ones are unvisited. (Source: Loreto et al. [4])

category is visited, new nodes connected to it (and maybe to other nodes) become visible and the graph expands. As Steven Johnson [18] stated:

> *The strange and beautiful truth about the adjacent possible is that its boundaries grow as one explores them.*

Network dynamics of innovation processes

The idea of representing the emergence of novelties as graph was recently used by Iacopini et al. [7] but their limit is still the fact that the number of nodes, which gives an upper bound to the the number of maximum elements observable, and links is predefined and there is no expansion of the graph during the process. They proposed a model based on a edge-reinforced random walk on a small-world network: the graph is represented as a ring in which each node is connected to all its closest m left neighbors and m right neighbours, and any other couple of nodes is connected with a small probability p. Each link has the same initial weight and the process starts on a generic node: the next node is chosen among the neighbors of the actual one, with probability proportional to the weight of the

link; once the link is crossed it receives a reinforcement on its weight, which is incremented by a tunable quantity δw. At first they empirically showed that the number of distinct visited nodes followed Heaps' law, in which the estimated exponent β was a decreasing function of δw. After that, they used the model in order to reproduce the growth of science knowledge analyzing the text of scientific articles published in years from 1991 to 2010, focusing on four distinct disciplines: astronomy, ecology, economy and mathematics. They observed for each of them the growth of new concepts through time and then, instead of a small-world network, they extracted the one underlined by the specific field and tuned the parameter δw such that could reproduce the same exponent of observed Heaps' law. At last they showed that this model intrinsically takes into account the correlations between similar concepts, using the distribution of inter-event times and the normalized entropy: the first one shows that, considering the time between two consecutive appearances of the same concept the distribution is concentrated in lower times with respect to the reshuffled sequence; the second one shows that the average entropy, a measure that counts how much equally distributed in time is each element a after the first appearance in the sequence, is lower than in the reshuffled sequence. It is interesting to notice that this model is not Markovian: indeed, due the edge-reinforcement mechanism, the number of visits of each nodes at a certain time is dependent on the whole history of the process and not only to the number of visits at the preceding time.
Therefore the correlations in this model are intrinsically determined by the network structure of the model, while in the extension of Polya's urn with triggering proposed by Tria et al. [5] the correlations are created with an artificial system based on labels and weight to the balls. In the thesis we will keep the simple definition of this model, as explained in the next section.

2.2 The Polya urn with triggering (PUT) model

The case of our interest is Polya's urn with triggering model, from now on indicates as PUT, introduced by Tria et al. [5] and whose process works with the following steps and rules:

1. the urn is represented as a set U, in which at time $n = 0$ there are N_0 initial balls of different colors;

2. at each time $n \in \mathbb{N}$ $(n > 0)$ a ball is drawn from U and replaced in it with ρ copies of the same color;

3. S is the ordered sequence that records all the draws, i.e. $S(n)$ represents the color drawn at time n;

4. if the drawn ball has never appeared in S, i.e. $S(n) \neq S(k)\ \forall k \in \{1, \dots, n-1\}$, then $\nu + 1$ balls of brand new colors are added to U;

5. in another version of the model the ρ copies are placed in the urn only starting from the second draw of each color.

A graphical representation of the process is pictured in figure 2.2. The process is intended

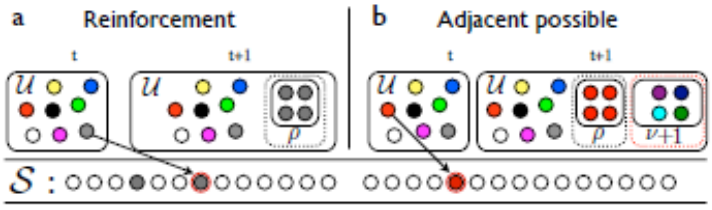


Figure 2.2: PUT model: visual representation of the reinforcement and triggering mechanisms. (Source: Loreto et al. [4])

to reproduce the *richer-gets-richer* mechanism for which the colors with more draws are more likely to be drawn, since the probability is proportional to the presence in the urn. While the parameter ρ empowers the reinforcement of the drawn colors, the parameter ν is intended to explore the *adjacent possible*, since every new color that appears in S triggers new colors and gives them the possibility to be drawn, expanding the space to explore. This last feature is quite different from all the other models presented: in Simon's models or Hoppe's urn the expansion of the explored space is obtained with some sort of artificial tools, while here it tries to reproduce what happens in reality, where each discovery opens the path to new ones.
Our interest now is to study two statistics of the process just described:

- $\left(D_n^{(k)}\right)_{n\geq 0}$, which counts the number of distinct colors appeared in S up to time n, drawn at least k times: notice that $\left(D_n^{(1)}\right)_{n\geq 0}$ represents the number of different colors drawn and the main focus will be on this one, which will be simply written as $(D_n)_{n\geq 0}$;
- $(K_n^j)_{n\geq 0}$, which counts the number of times the j-th color ever appeared in S up to time n has been drawn.

The first one clearly will be used to prove that the model reproduces Heaps' law, while the second will need more discussion in order to prove Zipf's law.

2.3 Some heuristic results

While Iacopini et al. [7] only numerically showed their result for the process $(D_n)_{n\geq 0}$, Loreto et al. [4], together with the definitions, also gave heuristic arguments to prove Heaps' and Zipf's law for the models presented. Defined D_n, the number of distinct elements

counted at time n in the specific model, and $f(r)$, the frequency of the element in position r of the rank, we have, for $n \to \infty$:

- for simple Simon's model
$$D_n \sim pn;$$
$$f(r) \propto r^{1-p};$$
- for Hoppe's urn model
$$D_n \sim \theta \log\left(1 - \frac{n}{\theta}\right);$$
$$f(r) \propto e^{-\frac{r-1}{\theta}};$$
- for classic sample-space reducing model
$$D_n \sim N - \sum_{i=1}^{N} e^{-\frac{nA}{i}};$$
$$f(r) \propto r^{-1};$$
where A is a normalization constant satisfying
$$A^{-1} = \sum_{i=1}^{n} i^{-1}.$$

Regarding PUT model, the following results were presented in the article, which we will formalize and slightly correct in next chapters. Setting the parameter $a = \nu + 1$ in the original version of the model and $a = \nu + 1 - \rho$ in the case where there is no reinforcement at the first draw of a color we have, for $n \to \infty$:

- $\rho > \nu$: $D_n \sim (\rho - \nu)^{\frac{\nu}{\rho}} n^{\frac{\nu}{\rho}}$;
- $\rho < \nu$: $D_n \sim \frac{\nu - \rho}{a} n$;
- $\rho = \nu$: $D_n \sim \frac{\nu}{a} \frac{n}{\log(n)}$.

The frequency-rank distribution instead is always
$$f(r) \propto r^{-\frac{\rho}{\nu}}.$$

We have observed that PUT is conceptually suitable for our modeling, since it reproduces *richer-gets-richer* mechanism and explores the *adjacent possible*, and also seems to present results that could reproduce Heaps' and Zipf's laws; for this reason we will focus on this model in the following chapters.

Chapter 3

Stochastic approximation of the PUT model

In this chapter, the stochastic approximation [20], whose lemma is shown in appendix A, is applied to the already presented urn model in order to study the behavior of two specific statistics of the process. It will be obtained an ODE, whose stability will be studied in the second part of the chapter.

3.1 Stochastic approximation

In this section the stochastic process $(D_n^{(1)})_{n\geq 0}$, counting the number of distinct colors appeared in S up to time n, will be considered. For simplicity we will use the lighter notation $D_n = D_n^{(1)}$.

At first it is useful to calculate the total number of balls in the urn at time n that is given by

$$|U|_n = N_0 + \rho n + (\nu + 1)D_n;$$

Indeed, after each draw, ρ balls are added to the urn, while every time a new ball is drawn (it happens D_n times), further $\nu + 1$ balls are inserted. For the variant of the model in which the copies of the drawn ball are only inserted from the second draw of that color, in that case

$$|U|_n = N_0 + \rho(n - D_n) + (\nu + 1)D_n = N_0 + \rho n + (\nu + 1 - \rho)D_n.$$

In a general case it will be considered

$$|U|_n = N_0 + \rho n + aD_n$$

where $a = \nu + 1$ or $a = \nu + 1 - \rho$, depending on the model used.

Focusing on $(D_n)_{n\geq 0}$, it is a discrete time Markov chain, since the probability distribution at each time n of D_n depends only on its value at the previous time $n-1$ and not on the whole path $\{D_0, D_1, \dots, D_{n-1}\}$. It is non-homogeneous since it depends also on the time n. The process clearly starts with $D_0 = 0$, then the probability that at the $(n+1)$-th draw a ball with a never-seen-before color is drawn is

$$P\left(D_{n+1} = D_n + 1 | D_n, n\right) = \frac{N_0 + \nu D_n}{N_0 + \rho n + a D_n}.$$

At denominator there is the total number of balls in the urn while at the numerator there is the total number of balls of colors that have never been drawn: indeed, at the beginning, there are only N_0 never drawn balls and every time a new color is drawn it immediately becomes "old", decreasing by one this quantity, but it triggers the insertion of $\nu + 1$ new balls, with a net increase of ν new colors, therefore obtaining $N_0 + \nu D_n$.
Indeed the number of distinct drawn colors increases by one at time $n+1$ if a ball whose color has not appeared yet up to $S(n)$ is drawn. It can also be rewritten:

$$D_{n+1} = D_n + \xi_{n+1}, \qquad n = 0,1,\dots \tag{3.1}$$

where ξ_{n+1} is a $\{0,1\}$-valued random variable such that

$$\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = p(D_n, n) = P\left(D_{n+1} = D_n + 1 | D_n, n\right),$$

where $\mathcal{F}_n$ is the σ-algebra generated by the events of the process. Deterministically it will be obtained $D_1 = 1$, but after that the process will be completely stochastic.

After this part, it is our purpose to consider the number of times that the j-th color ever occurred in the sequence S has been drawn. This is a more complex issue, since the time of first occurrence of a color, the starting point of the process, is a random variable itself, except for $j = 1$, since the first color clearly appears at the first draw, at time $n = 1$. Therefore assume that $D_n \geq j - 1$ and the j-th color has at least one ball in the urn U (just one if $D_n = j - 1$ since it has not appeared yet, more balls if $D_n > j - 1$) and define the stochastic process $(K_n^j)_{n\geq 0}$ that counts the number of times the color has been drawn; due to the previous observations $K_n^j > 0 \Leftrightarrow D_n \geq j$ and also $D_n \geq j \Rightarrow n \geq j$. Following the same reasoning as for D_n it can be written

$$P(K_{n+1}^j = K_n^j + 1 | K_n^j, D_n, n) = \frac{\rho K_n^j + 1}{N_0 + \rho n + a D_n}, \qquad j \leq D_n + 1.$$

At the numerator this time there is the total number of balls of j-th color in the urn: the first one, which was present from the beginning or added with triggering, and ρ copies of it every time it has been drawn. In the case when at the first draw the copies are not introduced that probability is different:

$$P(K_{n+1}^j = K_n^j + 1 | K_n^j, D_n, n) = \frac{\rho(K_n^j - 1) + 1}{N_0 + \rho n + a D_n}, \qquad j \leq D_n,\ K_n^j \geq 1;$$

if $D_n = j-1$ then $K_n^j = 0$, then the expression before is not correct while the following is valid:

$$P(K_{n+1}^j = K_n^j + 1 | K_n^j = 0, D_n, n) = \frac{1}{N_0 + \rho n + aD_n}, \quad j = D_n + 1.$$

Of course, if $j > D_n + 1$, it is always true

$$P(K_{n+1}^j = K_n^j + 1 | K_n^j, D_n, n) = 0.$$

As done before, it will be useful to write the more general form

$$K_{n+1}^j = K_n^j + \delta_{n+1}, \quad j \leq D_n + 1,\ n = 0,1,\dots \tag{3.2}$$

where δ_{n+1} is a $\{0,1\}$-valued random variable such that $\mathbb{E}[\delta_{n+1}|\mathcal{F}_n] = q(K_n^j, D_n, n)$, with $q(K_n^j, D_n, n) = P(K_{n+1}^j = K_n^j + 1 | K_n^j, D_n, n)$ and $\mathcal{F}_n$ is the σ-algebra generated by the events of the process.

Considering alone the process $(K_n^j)_{n\geq 0}$ it is not a Markov chain since the conditional probability depends also on D_n, but considered together with $(D_n)_{n\geq 0}$, the whole process is a non-homogeneous Markov chain, as it can be deduced from the previous discussion.

Let us now consider the stochastic process $(X_n)_{n\geq 0}$, where $X_n = \frac{D_n}{n}$; it represents the fraction of distinct colors appeared in S with respect to all draws up to time n and due to this definition $X_n \in [0,1]$. Now, taking the expression in (3.1) and using the same approach described in Appendix A we can obtain the same form as in (A.2):

$$X_{n+1} = X_n + \frac{1}{n+1}\left[p(D_n, n) - X_n + \xi_{n+1} - p(D_n, n)\right]. \tag{3.3}$$

Using the conditional expectation in the (3.3) it is possible to write

$$\mathbb{E}\left[\frac{X_{n+1} - X_n}{(n+1)^{-1}} | \mathcal{F}_n\right] = \mathbb{E}\left[p(D_n, n) - X_n + \xi_{n+1} - p(D_n, n) | \mathcal{F}_n\right] = p(D_n, n) - X_n;$$

this expression is an expected increment and we would like to have a formal proof that, in order to asymptotically study the process $(X_n)_{n\geq 0}$, it can be approximated by the solution $x(t)$ of an ODE of the form

$$\dot{x} = p(x) - x.$$

The stochastic approximation presented in Appendix A gives a formal proof of that and we are going to show that the process considered satisfies the assumptions needed. This allows us to state and prove the following result.

Proposition 1. *Consider the process of PUT model and the normalizations of D_n and K_n^j, namely X_n and Y_n^j. Consider the function $\bar{b}(t) = \begin{pmatrix} \bar{x}(t) \\ \bar{y}(t) \end{pmatrix}$, which is a piecewise linear*

function such that $\bar{x}(t(n)) = X_n$ *and* $\bar{y}(t(n)) = Y_n^j$, $\forall n \geq 0$, *with* $t(n) = \sum_{k=0}^{n-1} \frac{1}{n+1}$; *finally consider* $b^m(t) = \begin{pmatrix} x^m(t) \\ y^m(t) \end{pmatrix}$ *and* $b_m(t) = \begin{pmatrix} x_m(t) \\ y_m(t) \end{pmatrix}$, $m \in \mathbb{N}$, *the solutions of the ODE system*

$$\begin{cases} \dot{x}(t) = f(x(t), y(t)) = \dfrac{x(t)(\nu - \rho - ax(t))}{\rho + ax(t)}; \\ \dot{y}(t) = g(x(t), y(t)) = \dfrac{-ax(t)y(t)}{\rho + ax(t)} \end{cases} \tag{3.4}$$

respectively 'starting' and 'ending' at $t(m)$ *with the condition* $b(t(m)) = \bar{b}(t(m))$.
Then, for any $T > 0$ *it holds*

$$\lim_{m\to\infty} \sup_{t\in[t(m),t(m)+T]} ||\bar{b}(t) - b^m(t)|| = 0 \quad a.s. \tag{3.5}$$

$$\lim_{m\to\infty} \sup_{t\in[t(m)-T,t(m)]} ||\bar{b}(t) - b_m(t)|| = 0 \quad a.s. \tag{3.6}$$

Proof. First of all we observe that $p(D_n, n) = \frac{N_0+\nu D_n}{N_0+\rho n+aD_n}$ depends on D_n and also on n; it is possible to substitute D_n with X_n very easily: $p(D_n, n) = p(X_n, n) = \frac{N_0/n+\nu X_n}{N_0/n+\rho+aX_n}$. However for our purpose we need p not dependent on n, but it is useful to rewrite (3.3) as follows:

$$X_{n+1} = X_n + \frac{1}{n+1}\left[\frac{\nu X_n}{\rho + aX_n} - X_n + \xi_{n+1} - p(X_n, n) + \epsilon_1(n)\right] \tag{3.7}$$

where

$$\epsilon_1(n) = p(X_n, n) - \frac{\nu X_n}{\rho + aX_n} = \frac{N_0(\rho + (a-\nu)X_n)}{(\rho + aX_n)(N_0 + n(\rho + aX_n))} \longrightarrow 0, \qquad n \to \infty.$$

Now the (3.7) can be rewritten in the form which is useful in order to apply stochastic approximation:

$$X_{n+1} = X_n + \frac{1}{n+1}\left[f(X_n) + M_{n+1}^{(1)} + \epsilon_1(n)\right],$$

where $f(X_n) = \dfrac{\nu X_n}{\rho + aX_n} - X_n$ and $M_{n+1}^{(1)} = \xi_{n+1} - p(X_n, n)$.

Following the same approach used for D_n, define $Y_n^j = \frac{K_n^j}{n} \in [0,1]$ $\forall n$ and rewrite the (3.2) in the following way, considering $q(X_n, Y_n^j, n)$ instead of $q(D_n, K_n^j, n)$, just dividing numerator and denominator by n:

$$Y_{n+1}^j = Y_n^j + \frac{1}{n+1}\left[g(X_n, Y_n^j) + M_{n+1}^{(2)} + \epsilon_2(n)\right]$$

where

$$g(X_n, Y_n^j) = \frac{\rho Y_n^j}{\rho + aX_n} - Y_n^j,$$

$$M_{n+1}^{(2)} = \delta_{n+1} - q(X_n, Y_n^j, n),$$

$$\epsilon_2(n) = q(X_n, Y_n^j, n) - \frac{\rho Y_n^j}{\rho + aX_n}$$

and again it possible to show that $\epsilon_2(n) \to 0$ when $n \to \infty$.
Clearly the second process depends on the first one, so they can be considered as a whole process

$$\begin{bmatrix} X_{n+1} \\ Y_{n+1}^j \end{bmatrix} = \begin{bmatrix} X_n \\ Y_n^j \end{bmatrix} + \frac{1}{n+1} \left(\begin{bmatrix} f(X_n, Y_n^j) \\ g(X_n, Y_n^j) \end{bmatrix} + \begin{bmatrix} M_{n+1}^{(1)} \\ M_{n+1}^{(2)} \end{bmatrix} + \begin{bmatrix} \epsilon_1(n) \\ \epsilon_2(n) \end{bmatrix} \right) \tag{3.8}$$

where f in fact depends only on the first variable.
We can now prove that the process described in (3.8) is of the form from equation (A.4), $B_{n+1} = B_n + a(n)[h(Z_n) + M_{n+1} + \epsilon(n)]$, and satisfies all the hypothesis necessary for Lemma 1.

1. The map h should be Lipschitz: in this case $h(x,y) = \begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix}$, $h : [0,1]^2 \to \mathbb{R}^2$, can be proved to be a Lipschitz function in its domain. In fact since x and y should reproduce the behavior of X_n and Y_n^j it is sufficient to consider a domain in $[0,1]^2$. The functions f and g are fractional with polynomials at numerator and denominator so h is Lipschitz on a compact domain which does not include its discontinuity points (the zeros of denominator), which happens only when $x = -\frac{\rho}{a}$. It remains then to check that $-\frac{\rho}{a} \notin [0,1]$, which is true either when $a = \nu + 1$ or $a = \nu + 1 - \rho$, with ρ and ν positive integers.

2. The step-sizes $\{\frac{1}{n+1}\}_{n \geq 0}$ are positive scalars such that

$$\sum_{n=0}^{\infty} \frac{1}{n+1} = \infty, \quad \sum_{n=0}^{\infty} \left(\frac{1}{n+1} \right)^2 < \infty$$

   so they satisfy the requirement.

3. $\{M_n\}$, with $M_n = \begin{pmatrix} M_n^{(1)} \\ M_n^{(2)} \end{pmatrix}$ is a martingale difference sequence with respect to the increasing family of σ-fields

$$\mathcal{F}_n = \sigma(X_m, Y_m^j : m \leq n),$$

   which is easy to check:

$$\begin{aligned} E[M_{n+1}|\mathcal{F}_n] &= E\left[\begin{pmatrix} M_{n+1}^{(1)} \\ M_{n+1}^{(2)} \end{pmatrix} |\mathcal{F}_n \right] = E\left[\begin{pmatrix} \xi_{n+1} - p(X_n, n) \\ \delta_{n+1} - q(X_n, Y_n^j, n) \end{pmatrix} |\mathcal{F}_n \right] \\ &= \begin{pmatrix} p(X_n, n) - p(X_n, n) \\ q(X_n, Y_n^j, n) - q(X_n, Y_n^j, n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

a.s., $n \geq 0$. Indeed for both ξ_{n+1} and δ_{n+1} the conditional probability distribution is a Bernoulli of parameters $p(X_n, n)$ and $q(X_n, Y_n^j, n)$, which are $\mathcal{F}_n$-measurable and also represent the conditional expected values.
$\{M_n\}$ are also square-integrable:

$$\begin{aligned} E[||M_{n+1}||^2|\mathcal{F}_n] &= E[\xi_{n+1}^2 - 2\xi_{n+1}p(X_n,n) + p(X_n,n)^2 \\ &\quad + \delta_{n+1}^2 - 2\delta_{n+1}q(X_n,Y_n^j,n) + q(X_n,Y_n^j,n)^2|\mathcal{F}_n] \\ &= p(X_n,n) - p(X_n,n)^2 + q(X_n,Y_n^j,n) - q(X_n,Y_n^j,n)^2 \leq 2 \\ &\leq 2(1+\|X_n\|^2 + \|Y_n^j\|^2) = 2(1+\|B_n\|^2) \end{aligned}$$

then $K = 2$.

4. It should hold $\sup_n \|B_n\| = \sup_n \left\| \begin{pmatrix} X_n \\ Y_n^j \end{pmatrix} \right\| < \infty$, which is true since X_n and Y_n^j are between 0 and 1 for definition.

5. The last condition is that $\epsilon(n) = \begin{pmatrix} \epsilon_1(n) \\ \epsilon_2(n) \end{pmatrix} \longrightarrow 0$ when $n \to \infty$ and it follows directly from the construction of $\epsilon_1(n)$ and $\epsilon_2(n)$, as seen before.

This shows that Assumption 1 in Appendix A is satisfied so that the claim follows from Lemma 1. □

By studying the equilibria of the ODE system (3.4), which tracks the values of X_n and Y_n^j, it should be possible to give an approximation of the behavior of two processes when n is large, and consequently of D_n and K_n^j.

3.2 Analysis of the approximating ODEs

The aim of this section is to study the equilibria of the ODE (3.4) with respect to the possible values of ν and ρ.
At first let us focus on analyzing the equilibrium points of the first ODE of the system (3.4) and their stability, which is autonomous and does not depend on y.

$$\dot{x} = \frac{x(\nu - \rho - ax)}{\rho + ax} = f(x). \tag{3.9}$$

The zeros of $f(x)$ are $x = 0$ and $x = \frac{\nu-\rho}{a}$ and the points both do not create problems with the denominator, which is positive in both cases.
In order to determine their stability it is necessary to calculate the derivative of $f(x)$:

$$f'(x) = \frac{\rho(\nu-\rho) - 2\rho ax - a^2x^2}{(\rho+ax)^2};$$

it results $f'(0) = \frac{\nu-\rho}{\rho}$ and $f'(\frac{\nu-\rho}{a}) = \frac{\rho-\nu}{\nu}$. In the first case $x = 0$ is an asymptotic equilibrium point if and only if $\nu < \rho$, while in the second one $x = \frac{\nu-\rho}{a}$ is an asymptotic equilibrium point if and only if $\nu > \rho$. In this last case it is convenient to check whether $\frac{\nu-\rho}{a} \in [0,1]$: indeed either if $a = \nu + 1$ or $a = \nu + 1 - \rho$ the condition is satisfied when $\nu \geq \rho$ are positive integers.
What happens instead when $\nu = \rho$?

$$f(x) = \frac{-ax^2}{\nu + ax} \qquad f'(x) = \frac{-2\nu ax - a^2x^2}{(\nu + ax)^2};$$

the only equilibrium point is $x = 0$ but it is not possible to determine its stability from the first order derivative since it is equal to 0. It is necessary then to calculate the second order derivative:

$$f''(x) = \frac{-2a\nu^2}{(\nu + ax)^3}.$$

In order to check the stability we observe that $f''(0) = \frac{-2a}{\nu} < 0$ since either $a = \nu + 1 > 0$ or $a = \nu - \rho + 1 = 1 > 0$; therefore even in this particular case $x = 0$ is an asymptotically stable point.
Summarizing:

- if $\nu \leq \rho$ the only asymptotically stable point is $x = 0$, then for any initial condition $x_0 \in [0,1]$ the solution of the Cauchy problem will converge to 0 (figures 3.1 and 3.2);
- if $\nu > \rho$ the only asymptotically stable point is $x = \frac{\nu-\rho}{a} \in (0,1]$, then for any initial condition $x_0 \in [0,1]$ the solution of the Cauchy problem will converge to $\frac{\nu-\rho}{a}$ (figures 3.3 and 3.4).

Regarding the second ODE in (3.4), it is necessary to study the whole system since the second equation depends also on x:

$$\begin{cases} \dot{x} = \dfrac{x(\nu - \rho - ax)}{\rho + ax}; \\ \dot{y} = \dfrac{-axy}{\rho + ax}. \end{cases} \tag{3.10}$$

The first equation has already been studied, so we can use the obtained results for finding equilibria for both. First of all it is necessary to calculate the Jacobian matrix $J_h(x, y)$, in order to check that the eigenvalues in the equilibrium points are negative:

$$J_h(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix} = \begin{bmatrix} \dfrac{\rho(\nu-\rho) - 2\rho ax - a^2x^2}{(\rho + ax)^2} & 0 \\ \dfrac{-ay(\rho + ax) + a^2xy}{(\rho + ax)^2} & \dfrac{-ax}{\rho + ax} \end{bmatrix} \tag{3.11}$$

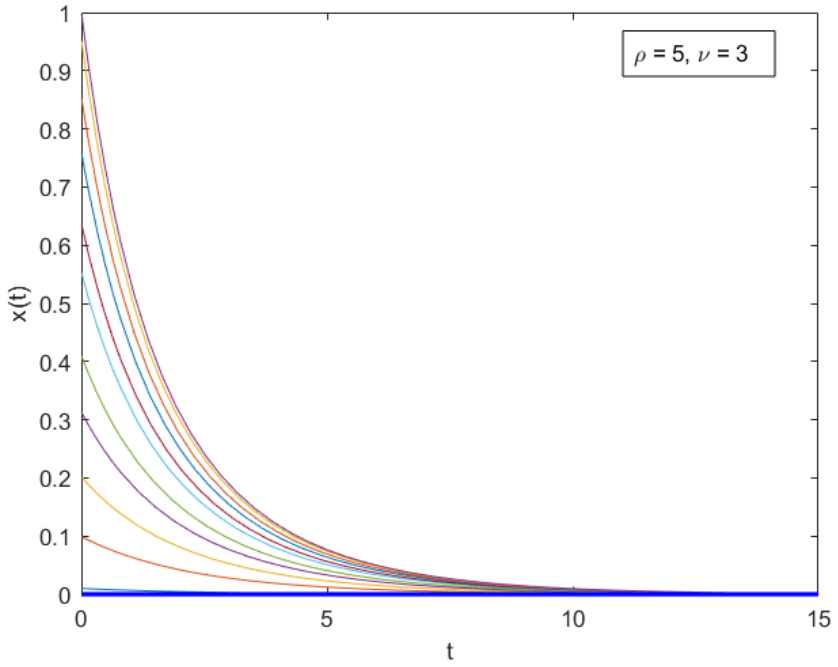


Figure 3.1: Time plot of solutions of the ODE in (3.9) in the case $\rho > \nu$: for any initial condition $x(0) \in [0,1]$, $x(t)$ exponentially goes to 0.

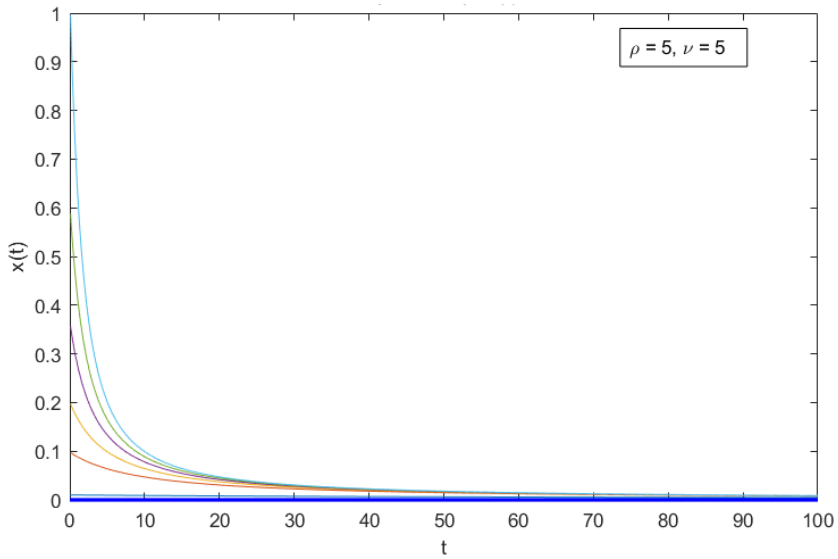


Figure 3.2: Time plot of solutions of the ODE in (3.9) in the case $\nu = \rho$: for any initial condition $x(0) \in [0,1]$, again $x(t)$ goes to 0, but slower.

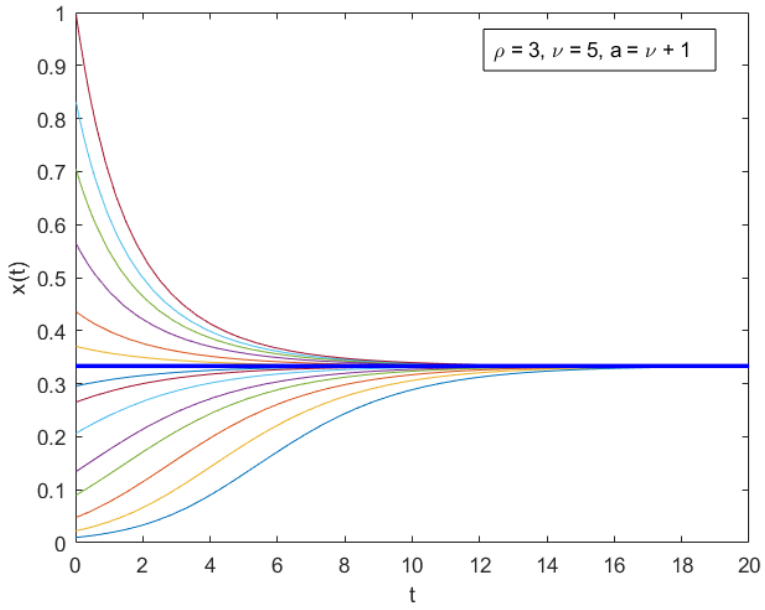


Figure 3.3: Time plot of solutions of the ODE in (3.9) in the case $\nu > \rho$: for any initial condition $x(0) \in (0,1]$, $x(t)$ converges to $\frac{\nu-\rho}{a}$; in this case $a = \nu + 1$ and $\frac{\nu-\rho}{a} = \frac{1}{3}$.

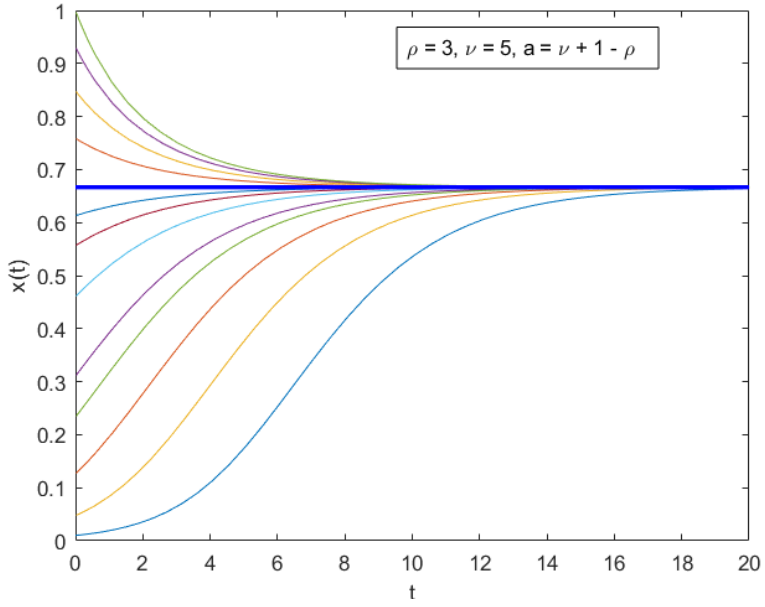


Figure 3.4: Time plot of solutions of the ODE in (3.9) in the case $\nu > \rho$: for any initial condition $x(0) \in (0,1]$, $x(t)$ converges to $\frac{\nu-\rho}{a}$; in this case $a = \nu + 1 - \rho$ and $\frac{\nu-\rho}{a} = \frac{2}{3}$.

- When $\nu > \rho$, in $x = \frac{\nu-\rho}{a}$ there is the only stable equilibrium for the first equation; for this reason the only way to have also the second equation equal to zero is to set $y = 0$. The stability is checked calculating the Jacobian in the equilibrium point and observing that both eigenvalues are negative, since $\rho < \nu$:

$$J_h\left(\frac{\nu-\rho}{a},0\right) = \begin{bmatrix} \frac{\rho-\nu}{\nu} & 0 \\ 0 & \frac{\rho-\nu}{\nu} \end{bmatrix}.$$

  We can see the result from the phase plot in figure 3.5.

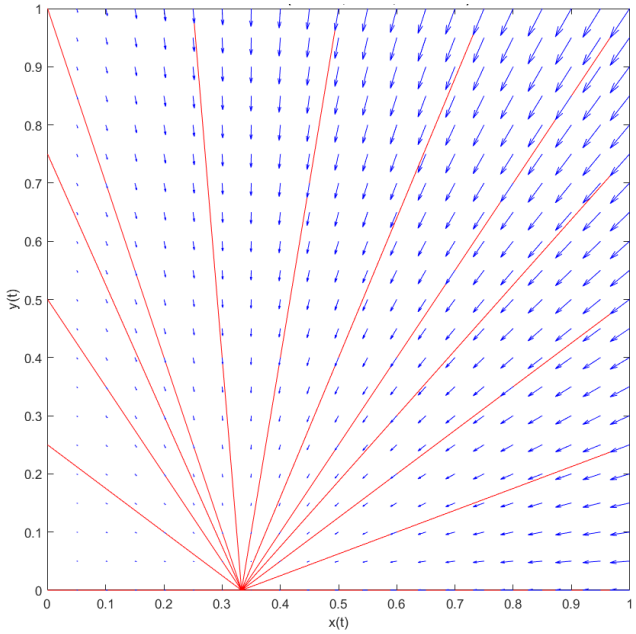


Figure 3.5: Phase plot in the case $5 = \nu > \rho = 3$ and $a = \nu + 1$: for any initial condition $(x(0), y(0)) \in (0,1] \times [0,1]$, the solution of the ODE in (3.10) converges to $(\frac{\nu-\rho}{a} = \frac{1}{3},0)$. The red lines are some trajectories, which are linear.

- When $\nu \leq \rho$ the only stable equilibrium for the first equation is when $x = 0$, which leads to an equilibrium also for the second one, for any value of y. The Jacobian gives only one negative eigenvalue, while the other is equal to zero:

$$J_h\left(0, y\right) = \begin{bmatrix} \frac{\nu-\rho}{\rho} & 0 \\ \frac{-ay}{\rho} & 0 \end{bmatrix}.$$

  For this reason it is expected that the asymptotic value of $y(t)$ depends on the initial condition (x_0, y_0); therefore it is necessary to determine the equilibrium value as a function of the initial point. In order to determine the trajectories observe that

substituting $\frac{x}{\rho+ax}$ in the second equation using the first one, that the latter ODE can be rewritten as

$$\dot{y} = -ay\frac{\dot{x}}{\nu-\rho-ax}$$

from which we obtain

$$\frac{\dot{y}}{\dot{x}} = \frac{dy}{dx} = \frac{-ay}{\nu-\rho-ax}. \tag{3.12}$$

This is a Cauchy problem where the result is a function $y(x)$ and the initial condition is $y(x_0) = y_0$.

It follows that the solution for the trajectory is $y(x) = \frac{y_0(\nu-\rho-ax)}{\nu-\rho-ax_0}$. It is interesting to notice that this trajectory is also valid for the case $\nu > \rho$ and it is easy to check that $y(x = \frac{\nu-\rho}{a}) = 0\ \forall(x_0, y_0) \in [0,1] \times [0,1]$, as found before.

When $\nu \leq \rho$ the value of x converges to 0, therefore the asymptotic value of y is given by $y(x = 0) = \frac{y_0(\nu-\rho)}{\nu-\rho-ax_0}$. When $\nu = \rho$ then $y(x = 0) = 0$, otherwise the numerator is negative and we need to verify that $y(x = 0) \in [0,1]$.

We should check that $\frac{y_0(\nu-\rho)}{\nu-\rho-ax_0} \in [0,1]$ and since we are considering the case $\nu < \rho$ the numerator is negative and it should be larger (smaller absolute value) than the denominator:

$$\begin{aligned} y_0(\nu-\rho) > \nu-\rho-ax_0 \quad &\iff \\ y_0 \leq 1 + \frac{a}{\rho-\nu}x_0. \end{aligned} \tag{3.13}$$

This condition is always true when $a \geq 0$, but in the case where $a = \nu+1-\rho < 0$ there are some initial conditions that lead to an asymptotic value of y which is greater than 1: we will see later that for our purpose the initial conditions, which are at sufficiently large times, cannot be considered among all values $(x_0, y_0) \in [0,1] \times [0,1]$ and therefore this condition will hold. The phase plots in figures 3.6, 3.7 and 3.8 show three different cases for this result.

The analysis and discussion of this section can be summarized in Proposition 2.

Proposition 2. *Consider the dynamical system described by the ODE* (3.4)*, then:*

(i) if $\nu > \rho$, $E = \{\frac{\nu-\rho}{a}\} \times \{0\}$ *is the set of equilibria of the system and* $\forall(x(0), y(0)) \in [0,1]^2$ *the solution converges to the point* $\left(\frac{\nu-\rho}{a},0\right)$*;*

(ii) if $\nu = \rho$, $E = \{0\} \times \{0\}$ *is the set of equilibria of the system and* $\forall(x(0), y(0)) \in [0,1]^2$ *the solution converges to the point* $(0,0)$*;*

(iii) if $\nu < \rho$*, given an initial condition* $(x(0), y(0)) \in [0,1]^2$ *the solution converges to* $\left(0, \frac{y(0)(\nu-\rho)}{\nu-\rho-ax(0)}\right)$*, therefore* $E = \{0\} \times [0, \rho-\nu]$ *is the set of equilibria of the system; in order to have* $y(t) \in [0,1]\ \forall t \geq 0$ *the initial condition should satisfy* $y(0) \leq 1 + \frac{a}{\rho-\nu}x(0)$*, which is always true only in the model where* $a = \nu + 1$*.*

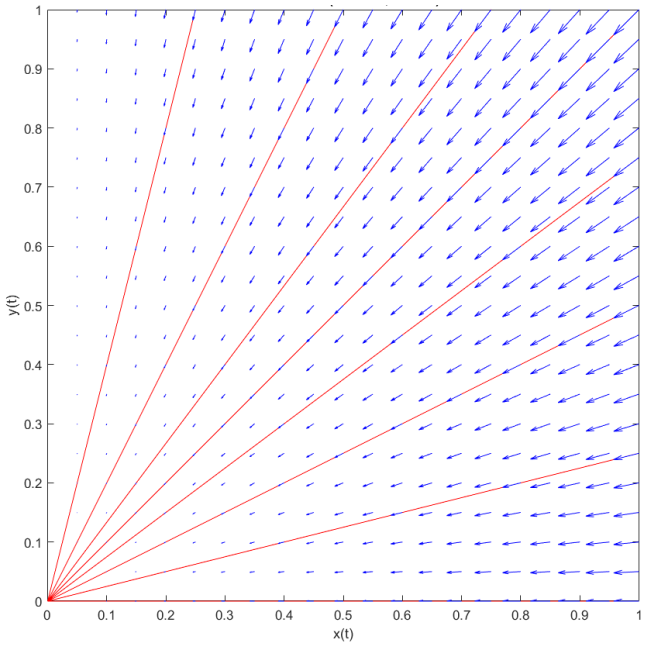


Figure 3.6: Phase plot in the case $\nu = \rho = 5$: for any initial condition $(x(0), y(0)) \in [0,1] \times [0,1]$ the solution of the ODE converges to (0,0). The red lines are some trajectories, which are linear.

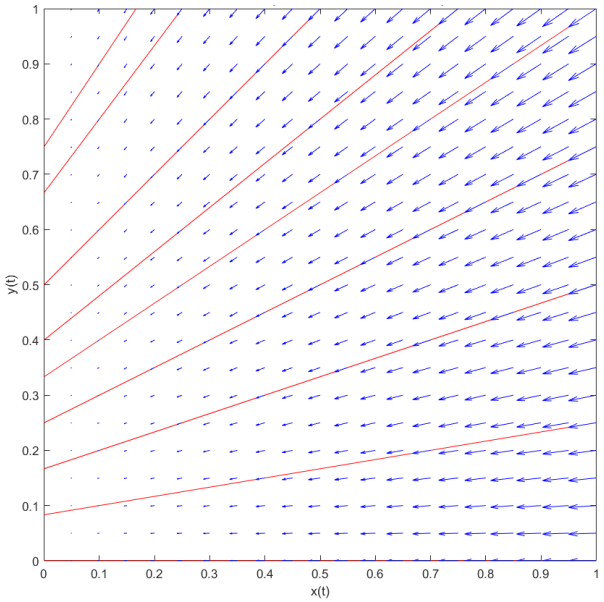


Figure 3.7: Phase plot in the case $3 = \nu < \rho = 5$ and $a = \nu + 1 = 4 > 0$: based on the initial condition $(x(0), y(0)) \in [0,1] \times [0,1]$ the solution of the ODE converges to a different value, where $x = 0$ and $y \in [0,1]$. The red lines are some trajectories, which are linear.

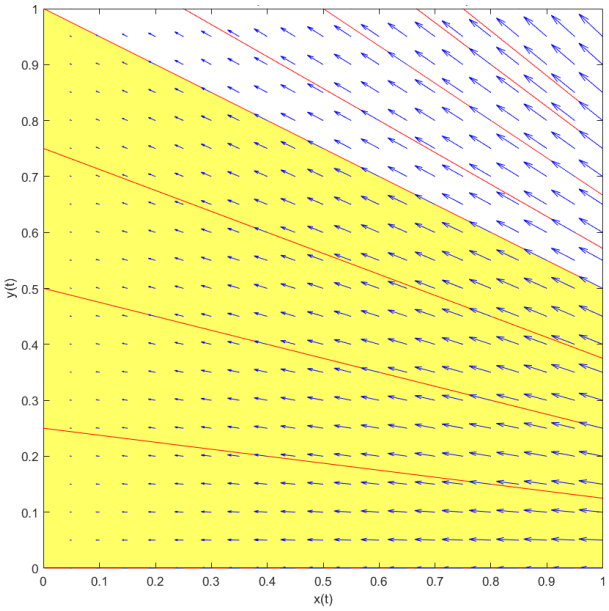


Figure 3.8: Phase plot in the case $3 = \nu < \rho = 5$ and $a = \nu - \rho + 1 = -1$: based on the initial condition $(x(0), y(0)) \in [0,1] \times [0,1]$ the solution of the ODE converges to a different value, where $x = 0$ and $y \in [0,1]$ only when the initial condition is in the yellow region. We will show that for our purpose the initial condition will be in that area. The red lines are some trajectories, which are linear.

Chapter 4

Main results

In this chapter, we build on the results obtained in Chapter 3 in order to find an approximation for the asymptotic behavior of the number of distinct elements D_n and on the frequency distribution K_n^j ($j >> 1$). After this, we generalize these results by studying the behavior of the variable $D_n^{(k)}$, counting the distinct colors drawn at least k times, for arbitrary k. In the last section, there is a study with the aim of maximizing $D_n^{(2)}$ and $D_n^{(3)}$, in the case $\nu > \rho$ and the sum $\nu + \rho$ is fixed.

4.1 Heaps' law

Due to the results in Chapter 3 it is now possible to describe the asymptotic behavior of the process $(D_n)_{n\geq 0}$.

Theorem 1. *Consider the process of PUT's model with parameters $\nu, \rho \in \mathbb{N}$. The statistic $(D_n)_{n\geq 0}$, which counts the number of distinct colors drawn up to each time n, has the following asymptotic behavior:*

(i) when $\nu > \rho$

$$D_n \sim \frac{\nu - \rho}{a} n, \quad \text{as } n \to \infty;$$

(ii) when $\nu < \rho$

$$D_n \asymp n^{\frac{\nu}{\rho}} \quad \text{as } n \to \infty;$$

(iii) when $\nu = \rho$

$$D_n \asymp \frac{\nu}{a} \frac{n}{\log(n)} \quad \text{as } n \to \infty.$$

Proof. Proposition 1 states that for very large times the function that interpolates the process $X_n = \frac{D_n}{n}$ is approximated by $x(t)$, where $x(t)$ is the solution of the Cauchy problem described by (3.9) with an initial condition at time $t(m)$, $m >> 1$, such that

$x(t(m)) = X_m$. Moreover Proposition 2 states that, for any initial condition in [0,1], $x(t)$ converges to a fixed equilibrium point, therefore we can just consider

$$X_n = \bar{x}(t(n)) \sim x(t(n)), \qquad n \to \infty$$

It is very important to notice that, as $n \to \infty$, $t(n) = \sum_{k=0}^{n-1} \frac{1}{k+1} \sim \log(n)$, so $X_n \sim x(\log(n))$ and

$$D_n = nX_n \sim nx(\log(n)).$$

Now we can use results from the analysis of ODE's convergence, from which we can infer the behavior of D_n when n is large. We distinguish again three cases based on the values of ρ and ν:

- When $\nu > \rho$ the asymptotic equilibrium point for the ODE in (3.9) is $x = \frac{\nu-\rho}{a}$ which means that, for large values of t, $x(t) \sim \frac{\nu-\rho}{a}$. Then

  $$D_n \sim \frac{\nu - \rho}{a} n,$$

  which means that the number of distinct colors drawn increases linearly with respect to time.

- When $\nu < \rho$, since the asymptotic equilibrium is in $x = 0$, it is convenient to use the McLaurin series to study how fast $x(t)$ goes to zero. Since $f(x) = f(0)+f'(0)x+o(x)$, as $x \to 0$, we solve then the ODE

  $$\dot{\tilde{x}} = f(0) + f'(0)\tilde{x} = \frac{\nu - \rho}{\rho}\tilde{x}, \tag{4.1}$$

  where $\tilde{x}$ is just to indicate that we are not really calculating $x(t)$ but an approximation when it is close to 0. The general solution is

  $$\tilde{x}(t) = \tilde{x}(0)e^{\frac{\nu-\rho}{\rho}t}.$$

  Therefore, we may write

  $$x(t) \asymp e^{\left(\frac{\nu}{\rho}-1\right)t}$$

  where the symbol "$\asymp$" has been given the following meaning:

  $$c_1 \leq \lim_{t\to+\infty} \frac{x(t)}{e^{\left(\frac{\nu}{\rho}-1\right)t}} \leq c_2$$

  where c_1 and c_2 are positive constants. We have then an expression for the asymptotic behavior of D_n:

  $$D_n \asymp ne^{\left(\frac{\nu}{\rho}-1\right)\log(n)} = n^{\frac{\nu}{\rho}}$$

  The behavior in this case is sub-linear, which is exactly what Heaps' law states.

- When $\nu = \rho$ the equilibrium is again in $x = 0$; we follow the same approach as the previous case but now the McLaurin series is until second order since $f'(0) = 0$:

$$f(x) = f(0) + f'(0)x + \frac{f''(0)x^2}{2} + o(x^2), \qquad x \to 0.$$

The approximating ODE is

$$\dot{\tilde{x}} = f(0) + f'(0)\tilde{x} + \frac{f''(0)\tilde{x}^2}{2} = \frac{-a}{\nu}\tilde{x}^2$$

whose solution is $\tilde{x}(t) = \dfrac{\nu}{at + c}$ for some constant $c \in \mathbb{R}$. It results then $x(t) \asymp \dfrac{\nu}{at}$ for large values of t. In this case we obtain

$$D_n \asymp \frac{\nu n}{a \log(n)};$$

the behavior is slower than the linear case, but faster than the case $\nu < \rho$, as one could expect.

□

4.2 Zipf's law

Our purpose is now to show that the tail of the frequency-rank distribution of the colors drawn from the urn, i.e. when j is large, follows Zipf's law with $\alpha = \frac{\varrho}{\nu}$.
We start then considering the j-th distinct color ever occurred in S, $j >> 1$; we start from the following approximation, obtained with the same considerations already done for X_n:

$$\frac{K_n^j}{n} = Y_n^j = \bar{y}(t(n)) \sim y(\log n),$$

and then we want to describe it as a function of j (and possibly n and D_n). In order to do that it is necessary to define the random variable

$$N_j = \min\{n \geq 0 : D_n = j\},$$

i.e. the time of first appearance of j-th color; using this definition it is possible to define an initial condition for the ODE solutions $x(t)$ and $y(t)$ for large times (the $t(m)$ of the Proposition 1). From the definition of N_j it follows

$$\begin{cases} X_{N_j} = \dfrac{D_{N_j}}{N_j} = \dfrac{j}{N_j}; \\ Y_{N_j}^j = \dfrac{K_{N_j}^j}{N_j} = \dfrac{1}{N_j} \end{cases}$$

and observing $X_{N_j} = x(t(N_j)) \sim x(\log N_j)$ and $Y^j_{N_j} = y(t(N_j)) \sim y(\log N_j)$ it is possible to set the initial conditions

$$\begin{cases} x(\log N_j) = \frac{j}{N_j}; \\ y(\log N_j) = \frac{1}{N_j}. \end{cases} \tag{4.2}$$

Now we can analyze the three cases depending on the respective values of ν and ρ. It is important to remember that those initial conditions are considered for j large, which means also the starting time N_j large.

Case $\nu > \rho$.

From the previous section we know that $D_n \sim \frac{\nu-\rho}{a} n$. By the definition of N_j, it results $j = D_{N_j} \sim \frac{\nu-\rho}{a} N_j$ and consequently we have that

$$N_j \sim \frac{a}{\nu - \rho} j$$

When $y \to 0$, i.e. it converges to the equilibrium, the Taylor series for $g(x, y)$ is

$$g(x,y) = g\left(\frac{\nu-\rho}{a},0\right) + \nabla_g \left(\frac{\nu-\rho}{a},0\right) \cdot \begin{bmatrix} x \\ y \end{bmatrix} + o\left(\left\| \begin{matrix} x \\ y \end{matrix} \right\|\right)$$

where ∇_g is the second row of J_h in (3.11). Considering now $\tilde{x}(t)$ and $\tilde{y}(t)$, approximations of $x(t)$ and $y(t)$ when they are close to the equilibrium point $(\frac{\nu-\rho}{a},0)$, we need to solve

$$\dot{\tilde{y}} = \begin{bmatrix} 0 & \frac{\rho-\nu}{\nu} \end{bmatrix} \cdot \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \frac{\rho - \nu}{\nu} \tilde{y},$$

whose solution is

$$\tilde{y}(t) = \tilde{y}(0) e^{\left(\frac{\rho}{\nu} - 1\right)t}.$$

Using initial condition from (4.2), intended for y but suitable also for its approximation $\tilde{y}$, we can rewrite

$$\tilde{y}(t) = \frac{1}{N_j} e^{\left(\frac{\rho}{\nu} - 1\right)(t - \log(N_j))}$$

It follows that, for sufficiently large n and j,

$$Y^j_n \sim y(\log(n)) \sim \tilde{y}(\log(n)) = \frac{n^{\frac{\rho}{\nu}}}{n N_j^{\frac{\rho}{\nu}}}$$

and then, substituting N_j with $\frac{a}{\nu-\rho} j$ and n with $\frac{a}{\nu-\rho} D_n$, we obtain the expression

$$Y^j_n \sim \frac{\left(\frac{a}{\nu-\rho}\right)^{\frac{\rho}{\nu}} D_n^{\frac{\rho}{\nu}}}{n \left(\frac{a}{\nu-\rho}\right)^{\frac{\rho}{\nu}} j^{\frac{\rho}{\nu}}} = \frac{1}{n} \left(\frac{j}{D_n}\right)^{-\frac{\rho}{\nu}} \propto j^{-\frac{\rho}{\nu}}. \tag{4.3}$$

Case $\nu < \rho$.

Using again the results from Section 4.1 and the definition of N_j we want to solve the same ODE for $\tilde{x}(t)$ as the one in (4.1), but with the initial condition for x in (4.2), obtaining

$$\tilde{x}(t) = \frac{j}{N_j} e^{\left(\frac{\nu}{\rho}-1\right)(t-\log(N_j))},$$

which leads to

$$D_n \sim nx(\log n) \sim n\tilde{x}(\log n) = \frac{j}{N_j^{\frac{\nu}{\rho}}} n^{\frac{\nu}{\rho}}.$$

Therefore it results $N_j \sim \left(\frac{j}{D_n}\right)^{\frac{\rho}{\nu}} n$. Moreover, we know that $y(t)$ converges to the value $\frac{y_0(\nu-\rho)}{\nu-\rho-ax_0}$, and using initial conditions from (4.2) we obtain

$$Y_n^j \sim \frac{\frac{1}{N_j}(\nu-\rho)}{\nu-\rho-a\frac{j}{N_j}} = \frac{1}{N_j - \frac{a}{\nu-\rho}j} \sim \frac{1}{j^{\frac{\rho}{\nu}} D_n^{-\frac{\rho}{\nu}} n - \frac{a}{\nu-\rho}j} = \frac{j^{-\frac{\rho}{\nu}}}{D_n^{-\frac{\rho}{\nu}} n - \frac{a}{\nu-\rho} j^{1-\frac{\rho}{\nu}}}.$$

As j grows large the term at denominator $\frac{a}{\nu-\rho} j^{1-\frac{\rho}{\nu}}$ goes to zero and becomes neglectable since $c_1 \leq D_n^{-\frac{\rho}{\nu}} n \leq c_2$ for c_1, c_2 positive constants, therefore we can conclude again that

$$Y_n^j \sim \frac{1}{n}\left(\frac{j}{D_n}\right)^{-\frac{\rho}{\nu}} \propto j^{-\frac{\rho}{\nu}}. \tag{4.4}$$

Special case $a = \nu + 1 - \rho < 0$

Previously, in the analysis of the ODEs, it was stated that only some initial condition in [0,1] × [0,1] led to an equilibrium value for y in the range [0,1]. Considering now initial conditions from (4.2) it can be proved that inequality in (3.13) is always verified even when $a = \nu - \rho + 1 < 0$.

$$\begin{aligned}
y_0 &\leq 1 + \frac{a}{\rho-\nu} x_0 \\
\frac{1}{N_j} &\leq 1 + \frac{\nu-\rho+1}{\rho-\nu}\frac{j}{N_j}; \\
\rho-\nu &\leq N_j(\rho-\nu) + (\nu-\rho+1)j; \\
0 &\leq (N_j-1)(\rho-\nu) + (\nu-\rho+1)j; \\
(\rho-\nu-1)j &\leq N_j(\rho-\nu); \\
1 - \frac{1}{\rho-\nu} &\leq \frac{N_j}{j} \sim j^{\frac{\rho}{\nu}-1} D_n^{-\frac{\rho}{\nu}} n
\end{aligned}$$

which is true for j sufficiently large, since $D_n^{-\frac{\rho}{\nu}} n$ is bounded ($D_n \asymp n^{\frac{\nu}{\rho}}$).

Case $\nu = \rho$.

In this case both $x(t)$ and $y(t)$ converge to 0. The asymptotic behavior for the former has already been calculated for D_n but without initial condition, while for the latter we need to proceed in the following way. Since we are considering the ODE $\dot{y} = g(x, y)$, in order to get the Taylor approximation we need again to calculate the $\nabla_g(x, y)$. It corresponds to the second row of $J_h(x, y)$ (3.11), calculated in (0,0): however it results to be null in both components, requiring the use of the Hessian matrix of g:

$$H_g(x,y)_{|(0,0)} = \begin{bmatrix} \dfrac{-2a^3xy}{(\nu+ax)^3} + \dfrac{2a^2x}{(\nu+ax)^2} & \dfrac{a^2x}{(\nu+ax)^2} + \dfrac{-a}{\nu+ax} \\ \dfrac{a^2x}{(\nu+ax)^2} + \dfrac{-a}{\nu+ax} & 0 \end{bmatrix}_{|(0,0)} = \begin{bmatrix} 0 & \dfrac{-a}{\nu} \\ \dfrac{-a}{\nu} & 0 \end{bmatrix}.$$

We now have to consider Taylor series of second order

$$g(x,y) = g\,(0,0) + \nabla_g\,(0,0) \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^T H_g(0,0) \begin{bmatrix} x \\ y \end{bmatrix} + o\left(\left\| \begin{matrix} x \\ y \end{matrix} \right\|^2 \right)$$

and solve

$$\dot{\tilde{y}} = \frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 0 & \dfrac{-a}{\nu} \\ \dfrac{-a}{\nu} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \frac{-a}{\nu} \tilde{x}\tilde{y}$$

with initial conditions from (4.2). Solving again $\dot{\tilde{x}} = \frac{-a}{\nu}\tilde{x}^2$ but also with conditions from (4.2) it becomes

$$\tilde{x}(t) = \frac{\nu}{at + \nu\frac{N_j}{j} - a\log(N_j)}$$

and it follows that

$$\tilde{y}(t) = \frac{1}{j} \frac{\nu}{at + \nu\frac{N_j}{j} - a\log(N_j)} = \frac{1}{j}\tilde{x}(t).$$

The value of $D_n \sim n\tilde{x}(\log(n))$ is

$$D_n \sim \frac{\nu n}{a\log(n) + \nu\frac{N_j}{j} - a\log(N_j)} \tag{4.5}$$

We can write then

$$Y_n^j \sim y(\log(n)) \sim \tilde{y}(\log(n)) = \frac{1}{j}\tilde{x}(\log(n)) \sim \frac{1}{j}\frac{D_n}{n}$$

obtaining again

$$Y_n^j \sim \frac{j^{-1}}{D_n^{-1}n} = \frac{1}{n}\left(\frac{j}{D_n}\right)^{-\frac{\rho}{\nu}} \propto j^{-\frac{\rho}{\nu}}. \tag{4.6}$$

The analysis and discussion of this section gives a proof to the following theorem.

Theorem 2. *Consider the process of PUT's model with parameters $\nu, \rho \in \mathbb{N}$. The statistic $(K_n^j)_{n\geq 0}$, which counts the number of balls of the j-th distinct color ever drawn up to time n, $j >> 1$, has the following asymptotic behavior:*

$$K_n^j = nY_n^j \sim \left(\frac{j}{D_n}\right)^{-\frac{\rho}{\nu}} \propto j^{-\frac{\rho}{\nu}}, \quad \text{as } n, j \to \infty.$$

Remarks and normalization

In every case we found out that the relative frequency of j-th color is proportional to $j^{-\frac{\rho}{\nu}}$. In the definition, j refers to the order at which a color is drawn for the first time with respect to the others. Indeed it can be observed in the aftermath that Y_n^j is decreasing in j, therefore j can also be considered as the rank of the color. This result suggests an important, maybe trivial, observation for the model: the sooner a color appears in the sequence, the more likely it is to be drawn with respect to the ones that appear later. This is a consequence of the *richer-gets-richer* mechanism.

At time n the maximum value that can be assumed by j is D_n (the lowest rank); then it should be true that $\sum_{j=1}^{D_n} Y_n^j = 1$. In order to check that we should see that the sum $\sum_{j=1}^{D_n} j^{-\frac{\rho}{\nu}}$ is equal to the normalization factor found in expressions (4.3), (4.4) and (4.6), i.e. $(D_n^{-\frac{\rho}{\nu}} n)^{-1}$. We will see that the results are not exactly the same but not completely different either; in fact the approximations considered only j large and the model is not reliable for principal ranks.

Let us define

$$H_{D_n,\frac{\rho}{\nu}} = \sum_{j=1}^{D_n} \frac{1}{j^{\frac{\rho}{\nu}}}$$

the generalized harmonic number of order D_n with exponent $\frac{\rho}{\nu}$. We should have

$$Y_n^j \sim \frac{j^{-\frac{\rho}{\nu}}}{H_{D_n,\frac{\rho}{\nu}}} \tag{4.7}$$

and therefore we have to check that normalization factors are at least similar to $H_{D_n,\frac{\rho}{\nu}}$.

- When $\nu > \rho$ and n is large we have

$$H_{D_n,\frac{\rho}{\nu}} \sim \int_1^{D_n} j^{-\frac{\rho}{\nu}} dj = \left[\frac{j^{1-\frac{\rho}{\nu}}}{1-\frac{\rho}{\nu}}\right]_1^{D_n} = \frac{\nu}{\nu-\rho}\left(D_n^{1-\frac{\rho}{\nu}} - 1\right) \sim \frac{\nu}{\nu-\rho} D_n^{1-\frac{\rho}{\nu}} \sim \frac{\nu}{a} D_n^{-\frac{\rho}{\nu}} n$$

  which is similar to what we expect but not the same. Observe that, as $n \to \infty$ (and so D_n), Y_n^j goes to zero for each j: it is an expected behavior due a larger triggering with respect to reinforcement.

- When $\nu < \rho$ and n is large we have

$$H_{D_n,\frac{\rho}{\nu}} \sim \int_1^{D_n} j^{-\frac{\rho}{\nu}} dj = \left[\frac{j^{1-\frac{\rho}{\nu}}}{1-\frac{\rho}{\nu}}\right]_1^{D_n} = \frac{\nu}{\nu-\rho}\left(D_n^{1-\frac{\rho}{\nu}} - 1\right) \sim \frac{\nu}{\rho-\nu}.$$

  Here the normalization factor converges to a finite value and therefore also the frequencies $\forall j$: the effect of a larger reinforcement than triggering is highlighted by this behavior.

- When $\nu = \rho$ and n is large we have

$$H_{D_n,1} \sim \int_1^{D_n} j^{-1} dj = [\log j]_1^{D_n} = \log(D_n);$$

  now we should check, comparing to (4.6), that $\log(D_n) \sim D_n^{-1} n$. Since $D_n \asymp \frac{\nu}{a}\frac{n}{\log(n)}$ it results

$$\log(D_n) \asymp \log\left(\frac{\nu}{a}\right) + \log(n) - \log\left(\log(n)\right) \sim \log(n)$$

  and

$$D_n^{-1} n \asymp \frac{a}{\nu}\log(n) \asymp \log(n).$$

  It follows that $\log(D_n) \asymp D_n^{-1} n$, which means that the inverse of normalization factor has a logarithmic behavior. Here again the frequency of j-th element goes to zero as n grows large, but slower than the case $\nu > \rho$, since there is a logarithmic term at denominator, instead of sub-linear.

4.3 Distinct colors with at least k draws

We will see in numerical simulations that, especially in the model with $a = \nu + 1 - \rho$, in which the copies of a color are inserted only starting from the second draw, there are a lot of colors drawn only once. These elements can be interpreted as unsuccessful innovations, whose discovery did not lead to further attention from the agent. For this reason it was developed the idea of minimizing the number of elements drawn only once or, better, maximizing the number of distinct elements drawn at least 2, 3 or more times.

Therefore we can consider the stochastic processes $(D_n^{(k)})_{n\geq 0}$, $k = 1,2,\ldots$, that count the number of distinct colors drawn at least k times: these are generalizations of $(D_n)_{n\geq 0} = (D_n^{(1)})_{n\geq 0}$, the process that represents the number of colors drawn at least once. At first we are interested in observing how the two processes $(D_n^{(2)})_{n\geq 0}$ and $(D_n^{(3)})_{n\geq 0}$ behave with large n, and after that we will give a recursion for $D_n^{(k)}$ as a fraction of $D_n^{(k-1)}$.

Let us consider now the process $(D_n^{(2)})_{n\geq 0}$: it starts with $D_0^{(2)} = 0$ and then it increases by one every time a ball in the urn, whose color has been drawn only once, is drawn again. The number of distinct colors drawn only once is given by the difference between the ones drawn at least once, D_n, minus the once drawn at least twice, $D_n^{(2)}$; for each of these colors

there is only one ball in the urn, since we are considering the model in which after the first draw there is no reinforcement and only the drawn ball is replaced in the urn. We can then define the conditional probability that the number of elements drawn at least twice increases by one with the next draw:

$$p_2(D_n, D_n^{(2)}, n) = P(D_{n+1}^{(2)} = D_n^{(2)} + 1 | D_n, D_n^{(2)}, n) = \frac{D_n - D_n^{(2)}}{N_0 + \rho n + aD_n};$$

Regarding $D_n^{(3)}$ we can use the same approach, reminding that it increases by one every time a color with exactly two draws is drawn again. This time we have to take into account that, with the second replacement, we introduce ρ copies of the same color into the urn, which means that every color with exactly two draws has $\rho + 1$ copies in the urn. The conditional probability that $D_n^{(3)}$ increases by one is

$$p_3(D_n, D_n^{(2)}, D_n^{(3)}, n) = P(D_{n+1}^{(3)} = D_n^{(3)} + 1 | D_n, D_n^{(2)}, D_n^{(3)}, n) = \frac{(D_n^{(2)} - D_n^{(3)})(\rho + 1)}{N_0 + \rho n + aD_n}.$$

Notice that the whole process $\left(\left[D_n, D_n^{(2)}, D_n^{(3)}\right]\right)_{n \geq 0}$ is a again a non-homogeneous discrete time Markov chain.

The idea now is to use the stochastic approximation again, but omitting all the details and formal proofs used before, summarized by the following steps for obtaining the ODE that track the process:

- define $X_n^{(1)} = \frac{D_n}{n}$, $X_n^{(2)} = \frac{D_n^{(2)}}{n}$, $X_n^{(3)} = \frac{D_n^{(3)}}{n}$;
- rewrite p_2 and p_3 as functions of $X_n^{(1)}$, $X_n^{(2)}$, $X_n^{(3)}$ and eliminate the term $\frac{N_0}{n}$, neglectable when n is large:

$$p_2(X_n^{(1)}, X_n^{(2)}, n) = \frac{X_n^{(1)} - X_n^{(2)}}{N_0/n + \rho + aX_n^1} \sim \frac{X_n^{(1)} - X_n^{(2)}}{\rho + aX_n^1};$$

$$p_3(X_n^{(1)}, X_n^{(2)}, X_n^{(3)}, n) = \frac{(X_n^{(2)} - X_n^{(3)})(\rho + 1)}{N_0/n + \rho n + aX_n^{(1)}} \sim \frac{(X_n^{(2)} - X_n^{(3)})(\rho + 1)}{\rho + aX_n^{(1)}};$$

- consider the corresponding continuous functions $x_1(t)$, $x_2(t)$ and $x_3(t)$ and define the ODE system, where each equation is of the form $\dot{x_i} = p_i(x_1, x_{i-1}, x_i) - x_i$ and the first one is the same studied for D_n alone:

$$\begin{cases} \dot{x_1} = \dfrac{\nu x_1}{\rho + ax_1} - x_1 = f_1(x_1) \\ \dot{x_2} = \dfrac{x_1 - x_2}{\rho + ax_1} - x_2 = f_2(x_1, x_2) \\ \dot{x_3} = \dfrac{(x_2 - x_3)(\rho + 1)}{\rho + ax_1} - x_3 = f_3(x_1, x_2, x_3). \end{cases} \tag{4.8}$$

ODE analysis

In order to study the stability points it is defined the Jacobian matrix of the right term of the ODE system (4.8):

$$J_{[f_1\ f_2\ f_3]}(x_1,x_2,x_3)=\begin{bmatrix}\frac{\nu(\rho+ax_1)-a\nu x_1}{(\rho+ax_1)^2}-1 & 0 & 0\\ \frac{\rho+ax_2}{(\rho+ax_1)^2} & -\frac{1}{\rho+ax_1}-1 & 0\\ \frac{-a(x_2-x_3)(\rho+1)}{(\rho+ax_1)^2} & \frac{\rho+1}{\rho+ax_1} & -\frac{\rho+1}{\rho+ax_1}-1\end{bmatrix} \tag{4.9}$$

The different interesting cases, $\nu>\rho$ and $\nu<\rho$, are then analyzed.

- When $\nu>\rho$ it has already been found the equilibrium $x_1=\frac{\nu-\rho}{a}$ and, substituting in (4.8), the resulting equilibrium point is

$$\begin{cases}x_1=\frac{\nu-\rho}{a};\\ x_2=\frac{\nu-\rho}{a}\cdot\frac{1}{\nu+1};\\ x_3=\frac{\nu-\rho}{a}\cdot\frac{1}{\nu+1}\cdot\frac{1+\rho}{1+\rho+\nu}.\end{cases} \tag{4.10}$$

  In order to check the stability of the point we substitute these values in the Jacobian (4.9):

$$J_{[f_1\ f_2\ f_3]}\left(\tfrac{\nu-\rho}{a},\tfrac{\nu-\rho}{a}\tfrac{1}{\nu+1},\tfrac{\nu-\rho}{a}\tfrac{1}{\nu+1}\tfrac{1+\rho}{1+\rho+\nu}\right)=\begin{bmatrix}\frac{\rho}{\nu}-1 & 0 & 0\\ \frac{\rho}{\nu(\nu+1)} & -\frac{1}{\nu}-1 & 0\\ \frac{(\rho+1)(\rho-\nu)}{\nu(\nu+1)(\nu+1+\rho)} & \frac{\rho+1}{\nu} & -\frac{\rho+1}{\nu}-1\end{bmatrix}.$$

  The eigenvalues, represented by the elements on diagonal of the matrix, are all negative, therefore the equilibrium is asymptotically stable and for any initial condition the solution converges to the point in (4.10).
  The three quantities D_n, D_n^2 and D_n^3 are obtained with the usual approach, knowing already that $D_n\sim\frac{\nu-\rho}{a}n$:

$$D_n^{(2)}=nX_n^{(2)}\sim nx_2(\log(n))\sim\frac{\nu-\rho}{a}\cdot\frac{1}{\nu+1}n\sim\frac{1}{\nu+1}D_n;$$

$$\begin{aligned}D_n^{(3)}=nX_n^{(3)}&\sim nx_3(\log(n))\sim\frac{\nu-\rho}{a}\cdot\frac{1}{\nu+1}\cdot\frac{1+\rho}{1+\rho+\nu}n\\&\sim\frac{1+\rho}{1+\rho+\nu}D_n^{(2)}\sim\frac{1}{\nu+1}\cdot\frac{1+\rho}{1+\rho+\nu}D_n.\end{aligned}$$

- When $\nu<\rho$, it has already been found the equilibrium $x_1=0$, with asymptotic behavior, as $t\to\infty$,

$$x_1(t)\asymp e^{\left(\frac{\nu}{\rho}-1\right)t}.$$

$x_1 = 0$ implies also $x_2 = x_3 = 0$, which substituted in (4.8) give the equilibrium point. Checking the stability with the Jacobian matrix:

$$J_{[f_1\ f_2\ f_3]}(0,0,0) = \begin{bmatrix} \frac{\nu}{\rho} - 1 & 0 & 0 \\ \frac{1}{\rho} & -\frac{1}{\rho} - 1 & 0 \\ 0 & \frac{1}{\rho} + 1 & -\frac{1}{\rho} - 2 \end{bmatrix}.$$

Again, on the diagonal there are negative eigenvalues. It confirms that, for any initial condition, the solution of ODEs in (4.8) converges to (0,0,0).
In order to study the asymptotic behavior of $x_2(t)$ and $x_3(t)$ we introduce again $\tilde{x_1}$, $\tilde{x_2}$ and $\tilde{x_3}$ that approximate the function when their values are close to the equilibrium point. The system to solve is

$$\begin{bmatrix} \dot{\tilde{x_1}} \\ \dot{\tilde{x_2}} \\ \dot{\tilde{x_3}} \end{bmatrix} = J_{[f_1\ f_2\ f_3]}(0,0,0) \cdot \begin{bmatrix} \tilde{x_1} \\ \tilde{x_2} \\ \tilde{x_3} \end{bmatrix}$$

that becomes

$$\begin{cases} \dot{\tilde{x_1}} = \left(\frac{\nu}{\rho} - 1\right) \tilde{x_1}; \\ \dot{\tilde{x_2}} = \frac{1}{\rho}\tilde{x_1} - \left(\frac{1}{\rho} + 1\right) \tilde{x_2}; \\ \dot{\tilde{x_3}} = \left(\frac{1}{\rho} + 1\right) \tilde{x_2} - \left(\frac{1}{\rho} + 2\right) \tilde{x_3}; \end{cases}$$

the generic solutions, written in a convenient recursive and approximated form, are

$$\begin{cases} \tilde{x_1}(t) = c_1 e^{\left(\frac{\nu}{\rho} - 1\right)t}; \\ \tilde{x_2}(t) = \frac{1}{\nu+1}\tilde{x_1}(t) + c_2 e^{-\left(\frac{1}{\rho}+1\right)t} \sim \frac{1}{\nu+1}\tilde{x_1}(t); \\ \tilde{x_3}(t) \sim \frac{\rho+1}{\rho+1+\nu}\tilde{x_2}(t) + c_3 e^{-\left(\frac{1}{\rho}+2\right)t} \sim \frac{\rho+1}{\rho+1+\nu}\tilde{x_2}(t) \end{cases}$$

where c_1, c_2 and c_3 are constants depending on the initial condition.
Again here we observe that the fraction of elements drawn at least two or three times with respect to all the distinct elements, is the same found for $\nu > \rho$:

$$\begin{cases} x_2(t) \sim \frac{1}{\nu+1} x_1(t) \\ x_3(t) \sim \frac{1+\rho}{1+\rho+\nu} \cdot x_2(t) \sim \frac{1+\rho}{1+\rho+\nu} \cdot \frac{1}{\nu+1} \cdot x_1(t) \end{cases} \tag{4.11}$$

as $t \to \infty$.

In the next section we are going to find a formula for $D_n^{(k)}$ $\forall k \geq 1$, focusing on the case $\nu > \rho$.

General case $k \geq 1$ when $\nu > \rho$

From now on we decide to focus on the case $\nu > \rho$, since the behavior of D_n, $D_n^{(2)}$ and $D_n^{(3)}$ is more predictable than the other case, in which everything becomes neglectable as n grows large. We want now to generalize the expression for the number of distinct elements drawn at least k times $D_n^{(k)}$. The aim is to find a recursion that gives the fraction of elements with k draws minimum with respect to the ones drawn at least $k-1$ times, as the following theorem states.

Theorem 3. *Consider the process of PUT's model with parameters $\nu, \rho \in \mathbb{N}$, in the case $\nu > \rho$ and $a = \nu + 1 - \rho$. The statistic $(D_n^{(k)})_{n\geq 0}$, which counts the number of distinct colors with at least k draws up to time n, is approximated by the recursion*

$$D_n^{(k)} \sim \frac{(k-2)\rho + 1}{(k-2)\rho + \nu + 1} D_n^{(k-1)}, \quad \text{as } n \to \infty \tag{4.12}$$

and, considering $D_n^{(1)} = D_n \sim \frac{\nu-\rho}{a} n$, it has the non-recursive formula:

$$D_n^{(k)} \sim D_n \prod_{i=2}^{k} \frac{(i-2)\rho + 1}{(i-2)\rho + \nu + 1} \sim \frac{\nu - \rho}{a} n \prod_{i=2}^{k} \frac{(i-2)\rho + 1}{(i-2)\rho + \nu + 1}, \quad \text{as } n \to \infty. \tag{4.13}$$

Proof. The conditional probability that $D_n^{(k)}$ increases by 1 at time n is

$$P(D_{n+1}^{(k)} = D_n^{(k)} + 1 | D_n, D_n^{(k-1)}, D_n^{(k)}, n) = \frac{(D_n^{(k-1)} - D_n^{(k)})((k-2)\rho + 1)}{N_0 + \rho n + a D_n};$$

indeed at the numerator there is the number of balls of colors with exactly $k-1$ draws, replaced with ρ copies only from the second draw, i.e. $k-2$ times. Defining $X_n^{(k)} = \frac{D_n^{(k)}}{n}$, $X_n^{(k-1)} = \frac{D_n^{(k-1)}}{n}$ and their respective approximating functions $x_k(t)$ and $x_{k-1}(t)$, the corresponding ODE for $X_n^{(k)}$ with the stochastic approximation is

$$\dot{x_k} = \frac{(x_{k-1} - x_k)((k-2)\rho + 1)}{\rho + a x_1} - x_k = f_k(x_1, x_{k-1}, x_k). \tag{4.14}$$

The asymptotic equilibrium point for x_1 is always $\frac{\nu-\rho}{a}$. Supposing to have an asymptotic equilibrium also for x_{k-1}, it results that, in order to have an equilibrium for the ODE (4.14),

$$x_k = \frac{(k-2)\rho + 1}{(k-1)\rho + 1 + a x_1} x_{k-1} = \frac{(k-2)\rho + 1}{(k-2)\rho + \nu + 1} x_{k-1}.$$

The stability is checked just calculating the partial derivative of f_k in (4.14) with respect to x_k . It just remains to check that the derivative is negative in the equilibrium point. In fact, since f_k does not depend on x_m with $m > k$, the resulting Jacobian matrix is lower triangular and the elements on the diagonal are eigenvalues:

$$\frac{\partial f_k}{\partial x_k} = \frac{-((k-2)\rho + 1)}{\rho + a x_1} - 1 = \frac{-((k-2)\rho + 1)}{\nu} - 1 < 0.$$

We can infer then that

$$D_n^{(k)} = nX_n^{(k)} \sim nx_k(\log(n)) \sim n\frac{(k-2)\rho+1}{(k-2)\rho+\nu+1}x_{k-1}(\log(n))$$

$$\sim n\frac{(k-2)\rho+1}{(k-2)\rho+\nu+1}X_n^{(k-1)} \sim \frac{(k-2)\rho+1}{(k-2)\rho+\nu+1}D_n^{(k-1)}.$$

Since, when $\nu > \rho$, $D_n \sim \frac{\nu-\rho}{a}$ we obtain also the full expression in (4.13). □

4.4 Optimization

In PUT's model, the success of a color is clearly defined by the number of times it is drawn; in a broad view, for example considering an ecosystem of start-ups, it can be useful to find a way to measure the global success of the model. An idea is then to consider $D_n^{(k)}$, $k \geq 2$, again in the model with $a = \nu + 1 - \rho$ and $\nu > \rho$. It is expected that the larger $D_n^{(k)}$ is, the better PUT generates "successful" colors.

The aim is now to optimize $D_n^{(2)}$, since the crucial point for a color is to be drawn for a second time, because with the first one there is no replacement of further copies. Moreover the first draw can be interpreted as the first discovery of an innovation, while the second one determines its ability to receive attention again. It will also be analyzed the case for $D_n^{(3)}$, while there will be only a general consideration for the cases with $k > 3$.
Considering $D_n^{(2)}$, the objective is

$$\max_{\nu>\rho} D_n^{(2)} \sim \max_{\nu>\rho} \frac{1}{\nu+1}D_n \sim \max_{\nu>\rho} \frac{1}{\nu+1}\cdot\frac{\nu-\rho}{\nu-\rho+1}n, \qquad \nu,\rho \in \mathbb{N}.$$

Since n is just a measure of time, the real optimization problem is

$$\max_{\nu>\rho} \frac{\nu-\rho}{(\nu+1)(\nu-\rho+1)}, \qquad \nu,\rho \in \mathbb{N},$$

whose solution, without any further constraints, can be found empirically and it is $\nu = 1$, $\rho = 0$ that give an optimum of $\frac{1}{4}$. This is not a very interesting result, since the colors do not receive reinforcement and the *richer-gets-richer* mechanism is not reproduced. We decide then to introduce another constraint, in order give a fixed value for the sum $\nu + \rho$.

We would like to define some sort of budget C to use, determining a trade-off between reinforcement and triggering. The optimization problem becomes, given $C \in \mathbb{Z}^+$,

$$\max_{\substack{\nu>\rho \\ \nu+\rho=C}} \frac{\nu-\rho}{(\nu+1)(\nu-\rho+1)}, \qquad \nu,\rho \in \mathbb{N},$$

which becomes, substituting $\rho = C - \nu$,

$$\max_{\frac{C}{2}<\nu<C} \frac{2\nu-C}{(\nu+1)(2\nu-C+1)} = \max_{\frac{C}{2}<\nu<C} \mathrm{obj}_C^{(2)}(\nu), \qquad \nu \in \mathbb{Z}^+. \tag{4.15}$$

The continuous optimal point, as a function of C, is easy to find analytically and it corresponds to

$$\tilde{\nu}^{(2)}_{\max}(C) = \frac{1}{2}\left(C + \sqrt{C+2}\right); \tag{4.16}$$

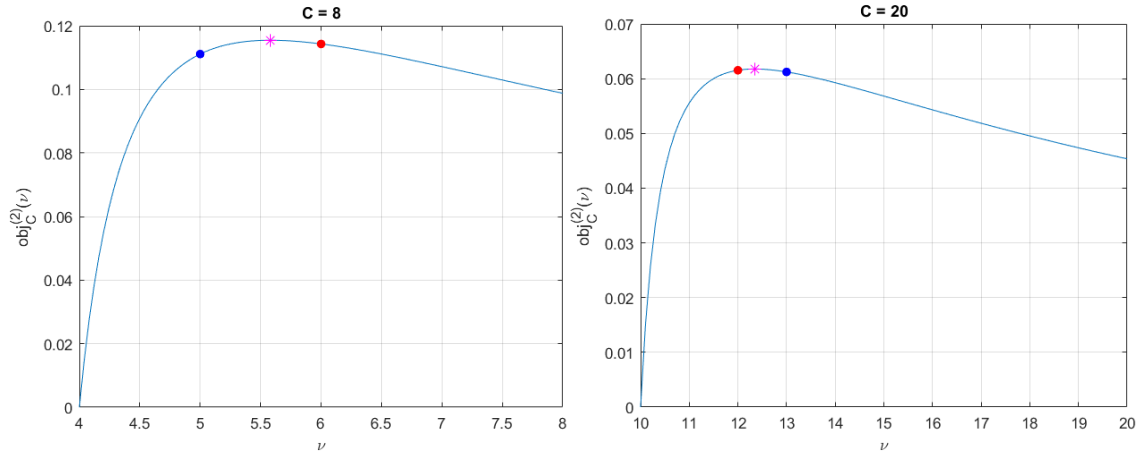
The integer solution is obtained by checking the values of floor and ceiling and choosing the one that gives the maximum value for the objective, as the following corollary states.

Corollary 1. *Consider the process of PUT's model with parameters $\nu, \rho \in \mathbb{N}$, when $\nu > \rho$ and $a = \nu + 1 - \rho$. Fixing $\nu + \rho = C \in \mathbb{Z}^+$, the statistic $(D^{(2)}_n)_{n\geq 0}$, which counts the number of distinct colors with at least two draws up to time n, is maximized by ν and ρ such that*

$$\nu^{(2)}_{max}(C) = \arg\max\{obj^{(2)}_C\left(\left\lfloor \tilde{\nu}^{(2)}_{max}(C)\right\rfloor\right), obj^{(2)}_C\left(\left\lceil \tilde{\nu}^{(2)}_{max}(C)\right\rceil\right)\}. \tag{4.17}$$

and $\rho^{(2)}_{max}(C) = C - \nu^{(2)}_{max}(C)$, where $\tilde{\nu}^{(2)}_{max}(C)$ is the one in (4.16) and $obj^{(2)}_C$ is defined in (4.15).

The plots of of $\text{obj}^{(2)}_C(\nu)$ for two different values of C are shown in figures 4.1a and 4.1b.



(a) $\nu^{(2)}_{\max}(8) = 6$.

(b) $\nu^{(2)}_{\max}(20) = 12$.

Figure 4.1: Plots of $\text{obj}^{(2)}_C(\nu)$ when $C = 8$ and $C = 20$: the purple stars represent respectively the global maxima with continuous ν: $\tilde{\nu}^{(2)}_{\max}(8) = 5.58$ and $\tilde{\nu}^{(2)}_{\max}(20) = 12.35$; the colored dots are the closest integer points and the red one is the solution to the integer optimization problems: $\nu^{(2)}_{\max}(8) = 6$ and $\nu^{(2)}_{\max}(20) = 12$.

The same approach can be used now for $D_n^{(3)}$ in order to calculate the maximum fraction, with respect to n, of elements with 3 or more draws. The optimization problem to solve is

$$\max_{\substack{\nu>\rho \\ \nu+\rho=C}} \frac{(\nu-\rho)(\rho+1)}{(\nu+1)(\nu-\rho+1)(\nu+\rho+1)}, \qquad \nu,\rho \in \mathbb{N};$$

and, with the substitution $\rho = C - \nu$, it becomes

$$\max_{\frac{C}{2}<\nu<C} \frac{(2\nu-C)(C-\nu+1)}{(\nu+1)(2\nu-C+1)(C+1)} = \max_{\frac{C}{2}<\nu<C} \mathrm{obj}_C^{(3)}(\nu), \qquad \nu \in \mathbb{Z}^+ \tag{4.18}$$

In this case the continuous optimal point is

$$\tilde{\nu}_{\max}^{(3)}(C) = \frac{(C+2)(2C+\sqrt{2(C+3)})-2}{4C+10} \tag{4.19}$$

The integer solution is obtained again by checking the values of floor and ceiling and choosing the one that gives the maximum value for the objective.

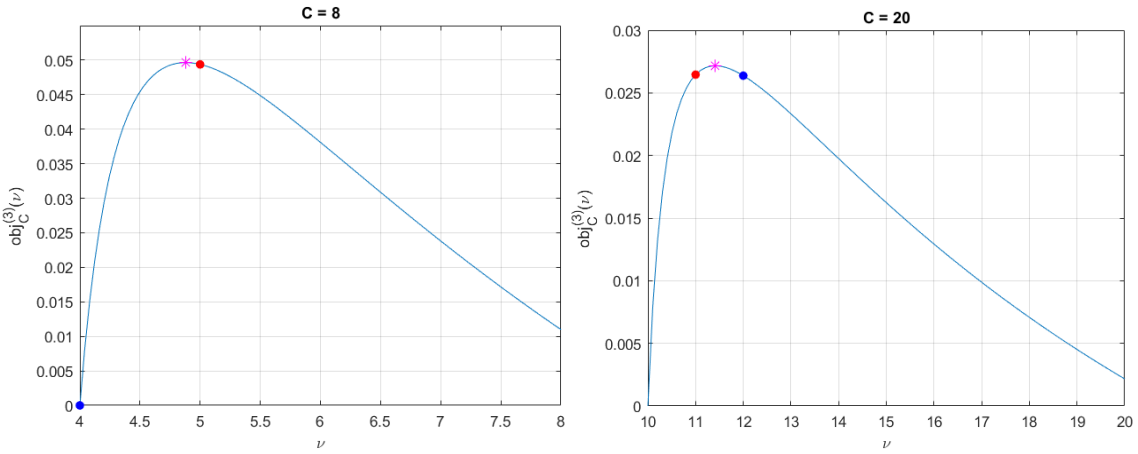
Corollary 2. *Consider the process of PUT's model with parameters $\nu, \rho \in \mathbb{N}$, when $\nu > \rho$ and $a = \nu+1-\rho$. Fixing $\nu+\rho = C \in \mathbb{Z}^+$, the statistic $(D_n^{(3)})_{n\geq 0}$, which counts the number of distinct colors with at least 3 draws up to time n, is maximized by ν and ρ such that*

$$\nu_{max}^{(3)}(C) = \arg\max\{obj_C^{(3)}\left(\left\lfloor \tilde{\nu}_{max}^{(3)}(C)\right\rfloor\right), obj_C^{(3)}\left(\left\lceil \tilde{\nu}_{max}^{(3)}(C)\right\rceil\right)\}. \tag{4.20}$$

and $\rho_{max}^{(3)}(C) = C - \nu_{max}^{(3)}(C)$, where $\tilde{\nu}_{max}^{(3)}(C)$ is the one in (4.19) *and $obj_C^{(3)}$ is defined in* (4.18).

The plots of $\mathrm{obj}_C^{(3)}(\nu)$ for the same values of C as figure 4.1 are shown in figures 4.2a and 4.2b.

The analysis could continue for $D_n^{(k)}$, $k \geq 4$, but it becomes less interesting. We already observed that the optimal value for ν decreases from the case $k = 2$ to the case $k = 3$. This pattern is due to the general term of the product in (4.13): in fact, from $k-1$ to k, the objective to maximize is multiplied by $\frac{(k-2)\rho+1}{(k-2)\rho+1+\nu}$. This value is penalized by large ν and it is increased by large ρ, bringing the the optimal values of the two parameters closer and closer to each other.



(a) $\nu^{(3)}_{\max}(8) = 5$.

(b) $\nu^{(3)}_{\max}(20) = 11$.

Figure 4.2: Plots of $\text{obj}^{(3)}_C(\nu)$ when $C = 8$ and $C = 20$: the purple stars represent respectively the global maxima with continuous ν: $\tilde{\nu}^{(3)}_{\max}(8) = 4.88$ and $\tilde{\nu}^{(3)}_{\max}(20) = 11.41$; the red and blue dots are the closest integer points and the red one is the solution to the integer optimization problems: $\nu^{(3)}_{\max}(8) = 5$ and $\nu^{(3)}_{\max}(20) = 11$. Moreover here we can observe that the shape of the objective functions is more concentrated with a peak on smaller values of ν, closer to ρ with respect to $\text{obj}^{(2)}_C(\nu)$.

Chapter 5

Numerical simulations of the PUT model

In this chapter, there are the results and plots from the numerical simulations of the PUT model, with the aim of observing that the analytical results of Chapter 4 are really observable once the model is applied. We need at first to point out that the results have asymptotic validity while the simulations have finite time. However, setting a number of draws large enough, it is possible to observe the expected generic patterns that do not depend on the single run.
The simulations were run on Matlab: the urn is represented by a vector whose index represents the color and the value the number of balls; the dimension of the vector increases every time new colors are added into the urn. The number of draws $n_{\max}$ was set from 10^4 to 10^6, depending on the time required from the specific model: generally when $\nu > \rho$ the computation is slower due to the increasing size of vectors that store informations about each color.

5.1 Heaps' law

First of all we show that the fraction of distinct elements with respect to the time, X_n, converges to 0 when $\nu \leq \rho$, while it goes to $\frac{\nu-\rho}{a}$ when $\nu > \rho$, as it was predicted from the analysis of ODEs. The plots are shown in figures 5.1, 5.2 and 5.3.
After that we would like to observe how D_n increases in time; the cases of main interest are when $\nu \neq \rho$, while the case $\nu = \rho$ is quite border-line and less interesting to plot. The following observations can be made:

- in the case $\nu < \rho$

$$D_n \sim cn^{\frac{\nu}{\rho}} \Longrightarrow \log_{10}(D_n) \sim \log_{10}(c) + \frac{\nu}{\rho}\log_{10}(n)$$

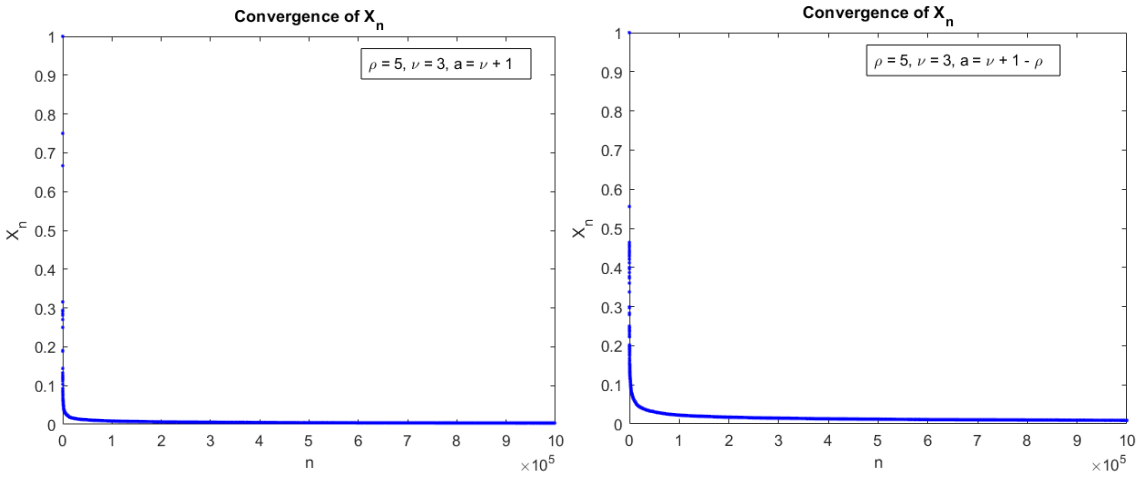


Figure 5.1: Evolution of $X_n = \frac{D_n}{n}$ after 10^6 draws, when $\nu < \rho$: as expected from the analysis of the ODE it converges to 0; we observe that on the left image when $a = \nu + 1$, and therefore the reinforcement starts from the first draw, the convergence is slightly faster than the case in which $a = \nu + 1 - \rho$.

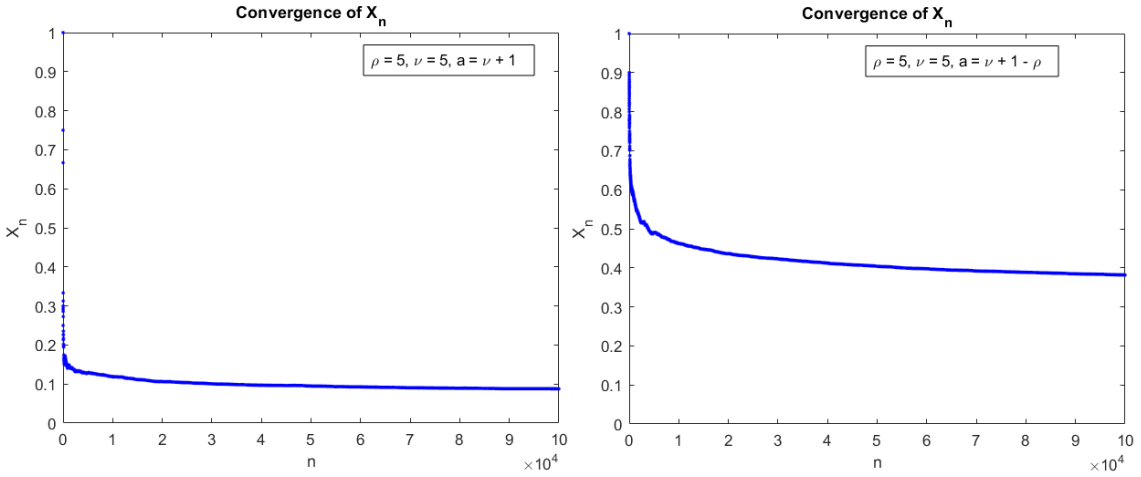


Figure 5.2: Evolution of $X_n = \frac{D_n}{n}$ after 10^5 draws, when $\nu = \rho$: as expected from the analysis of the ODE, it seem to go to 0, even though we don't observe it yet because the convergence is slow. Again, on the left image, when $a = \nu + 1$ and therefore the reinforcement starts from the first draw, the convergence is faster than the case in which $a = \nu + 1 - \rho$.

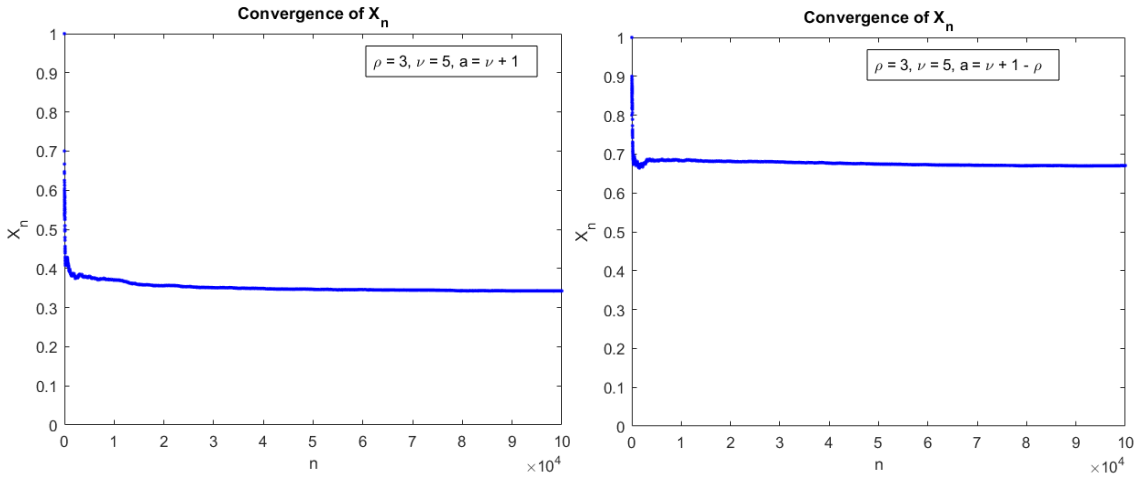


Figure 5.3: Evolution of $X_n = \frac{D_n}{n}$ after 10^5 draws, when $\nu > \rho$: as expected from analysis of the ODE it converges to $\frac{\nu-\rho}{a}$. On the left image, when $a = \nu + 1$, the convergence value is $\frac{\nu-\rho}{a} = \frac{1}{3}$, while on the right $a = \nu + 1 - \rho$ and the convergence value is $\frac{\nu-\rho}{a} = \frac{2}{3}$.

for some constant c;

- in the case $\nu < \rho$

$$D_n \sim \frac{\nu - \rho}{a} n \Longrightarrow \log_{10}(D_n) \sim \log_{10}\left(\frac{\nu - \rho}{a}\right) + \log_{10}(n).$$

If in the simulation, for each color j, the point $\Big(\log_{10}(N_j), \log_{10}(D_{N_j})\Big)$, $j = 1, \ldots, D_{n_{\max}}$, is recorded, we can use linear regression to estimate the coefficients. When $\nu > \rho$ the intercept should be about $\log_{10}\left(\frac{\nu-\rho}{a}\right)$ while the angular coefficient about 1; when $\nu < \rho$ the angular coefficient should be instead $\frac{\nu}{\rho}$. Since the analytical proof was for $n \to \infty$, the regression is done considering only the 15% of most recently appeared colors, in order to avoid initial uncertainty that could considerably change the results. Plots 5.4 and 5.5 show the the evolution of D_n in logarithmic scale for both the cases $\nu < \rho$ and $\nu > \rho$.

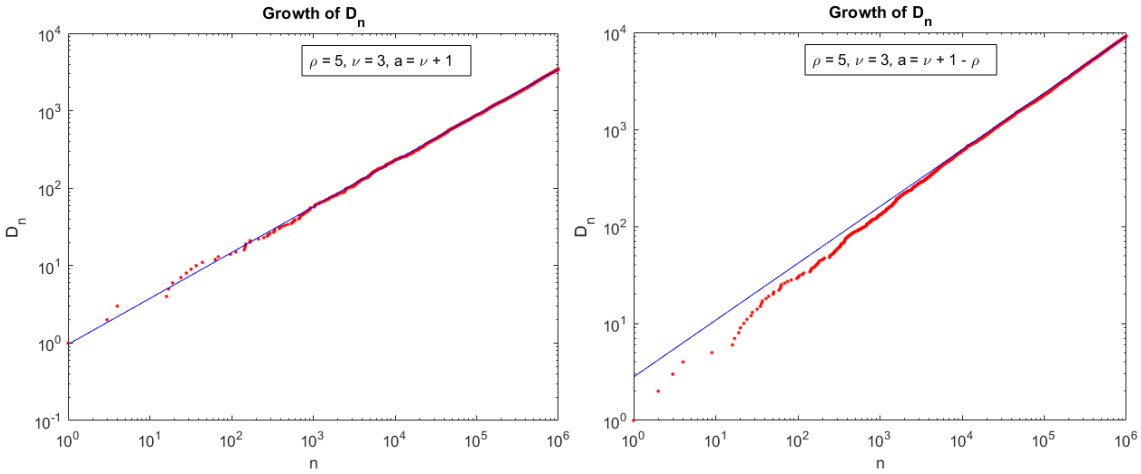


Figure 5.4: Evolution of D_n after 10^6 draws, when $\nu < \rho$: the scale of the plots is logarithmic, therefore they appear as lines. The blue line is fitted through linear regression on the latest 15% of colors. The predicted exponents are 0.5924 for the left plot and 0.5855 for the right one, which means that both of them are closer to the value $\frac{\nu}{\rho} = \frac{3}{5} = 0.6$. We observe a larger intercept on the right, when $a = \nu + 1 - \rho$, since without reinforcement at the first draw, more colors are drawn and D_n initially grows faster.

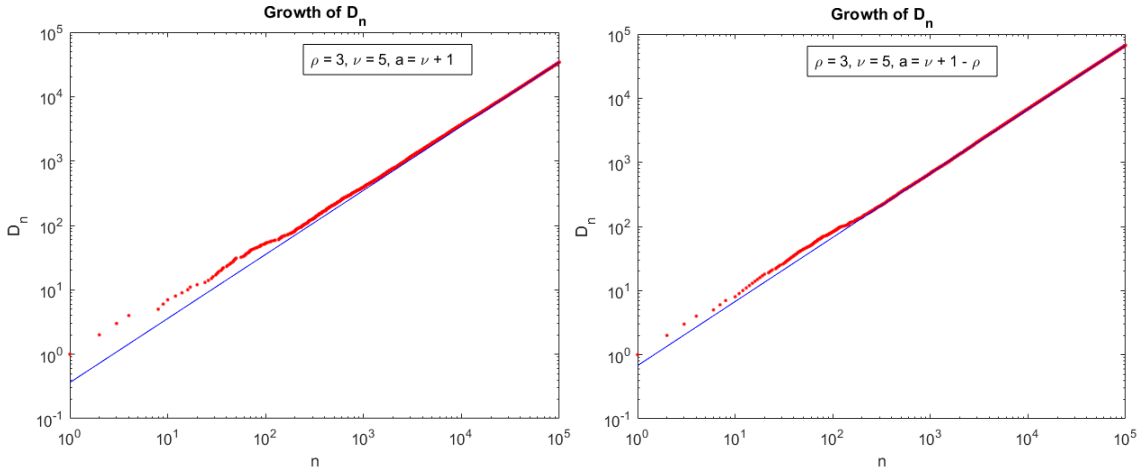


Figure 5.5: Evolution of D_n after 10^5 draws, when $\nu > \rho$: the scale of the plots is logarithmic and also here they are lines. The blue line is fitted through linear regression on the latest 15% of colors. The predicted exponents are 0.9950 for the left plot, and 0.9992 for the right one, both of them close to 1 as expected. In order to calculate the estimate of $\frac{\nu-\rho}{a}$ we calculate 10 to the power of the intercept: in the case of $a = \nu+1$ we have $10^{-0.4401} \approx 0.3630$ (expected $\frac{1}{3} \approx 0.33$), while for the case $a = \nu+1-\rho$ we have $10^{-0.1699} \approx 0.6762$ (expected $\frac{2}{3} \approx 0.67$).

5.2 Zipf's law

At first it is interesting to observe that the frequency of each color, after an initial uncertainty, goes to zero when $\nu > \rho$, while it stabilizes on a constant value when $\rho > \nu$. For this reason in figure 5.6 there are the plots of the five colors with most draws in both cases, for which we can observe the theoretic behavior.

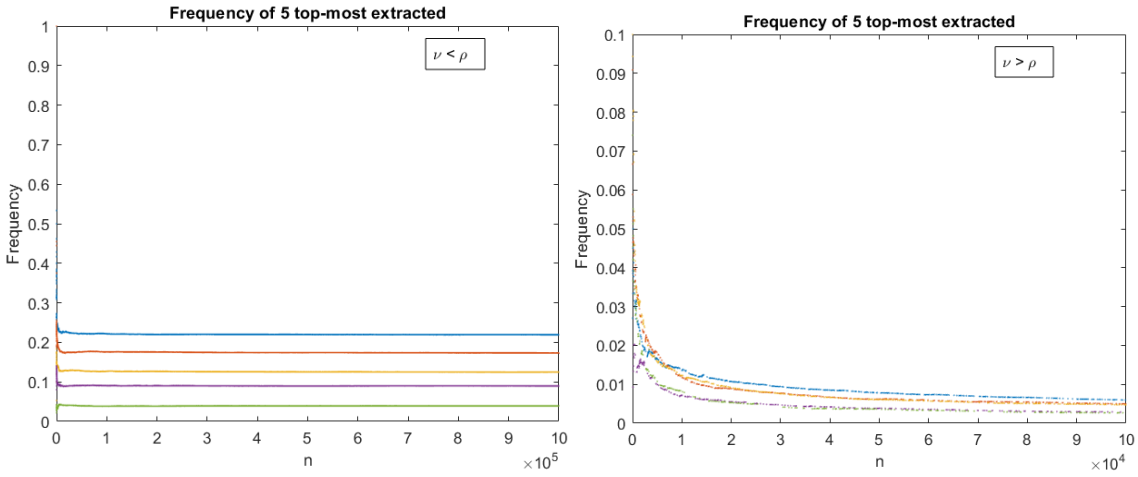


Figure 5.6: Evolution of the frequency of the five top-most drawn colors: when $\nu < \rho$ each color reaches a constant frequency, while in the case $\nu > \rho$ that fraction goes to zero for each color. Another interesting pattern is that on the right plot the lines are dotted: indeed the times between two consecutive draws of the same color are more interleaved, as expected.

The analysis in the previous section stated that the frequency of each color depends on the order it has been drawn for the first time and we would like to observe this pattern: however this is a stochastic model, therefore we can observe in figure 5.7 that the observed frequency has a great amount of noise with respect to the expected one.
This pattern does not contradict Zipf's law: if the frequencies are sorted and plotted against their rank we observe a power law in the tail of the distribution (figures 5.8 and 5.9). In order to estimate the exponent of the power law we use again a the linear regression on the logarithm of the frequencies: since the analysis described the tail of the distribution (j large), the first 100 elements in the rank have not been considered in the regression, and neither the elements with less than 4 draws. This last is a common practice due to the fact that the colors with few draws are too many and they could distort the result.
The plots in figure 5.9 represent the cases where $a = \nu + 1 - \rho$ and they have a singular pattern: there is a huge tail of elements drawn only once; this pattern was the main motivation of sections 4.3 and 4.4.

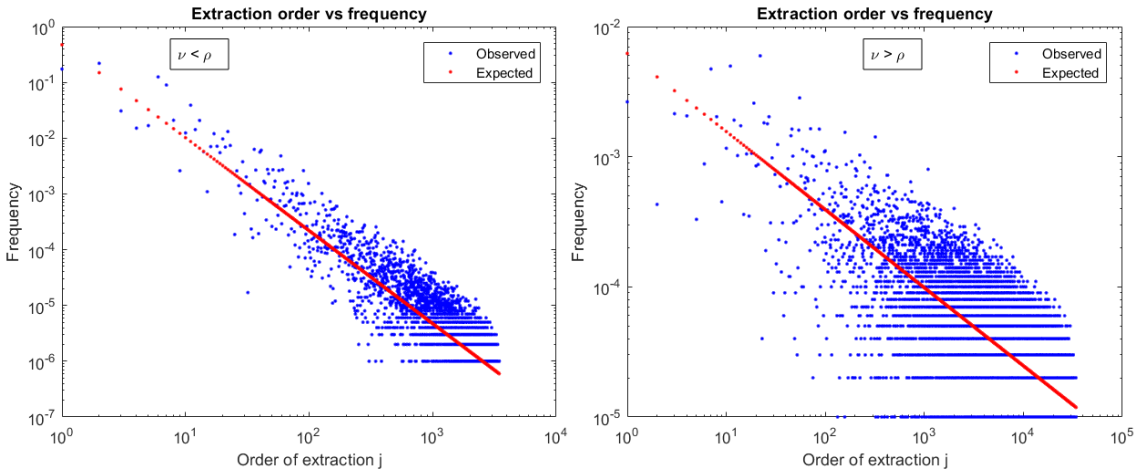


Figure 5.7: The blue dots represent the frequencies of each color plotted against their order of appearance, while the red ones are theoretical, proportional to $j^{-\frac{\rho}{\nu}}$. It is interesting to observe that when $\nu > \rho$ there is much more noise, since there are more colors and with very little frequency.

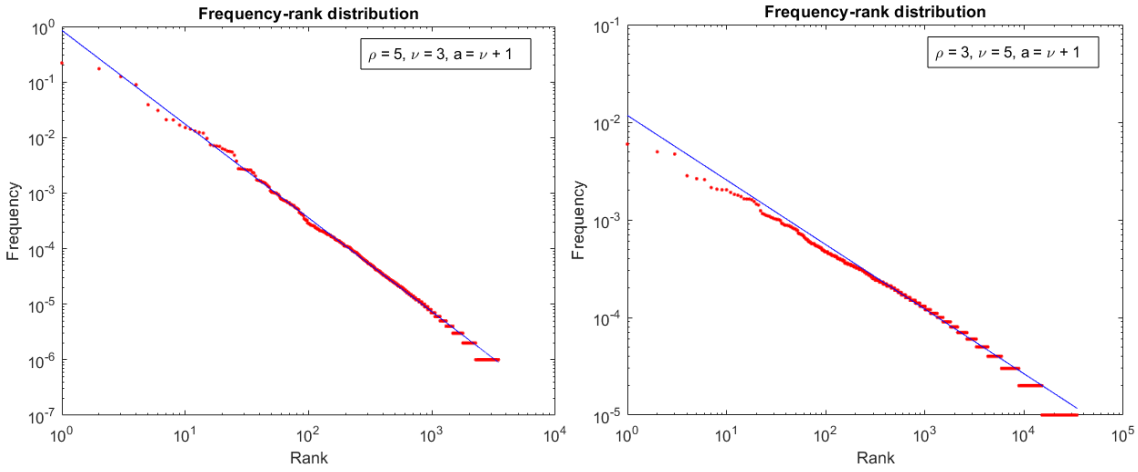


Figure 5.8: The red dots represent the frequencies of each color plotted against their rank, while the blue line is the fitting through the regression described. The exponent of the power law should be $-\frac{\rho}{\nu}$: on the left plot we have $-\frac{\rho}{\nu} = -\frac{5}{3} \approx -1.67$ and the estimate is equal to -1.6884; on the right one $-\frac{\rho}{\nu} = -\frac{3}{5} = -0.6$ and the estimate is equal to -0.6618.

5.3 Distinct colors with at least k draws

We now compare the formula obtained in expression (4.13) of Chapter 4 with numerical simulations. However we will see that, given $D_n^{(1)} = D_n$, the recursion (4.12) will be valid

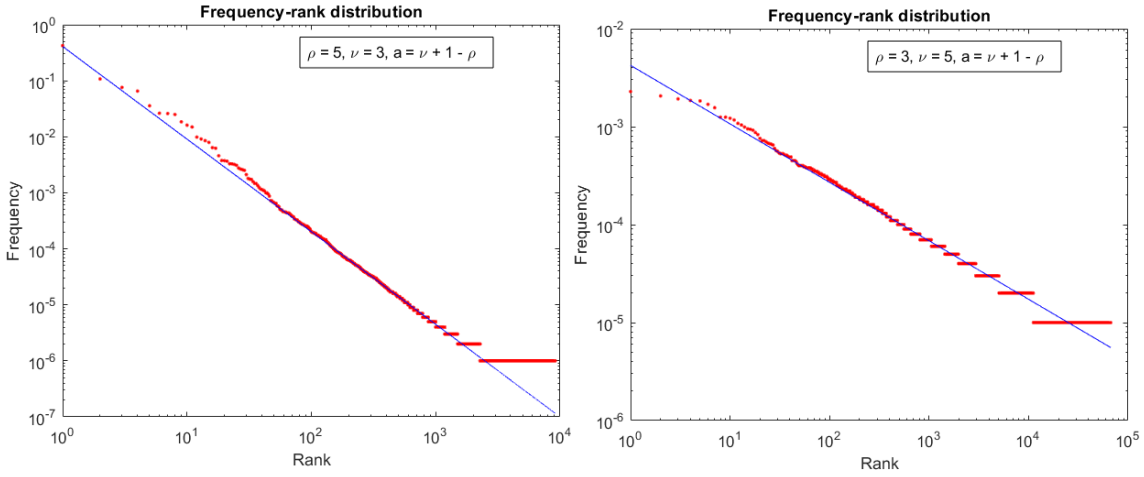


Figure 5.9: The red dots represent the frequencies of each color plotted against their rank, while the blue line is the fitting through the regression described. The exponent of the power law should be $-\frac{\rho}{\nu}$: on the left plot we have $-\frac{\rho}{\nu} = -\frac{5}{3} \approx -1.67$ and the estimate is equal to -1.6533; on the right one $-\frac{\rho}{\nu} = -\frac{3}{5} = -0.6$ and the estimate is equal to -0.5968. Since $a = \nu + 1 - \rho$ and there is no reinforcement at the first draw, many elements get to be drawn only once and it is clearly observable from the plots.

also for $\nu < \rho$. In order to verify the result, the usual simulation was run with $\rho = 3$ and $\nu = 5$ with 10^5 draws. The theoretic values of $D_n^{(k)}$, $k = 1, \dots, 10$, were compared with the observed one and the results are shown in table 5.1, confirming the analysis done.

k	1	2	3	4	5
$D_n^{(k)}$ observed	67035	11184	5054	2950	1978
$D_n^{(k)}$ predicted from (4.13)	66667	11111	4938	2881	1920
k	6	7	8	9	10
$D_n^{(k)}$ observed	1441	1056	809	654	559
$D_n^{(k)}$ predicted from (4.13)	1387	1058	837	682	568

Table 5.1: Comparison between expected $D_n^{(k)}$, $k \in \{1, \dots, 10\}$, and the observed values.

As announced before, the recursion (4.12) can be observed also when $\nu < \rho$ and given D_n it is possible to predict $D_n^{(k)}$ $\forall k \geq 2$. Running the usual simulation with $\rho = 5$ and $\nu = 3$ with 10^6 draws the results in the table 5.2 were observed.

k	1	2	3	4	5
$D_n^{(k)}$ observed	9148	2273	1503	1182	999
$D_n^{(k)}$ obtained from (4.12)	-	2287	1525	1198	1009
k	6	7	8	9	10
$D_n^{(k)}$ observed	872	784	714	664	621
$D_n^{(k)}$ obtained from (4.12)	883	791	722	666	621

Table 5.2: Given $D_n^{(1)}$, comparison between expected $D_n^{(k)}$, $k \in \{2, \dots, 10\}$, and the observed values.

5.4 Optimization

Now it is possible to show that the results of Chapter 4, in which the number of element with at least 2 or 3 draws were maximized, are mainly correct. In order to show that, different simulations were run, for all the possible values of ν and ρ such that $\nu > \rho > 0$ and $\nu + \rho = C$. The parameter C was set equal to 8 or 20, as in examples of figures 4.1 and 4.2. The number of total draws was set to 10^4, since large values of ν, for example 19, really slow down the computational time.

The results are shown in the tables 5.3 and 5.4: even though there is a different value as the one expected for $D_n^{(2)}$ in the case $C = 20$, they are coherent with analysis performed.

ν	*5*	*6*	*7*
ρ	*3*	*2*	*1*
$D_n^{(2)}$	1098	**1140**	1067
$D_n^{(3)}$	**476**	368	212

Table 5.3: Values of $D_n^{(2)}$ and $D_n^{(3)}$ for all possible combinations when $C = 8$; in bold the maximum values for each of them: they are as expected.

ν	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
ρ	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*
$D_n^{(2)}$	590	604	**613**	592	542	541	519	470	469
$D_n^{(3)}$	**288**	262	241	190	156	122	108	68	49

Table 5.4: Values of $D_n^{(2)}$ and $D_n^{(3)}$ for all possible combinations when $C = 20$; in bold the maximum values for each of them. We observe that the maximum value for $D_n^{(2)}$ is obtained when $\nu = 13$ instead of $\nu = 12$, the one predicted, but we have seen in figure 4.1b that the difference is very small between the two values for calculating the objective. In order to observe the expected behavior it may be necessary to run the simulations for more draws.

Chapter 6

Extensions of the model

In this chapter, the aim is to introduce some changes to the PUT model studied until now, in order to reproduce some different phenomena of the real world. In the first section the reinforcement of a color is no more constant from one color to another, but each element i has its own ρ_i that represents the number of balls to reintroduce. A different parameter for each color should represent its capacity to attract further attention, making it more likely to be drawn with respect to colors appeared before, but with smaller ρ_i.
The other section instead focuses on the urn, imagining a limited life with geometric distribution for each ball inside it: the analysis focuses on what happens inside the urn and what are the consequences for the draws and for the already analyzed statistics.

6.1 Reinforcements depending on colors

In the original model every color receives the same reinforcement once it is drawn, without considering a measure of quality for the color. The idea behind the model we are going to propose is that each innovation should receive a reinforcement based on how good it is evaluated, and its success should also depend on the quality and not only on how much early it is discovered. Monechi et al. [6] proposed a model that made possible what they call *waves of novelties*, in which even younger innovations could become popular. Their approach is a more sophisticated extension of the PUT model, taking into account correlations between semantically related innovations and modeling a multiple agent instead of an average one with the introduction of weights on balls based on some functions.
Our approach goes in another direction: it proposes the elimination of the constant reinforcement parameter ρ and it introduces a different parameter λ, the mean of a generic distribution from which reinforcement parameters ρ_i, $i \in \{1,2,\dots\}$, are sampled. The proposed model works in the following way:

- there are N_0 initial balls in the urn, each one with a different color and every color i has its own ρ_i, sampled from a random distribution with mean λ;
- at each time n, $n \geq 1$, a ball is drawn from the urn and, based on color i, ρ_i copies of the ball are inserted together with the drawn ball;
- if it is the first draw for a ball, $\nu + 1$ balls of new colors are placed into the urn and, concurrently, each new ball's ρ_i is determined by random sampling from the given distribution.

The goal now is to study again the behavior of $(D_n)_{n\geq 0}$ and $(K_n^j)_{n\geq 0}$. Specifically, we would like to determine the behavior of the former, while, for the latter, show that the frequency does not depend only on the order of draw j, but also on the specific reinforcement parameter ρ_j. The model is more interesting to study through by simulations, anyway it was tried to explain it analytically, as done in Chapter 2, but with some heuristic arguments.
At first, the total number of balls in the urn is different from PUT:

$$|U|_n = N_0 + \sum_{j=1}^{D_n} \rho_j K_n^j + (\nu + 1)D_n; \tag{6.1}$$

in fact, instead of considering ρ balls introduced at each time n, we should consider the number of time each color has been drawn and for each of them its own ρ_j. At first it is interesting to observe:

$$\begin{aligned}\frac{|U|_n}{n} &= \frac{N_0}{n} + \sum_{j=1}^{D_n} \rho_j \frac{K_n^j}{n} + (\nu + 1)\frac{D_n}{n} \sim \sum_{j=1}^{D_n} \rho_j Y_n^j + (\nu + 1)X_n \\ &\sim \rho_{\text{avg}} + (\nu + 1)X_n, \qquad \text{as } n \to \infty,\end{aligned}$$

where ρ_{avg} is the weighted sum of the ρ_j, based on the fraction of times their color has been drawn, since $\sum_{j=1}^{D_n} Y_n^j = 1$. This approach is useful to approximate the usual probabilities that D_n or K_n^j increase by one, which would depend both on all K_n^j, $j \in \{1, \ldots, D_n\}$, since at the denominator there would be the expression in (6.1). The use of ρ_{avg} is clearly a further approximation, since it is considered as a constant while it is not, but we will make some assumptions on the value it takes, based on the distribution and on ν. Therefore the approximated probabilities are

$$P(D_{n+1} = D_n + 1 | X_n) \sim \frac{\nu X_n}{\rho_{\text{avg}} + (\nu + 1)X_n};$$

$$P(K_{n+1}^j = K_n^j + 1 | X_n, Y_n^j) \sim \frac{\rho_j Y_n^j}{\rho_{\text{avg}} + (\nu + 1)X_n}.$$

Using the stochastic approximation in the same way as in Chapter 2 the ODE system to study is

$$\begin{cases} \dot{x} = \dfrac{\nu x}{\rho_{\text{avg}} + (\nu+1)x} - x = \dfrac{x(\nu - \rho_{\text{avg}} - (\nu+1)x)}{\rho_{\text{avg}} + (\nu+1)x}; \\ \dot{y} = \dfrac{\rho_j y}{\rho_{\text{avg}} + (\nu+1)x} - y = \dfrac{y(\rho_j - \rho_{\text{avg}} - (\nu+1)x)}{\rho_{\text{avg}} + (\nu+1)x}. \end{cases} \tag{6.2}$$

The study of the first equation is the same already done in Chapter 2, except that ρ is substituted with ρ_{avg}. When $n \to \infty$ it follows:

- in the case $\rho_{\text{avg}} > \nu$
$$D_n \asymp n^{\frac{\nu}{\rho_{\text{avg}}}};$$
- in the case $\rho_{\text{avg}} < \nu$
$$D_n \sim \frac{\nu - \rho_{\text{avg}}}{\nu + 1} n.$$

Regarding instead the number of occurrences of each color, the result is presented without any proof since it is not exact, but in general it should hold

$$K_n^j \sim \left(\frac{j}{D_n}\right)^{-\frac{\rho_j}{\nu}}. \tag{6.3}$$

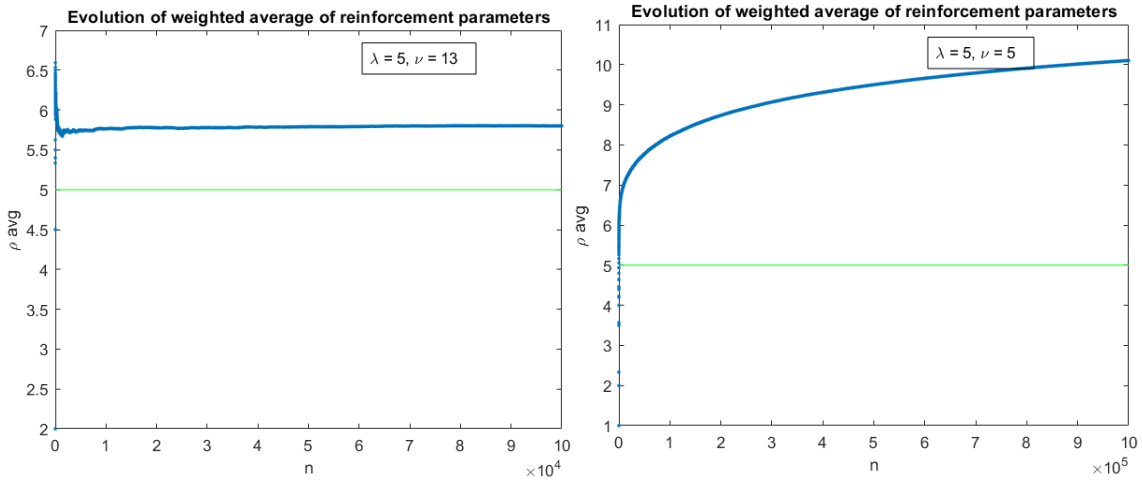
It follows that the occurrences of a color are again penalized by larger j, but now they are also increased by larger values of ρ_j.
It is necessary to make some assumptions about the value ρ_{avg}: it is the weighted average of the reinforcement parameters of the drawn colors and it is not a constant set from the beginning, but we would like to consider it as if it were, when n is large. We would try to understand which is its convergence value: since we know that each ρ_j, $j \geq 1$, is a non-negative integer random variable with mean λ, we expect $\rho_{\text{avg}} \geq \lambda$, assuming that the colors with larger ρ_j are more likely to be drawn and therefore their reinforcement parameter will have heavier weight in the average. For these reasons the result (6.3) is reported without a clear proof: the procedure is similar to the one in Chapter 2 (studying equilibria and approximating the ODE solution asymptotically), but it is based on some assumptions about ρ_j, $j \geq 1$, that are not always true. In fact the assumptions made in order to have stable equilibria in (6.2) are:

- if $\rho_{\text{avg}} < \nu$ it should hold $\rho_j < \nu \ \forall j \geq 1$, which may be not true if λ is not much smaller than ν;
- if $\rho_{\text{avg}} > \nu$ it should hold $\rho_j < \rho_{\text{avg}} \ \forall j \geq 1$, which may be true for most of the colors, but of course, being ρ_{avg} a weighted average, there should be at least one color j with $\rho_j \geq \rho_{\text{avg}}$.

Due to this uncertainty in the analytical study, it becomes necessary to study the simulations of this model. Three different distributions for the sampling of the reinforcement parameters have been proposed: Poisson, discrete uniform and rounded log-normal, since non-negative integers are needed. We would like to observe again how the number of distinct elements increases in time, check if the frequency-rank distribution is still la power law, and if it is possible for late innovations to emerge and reach high positions in the rank.

6.1.1 Poisson reinforcements

In this case ρ_j, $1 \leq j \leq D_n$, are i.i.d. random variables with Poisson distribution with parameter λ, corresponding to mean and variance. First of all we show the evolution of ρ_{avg} at each draw. We observe that, in the case λ is smaller enough than ν, ρ_{avg} converges to something little larger than λ, while in the case $\lambda \geq \nu$ or little smaller, ρ_{avg} increases to about the average value of the most drawn elements, which have quite larger values than λ (figures 6.1a and 6.1b).



(a) Final $\rho_{\text{avg}} = 5.8010$

(b) Final $\rho_{\text{avg}} = 10.1039$, but we see that it is still increasing.

Figure 6.1: The plots show how the weighted average of ρ_j, $1 \leq j \leq D_n$, evolves through time: when ν is too large ρ_{avg} converges to a value little larger than λ, here represented with the green line, while when ν is closer to λ, ρ_{avg} is biased towards the ρ_j of first positions, generally larger than ν.

Now we can also observe the evolution of D_n in the two cases and, as predicted, when $\nu > \rho_{\text{avg}}$ we have linear growth, on the other hand when $\nu < \rho_{\text{avg}}$ the growth is sublinear with coefficient $\frac{\nu}{\rho_{\text{avg}}}$. The estimate of the parameters is again obtained with linear regression on the logarithms, on the latest 15% of colors (figure 6.2).

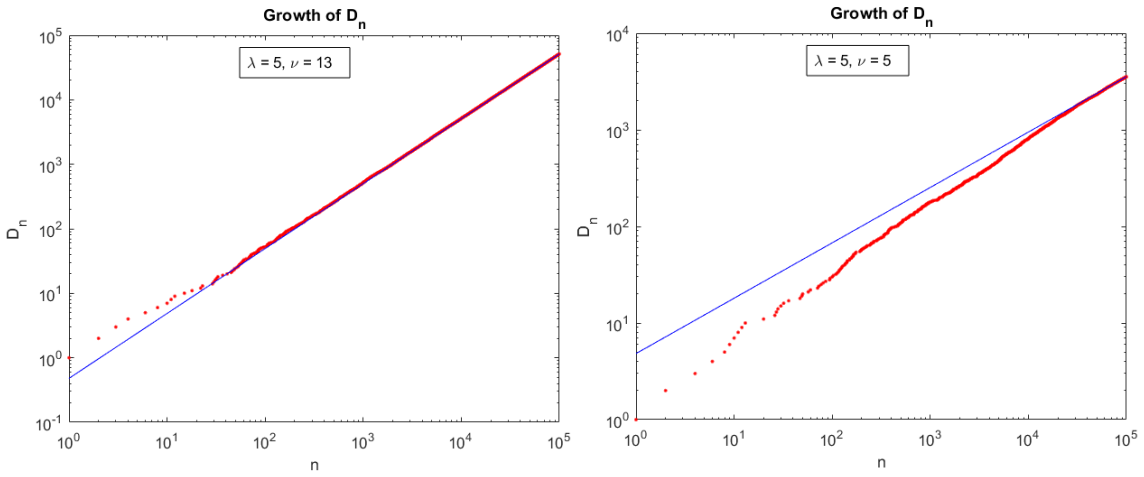
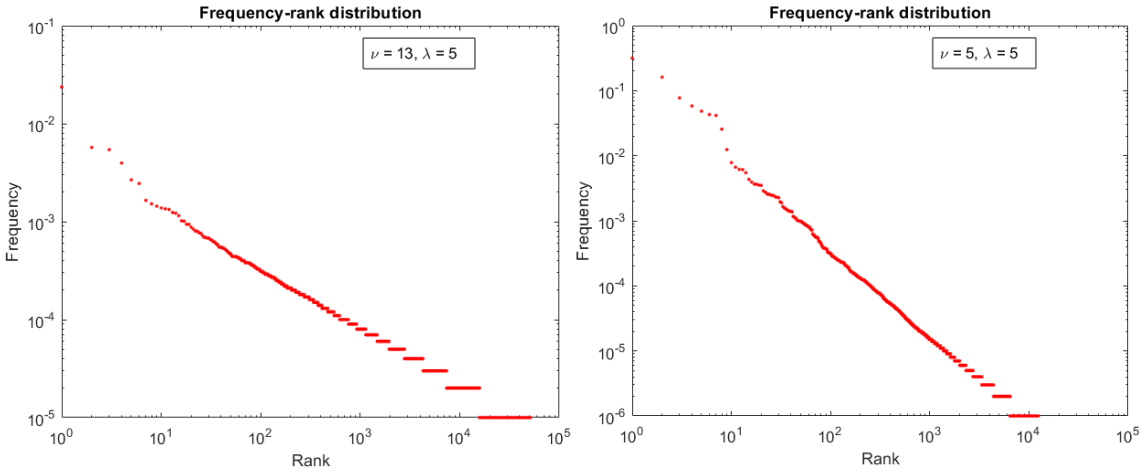


Figure 6.2: On the left the growth should be linear and in fact the estimated exponent is 1.0058; for more precision the coefficient is just calculated with the final fraction $\frac{D_n}{n} = 0.5131$ and it is close to the expected $\frac{\nu-\rho_{\text{avg}}}{\nu+1} = 0.5142$. On the right we see that the fitting blue line fits only the last part, since ρ_{avg} evolves through time and consequently the exponent $\frac{\nu}{\rho_{\text{avg}}}$ decreases. At the end $\frac{\nu}{\rho_{\text{avg}}} = 0.4949$, while the one predicted from the regression is 0.4953.

Regarding the frequency-rank distribution we only observe that it is again a power law, i.e. the frequency of an element in position i is proportional to $i^{-\alpha}$, $\alpha > 0$. It was also estimated α and observed that $\alpha \geq \frac{\lambda}{\nu}$ (figure 6.3).



(a) Estimated Zipf's coefficient: -0.5657

(b) Estimated Zipf's coefficient: -1.2741

Figure 6.3: Frequency-rank distributions.

Besides these observations, the crucial point of this model is to observe that even late colors are able to be in the first positions in the rank, if they have large reinforcement parameters. In table 6.1 we observe informations about the first five colors, with relative frequency, order and reinforcement parameter considering the case $\nu = \lambda = 5$. Moreover we see how they evolved in time in figure 6.4.

It interesting to observe how for early colors with smaller reinforcement parameter the frequency is decreasing, while later ones but with larger parameter are increasing their frequency: the most significant is the first one, appeared only as 88-th distinct element but with very large $\rho_j = 14$ that allowed to reach the top of the rank. In order to highlight the fact that in the first positions there are the colors with larger reinforcement there is also the plot in figure 6.5: each point represents the non-weighted average of the reinforcement parameter from the first up to that position in the rank; it can be observed that at the beginning the average is around ρ_{avg} and it decreases until reaching the mean λ of the distribution, once all the distinct drawn elements are considered.

Rank	Frequency	Order j	ρ_j
1	0.3158	88	14
2	0.1612	38	10
3	0.0773	22	9
4	0.0581	8	7
5	0.0483	5	7

Table 6.1: Informations about first 5 colors in the rank.

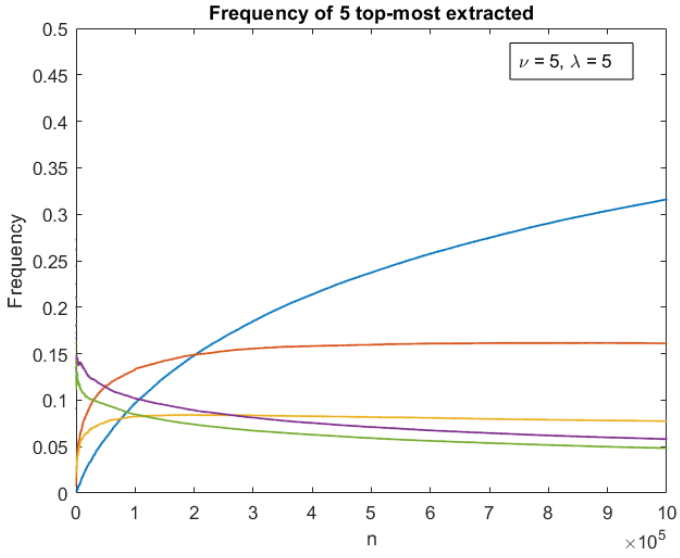


Figure 6.4: Evolution of the frequency of the first 5 colors through time.

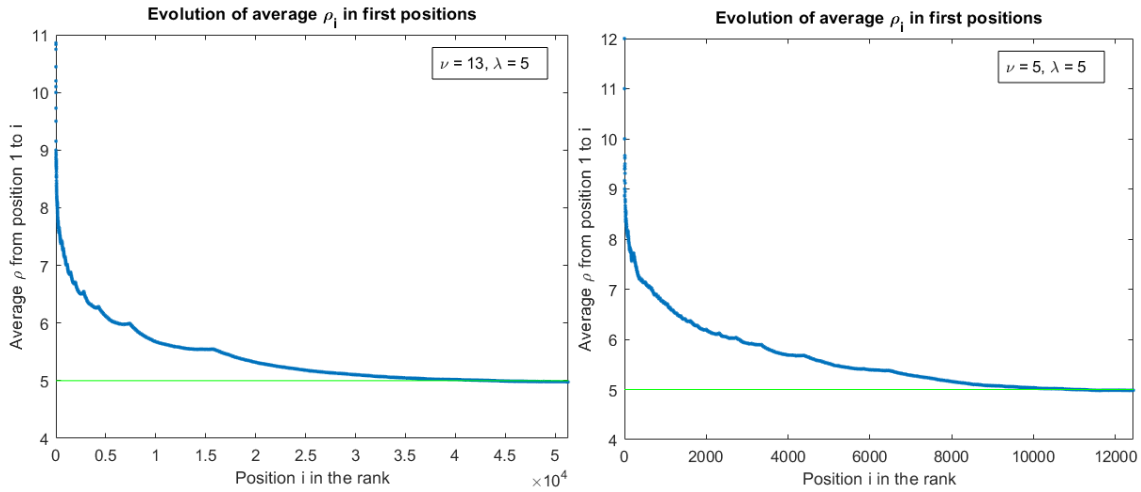


Figure 6.5: In the plots each point represents the non-weighted average of the reinforcement parameters from the first in the rank up to the position indicated on the x-axis; the green line is the value of λ.

The last analysis is intended to check if also very late elements are able to reach a significant frequency of draws. In order to observe that it was plotted the frequency versus the order of draw of each color (figure 6.6): we would like to see a uniform pattern instead of the decreasing one observed in figure 5.7. Actually, more than observing the success of very late elements, we observe more failures for the initial ones with small reinforcement.

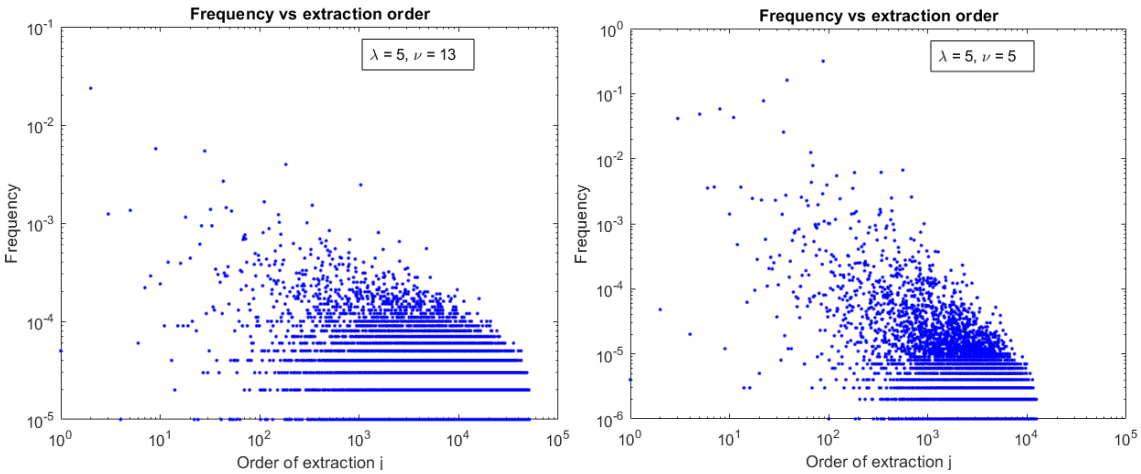


Figure 6.6: Poisson reinforcements frequency-order plots: each point represents a color, plotting the frequency versus the order of draw. There is still a decreasing pattern, even though it is sparser than the PUT.

6.1.2 Discrete uniform reinforcements

In this case ρ_j, $1 \leq j \leq D_n$, are i.i.d. random variables with discrete uniform distribution among values from 0 to 2λ, therefore with mean λ and variance $\frac{(2\lambda+1)^2-1}{12}$. Most of the observation already done for the Poisson distribution are quite the same, the difference here is that there is a maximum value for the reinforcement parameters and the distribution is not concentrated around its mean: therefore larger values are less rare and the variance is a bit larger, even though, setting for example $\lambda = 5$, they cannot be more than 10 while before we observed a value of 14 being among the first positions. As a consequence in the principal positions most of the colors have parameter equal to 10. This is a limit of this model, since it does not give the possibility to exist to rare but very successful innovations. In figure 6.7 there are the plots of the frequencies of each color against its order of appearance, trying to observe if they have a more uniform pattern than the Poisson case.

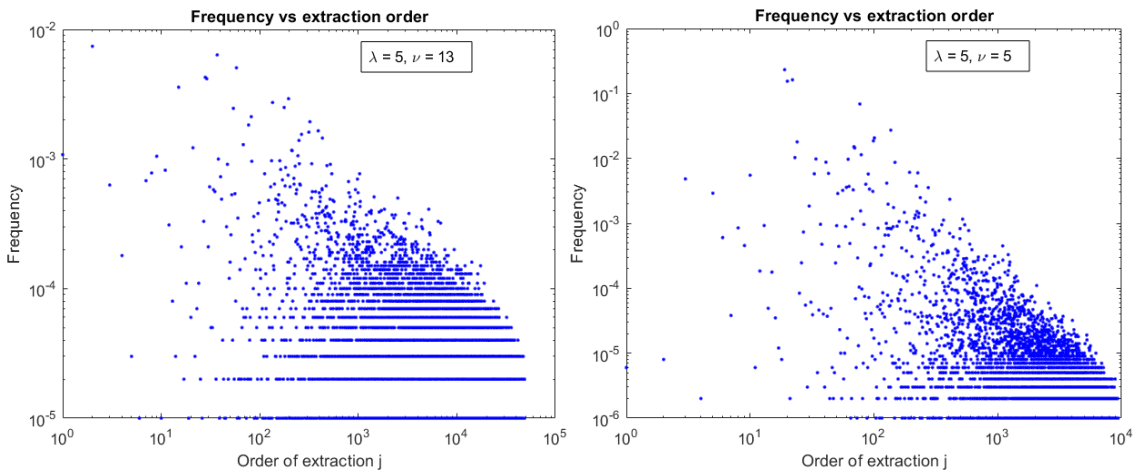


Figure 6.7: Discrete uniform reinforcements frequency-order plots: each point represents a color, plotting the frequency versus the order of draw. There is still a decreasing pattern, even though also here is sparser than the PUT and quite similar to Poisson case.

6.1.3 Rounded lognormal reinforcements

In both the previous distributions the variance was uniquely determined by the mean of the distribution, while we would like to have it larger in order to obtain very far values from the mean that could have a breakthrough among the others. Therefore now each ρ_j, $1 \leq j \leq D_n$, is obtained sampling from log-normal distribution, with parameters $\mu = 1.1094$ and $\sigma = 1$ such that the mean of the distribution is $\lambda = 5$ and the variance is sufficiently large (42.9570), and rounding to the closest integer.

We observe here that, even though we set $\nu = 13$ and $\lambda = 5$, due to the high variance introduced, at the end we have $\rho_{\text{avg}} = 46.0816$ and then a sub-linear growth for D_n. In figure 6.8 there is the plot of the frequencies versus the order of appearance and in this case we observe a more uniform pattern. In the first positions it is remarkable to observe elements drawn for the first time after the 1000th distinct color (table 6.2).

Rank	Frequency	Order j	ρ_j
1	0.5663	93	48
2	0.2529	363	53
3	0.0674	50	25
4	0.0327	145	36
5	0.0200	10	18
6	0.0067	1571	71
7	0.0064	2897	85
8	0.0050	1135	51
9	0.0033	55	21
10	0.0023	479	34

Table 6.2: Informations about first 10 colors in the rank.

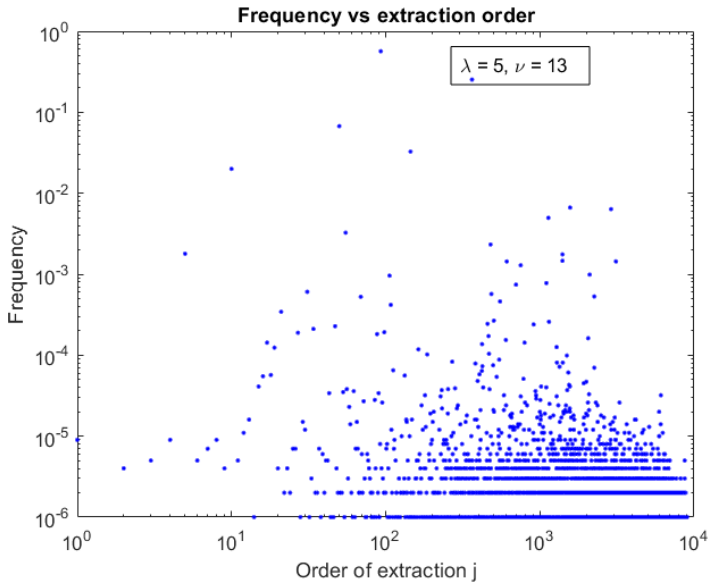


Figure 6.8: Rounded log-normal reinforcements frequency-order plot: each point represents a color, showing its frequency versus the order of draw. The decreasing pattern here is almost absent, allowing elements with late appearance to have many draws.

6.2 Elimination of balls from the urn

In the previous versions of the model the urn grows indefinitely, because ν or ρ new balls are always added at each draw. If we want to represent the urn as a system with limited capacity and where every ball has a limited life, we should think about something different. Moreover the probability of drawing any color in the urn is always positive, even though close to zero for the less represented ones. In the real world, when an innovation does not get enough attention, it generally disappears from the world, so the possibility of drawing again that "color" should be zero.

The idea developed for this model is to introduce the probability, for each ball, to disappear from the urn between two consecutive draws: this probability could be set constant for every ball at any time and quite small, which means that each ball has a lifetime with geometric distribution with parameter p.

6.2.1 Inside the urn

At first it is interesting to study the behavior of the number of balls in the urn and, at this purpose, the number of never drawn colors in the urn. Before studying analytically the problem, some simulations have been run and two different behaviors have been observed. A reasonable choice of p is to set it always smaller than 10^{-2}, in our cases we generally used 10^{-3} and the simulations brought the following observations.

- When $\rho > \nu$ the number of balls in the urn fluctuates around $p^{-1}\rho$, while the number of never drawn colors goes to zero. This means that it is impossible for D_n to increase, leading to a situation where there are just few colors in the urn and no other novelties can arise.

- When $\rho < \nu$ the number of balls in the urn fluctuates around $p^{-1}\nu$, while a quite fixed portion of them is represented by never drawn colors. In this case D_n can always increase.

Before giving more attention to numerical simulations, we tried to get some analytical results. Let us define two statistics: the stochastic processes $(U_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$ that describe respectively the number of balls in the urn and the number of balls in the urn of never drawn colors; clearly $V_n \leq U_n \ \forall n \geq 0$ and $U_0 = V_0 = N_0$. In detail the whole process consists of two steps at each time n:

1. The first part is the usual PUT: draw a ball from the urn and check the color, if it is the first draw replace it in the urn with $\nu + 1$ new colors; moreover always place in the urn ρ further copies of that color; there is again the slightly different model in which the copies are inserted from the second draw.

2. The second part consists of eliminating each ball in the urn with probability p. Given a certain number u of balls in the urn before this procedure, the number of balls surviving after this operation is a discrete random variable with binomial distribution with parameters u and $1 - p$.

Consider now that at each time n we apply together the two steps just described and define $n + \frac{1}{2}$ as the time between the two steps: it means that at time $n - \frac{1}{2}$ we apply step 1, at time n we apply step 2, at time $n + \frac{1}{2}$ we apply step 1 and so on.
Considering the processes $(U_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$, we can describe their conditional probability distribution when first step is applied. In the model in which the ρ copies are inserted from the first draw we have:

$$\begin{cases} P(U_{n+\frac{1}{2}} = U_n + \rho + \nu + 1 | U_n, V_n) = \dfrac{V_n}{U_n}; \\ P(U_{n+\frac{1}{2}} = U_n + \rho | U_n, V_n) = 1 - \dfrac{V_n}{U_n}. \end{cases}$$

In fact, after the first step, the number of balls in the urn always increases by ρ balls, and by further $\nu + 1$ balls if a new color is drawn. On the other hand in the case when no reinforcement is applied at the first draw the probabilities are

$$\begin{cases} P(U_{n+\frac{1}{2}} = U_n + \nu + 1 | U_n, V_n) = \dfrac{V_n}{U_n}; \\ P(U_{n+\frac{1}{2}} = U_n + \rho | U_n, V_n) = 1 - \dfrac{V_n}{U_n}. \end{cases}$$

For $(V_n)_{n\geq 0}$ instead there is only one case:

$$\begin{cases} P(V_{n+\frac{1}{2}} = V_n + \nu | U_n, V_n) = \dfrac{V_n}{U_n}; \\ P(V_{n+\frac{1}{2}} = V_n | U_n, V_n) = 1 - \dfrac{V_n}{U_n}. \end{cases}$$

Here the number of never drawn colors can increase or remain the same: if one of them is drawn, that number decreases by one but $\nu + 1$ new colors are added, bringing the net increase to ν; if an old color is drawn V_n stays still. These results lead to two different conditional expectations for $U_{n+\frac{1}{2}}$:

$$E\left[U_{n+\frac{1}{2}} | \mathcal{F}_n\right] = U_n + (\rho + \nu + 1)\frac{V_n}{U_n} + \rho\left(1 - \frac{V_n}{U_n}\right) = U_n + (\nu + 1)\frac{V_n}{U_n} + \rho$$

when the reinforcement is from the first draw, while

$$E\left[U_{n+\frac{1}{2}} | \mathcal{F}_n\right] = U_n + (\nu + 1)\frac{V_n}{U_n} + \rho\left(1 - \frac{V_n}{U_n}\right) = U_n + (\nu + 1 - \rho)\frac{V_n}{U_n} + \rho$$

when the reinforcement is from the second one. Summarizing and including the conditional expectation for $V_{n+\frac{1}{2}}$ it follows

$$\begin{cases} E\left[U_{n+\frac{1}{2}} | \mathcal{F}_n\right] = U_n + a\dfrac{V_n}{U_n} + \rho; \\ E\left[V_{n+\frac{1}{2}} | \mathcal{F}_n\right] = V_n + \nu\dfrac{V_n}{U_n}, \end{cases} \tag{6.4}$$

where $a = \nu + 1$ or $a = \nu + 1 - \rho$ depending on the model, as in the PUT, and $\mathcal{F}_n$ is the σ-algebra generated by the events of the process until time n.
In the second step each ball in the urn disappears with constant small probability p; therefore, conditional to the σ-algebra $\mathcal{F}_{n+\frac{1}{2}}$ generated by the events of the process until time $n + \frac{1}{2}$, U_{n+1} (V_{n+1}) is a binomial random variable with parameters $U_{n+\frac{1}{2}}$ ($V_{n+\frac{1}{2}}$) and $1 - p$. The conditional expectations are:

$$\begin{cases} E\left[U_{n+1} | \mathcal{F}_{n+\frac{1}{2}}\right] = U_{n+\frac{1}{2}}(1-p) \\ \\ E\left[V_{n+1} | \mathcal{F}_{n+\frac{1}{2}}\right] = V_{n+\frac{1}{2}}(1-p). \end{cases} \tag{6.5}$$

The idea now is to unify the two steps of the process as a whole one, considering only integer times. Putting together results from (6.4) and (6.5) we can summarize:

$$\begin{cases} E\left[U_{n+1} | \mathcal{F}_n\right] = \left[U_n + a\dfrac{V_n}{U_n} + \rho\right](1-p); \\ \\ E\left[V_{n+1} | \mathcal{F}_n\right] = \left[V_n + \nu\dfrac{V_n}{U_n}\right](1-p). \end{cases} \tag{6.6}$$

The use of stochastic approximation for this case results quite complicated, since the fact that we expect the quantities U_n and V_n to fluctuate around a constant value makes difficult to use the same procedure as done before for D_n and K_n^j and divide them by n, which will bring everything to zero. The approach now is heuristic, with the idea of using the conditional expectations of the difference between two subsequent values of the process as a discrete derivative. In this way we will define two functions that approximate the processes $(U_n)_{n\geq 0}$ and $(V_n)_{n\geq 0}$ and obtain an ODE system again. By studying the asymptotically stable equilibria of that ODE it should be possible to find the values around which the processes fluctuate or converge.

Rewriting the conditional expectations in (6.6) we obtain

$$\begin{cases} E\left[U_{n+1} - U_n | \mathcal{F}_n\right] = -pU_n + \left(a\dfrac{V_n}{U_n} + \rho\right)(1-p); \\ \\ E\left[V_{n+1} - V_n | \mathcal{F}_n\right] = -pV_n + \nu\dfrac{V_n}{U_n}(1-p). \end{cases} \tag{6.7}$$

The left term of each expression can be considered as a derivative with respect to the time and, defining the functions $u(t)$ and $v(t)$, approximations of U_n and V_n, the (6.7) is then approximated by the following ODE system:

$$\begin{cases} \dot{u} = -pu + \left(a\dfrac{v}{u} + \rho\right)(1-p); \\ \\ \dot{v} = -pv + \nu\dfrac{v}{u}(1-p). \end{cases}$$

By studying the ODE convergence, it follows that the asymptotically stable equilibrium points are

- $u = \frac{1-p}{p}\rho$ and $v = 0$ when $\nu < \rho$;
- $u = \frac{1-p}{p}\nu$ and $v = \frac{\nu-\rho}{a}\frac{1-p}{p}\nu = \frac{\nu-\rho}{a}u \leq u$ when $\nu > \rho$.

Therefore we expect the values of U_n and V_n to fluctuate around those values, except for V_n in the case it reaches zero and does not increase anymore (case $\nu < \rho$). These results are confirmed by the simulations, as we can observe from figures 6.9 and 6.10; only the case with $a = \nu + 1$ is shown, since there is no significant difference with the case $a = \nu + 1 - \rho$.

6.2.2 Distinct elements drawn

If now we consider D_n, the number of distinct drawn colors from the urn, we can use the same previous approach to describe it, heuristic again. First of all we have

$$P(D_{n+1} = D_n + 1 | U_n, V_n) = \frac{V_n}{U_n}$$

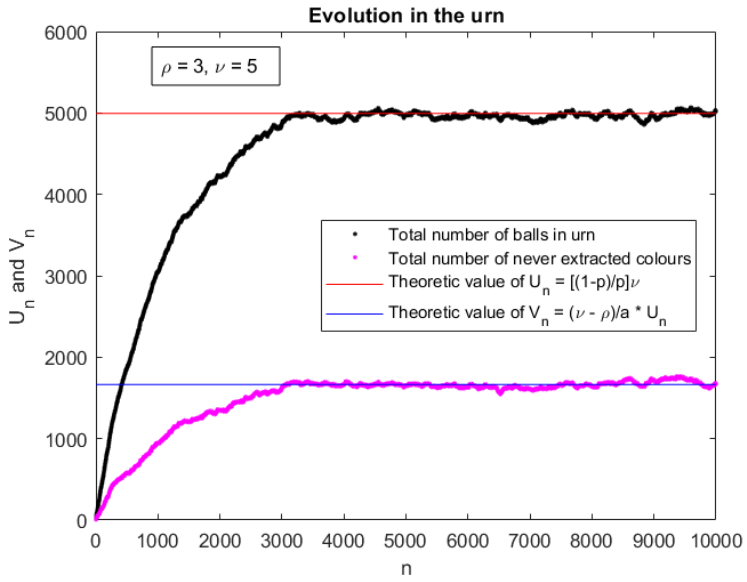


Figure 6.9: The plot shows how the number of balls in urn evolves in time and the number of balls (i.e. colors) that have never been drawn. The case considered is $\nu > \rho$ and $a = \nu+1$; the probability p is set to 10^{-3}. Red and blue lines show the theoretic value obtained in the analytical analysis.

and

$$E\left[D_{n+1} - D_n | \mathcal{F}_n\right] = \frac{V_n}{U_n}$$

that again can be thought as an ODE with corresponding approximating function $d(t)$:

$$\dot{d} = \frac{v}{u}.$$

From the previous analysis we know that in the case $\rho > \nu$, v goes to zero; this means that, after an initial increase, D_n will stay constant due to the disappearing of new colors in the urn. When $\nu > \rho$, at the equilibrium $\frac{v}{u} = \frac{\nu-\rho}{a}$, which means constant derivative for d and linear growth $D_n \sim \frac{\nu-\rho}{a} n$. This is the same result as the one obtained in the PUT model. Again both results are observed also in the simulations, in figures 6.11a and 6.11b.

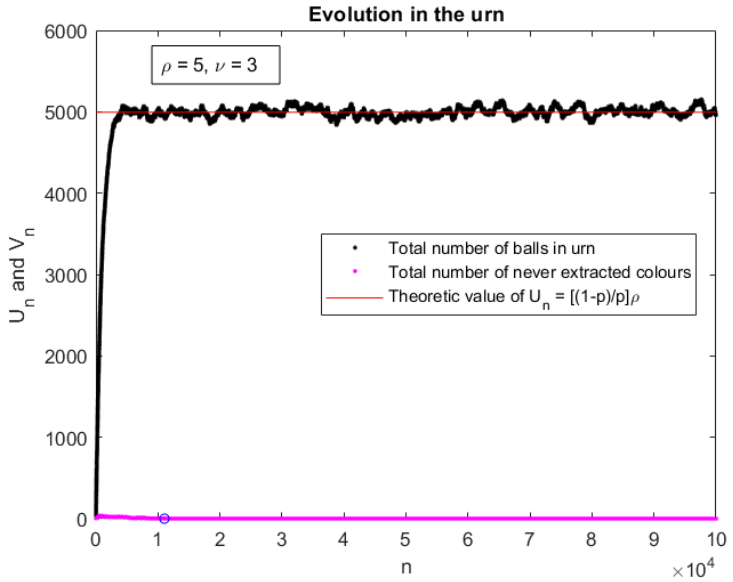


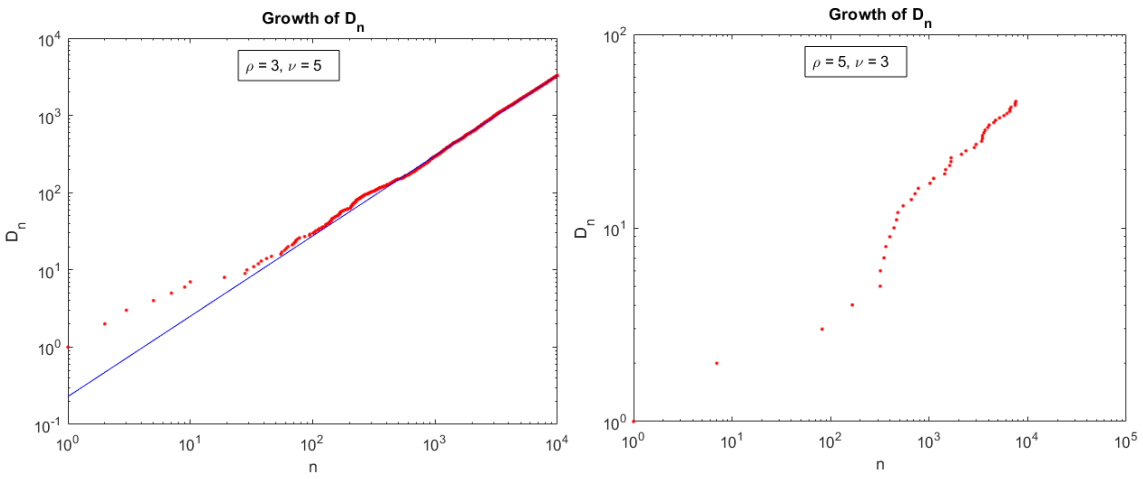
Figure 6.10: The plot shows how the number of balls in urn evolves in time and the number of balls (i.e. colors) that have never been drawn, which goes to zero. The case considered is $\nu < \rho$ and $a = \nu + 1$; the probability p is set to 10^{-3}. Red line shows the theoretic value obtained in the analytical analysis for U_n, while the blue circle is the time at which there are no more new balls in urn and V_n becomes 0, causing the interruption of innovations process.

6.2.3 Presence of a color in urn and frequency-rank distribution

Studying the presence of a general color i already present in the urn (notice that i in this case does not indicate the i-th distinct drawn), we can define $(C_n^i)_{n\geq 0}$, the stochastic process that describes it. Using the same approach as before, imagining at least one ball of color i in the urn at time n and, for simplicity, considering the model with the reinforcement from the first draw, we obtain

$$E\left[C_{n+1}^i - C_n^i | \mathcal{F}_n\right] = -pC_n^i + \rho\frac{C_n^i}{U_n}(1-p) = C_n^i\left(-p + \rho\frac{1-p}{U_n}\right). \tag{6.8}$$

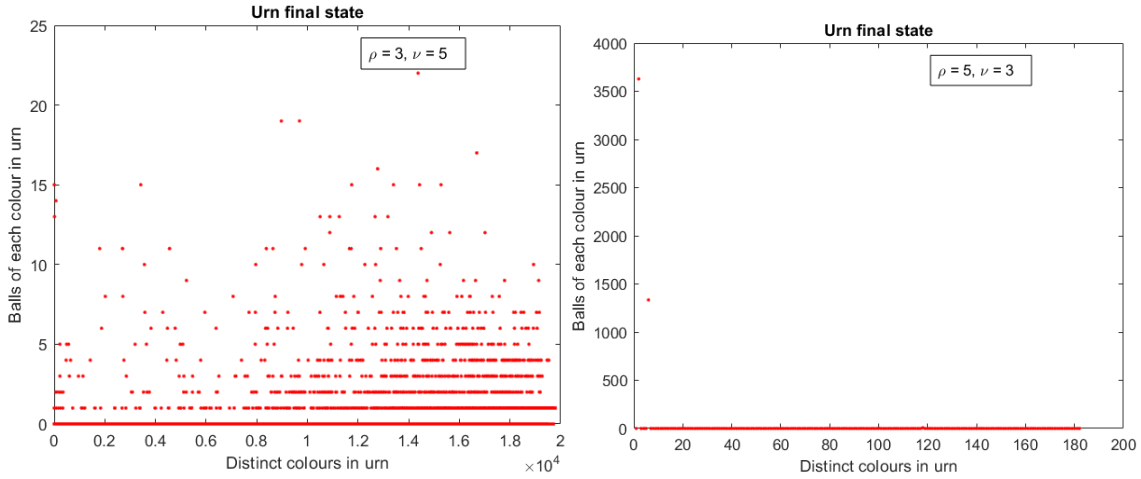
We can now make some observations. Since $C_n^i \geq 0\ \forall n \geq 0$ we expect it to increase if $U_n < \frac{1-p}{p}\rho$ and decrease if $U_n > \frac{1-p}{p}\rho$, following the expression (6.8). If $\rho > \nu$ we know that U_n fluctuates around $\frac{1-p}{p}\rho$, which brings the increment of C_n^i to zero and lets C_n^i fluctuate around the value it has reached. What really happens in the simulations is that some colors have a significant fraction of balls in the urn that allows them to survive and oscillate around the same value, while most of the colors disappear from the urn. If $\nu > \rho$ then $U_n \sim \frac{1-p}{p}\nu$, which means that as soon as U_n is above $\frac{1-p}{p}\rho$ the expected increment of C_n^i becomes negative, leading it to zero for every color i. This is also what happens in the simulations, where each color sooner or later disappears from the urn to make space



(a) $\nu > \rho$ (b) $\nu < \rho$

Figure 6.11: Evolution of D_n in the two usual cases: in the case $\nu > \rho$ we still observe a linear growth (estimated exponent 1.0397) as in the original model and a final fraction of distinct elements $\frac{D_n}{n} = 0.3298$, close to $\frac{\nu-\rho}{a} = \frac{1}{3}$. On the right $\nu < \rho$ and from about time $n = 10^4$ we do not observe any more new elements: the time it stops is about the time V_n becomes zero, as seen from figure 6.10.

for new colors. Figures 6.12a and 6.12b show the final number of balls of each color that has ever been in the urn, while figure 6.13 represents the evolution in time of the number of balls of the most frequent color in urn when $\rho > \nu$.
The frequency-rank distribution instead has only been observed with numerical simulations (figures 6.14 and 6.15) and not analytically. It is observed that when $\rho > \nu$ there are too few colors to estimate the power law. Otherwise, when $\nu > \rho$, it is observed the same pattern of the original model with the infinite urn and alive and dead colors both seem to follow the power law with exponent $-\frac{\rho}{\nu}$.



(a) $\nu > \rho$ (b) $\nu < \rho$

Figure 6.12: Final number of balls in the urn: each element on the x-axis represent color i in the urn; the value i can be also considered as an order of appearance in urn, since every time the triggering was applied the new colors were added extending the vector representing the urn at the end. We notice on the left that the most recent colors are generally more than the ones appeared first; on the right instead there are only 3 colors still alive, two of them among the initial ones, and all the other have no more balls.

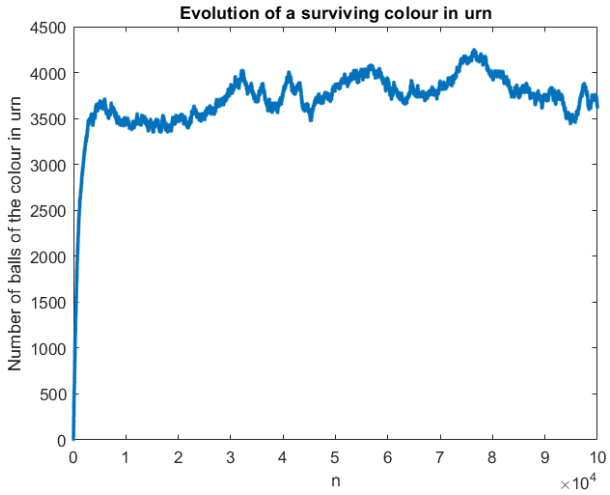


Figure 6.13: The plot shows the evolution of the number of balls of the most frequent color in urn in the case $\rho > \nu$: it fluctuates around a value but has a more oscillating behavior than the total number of balls.

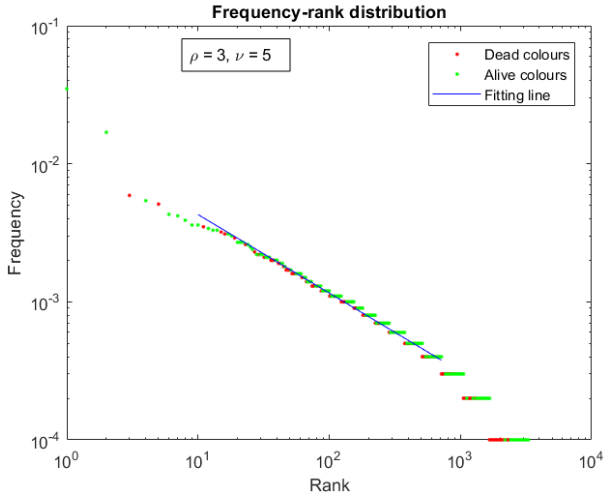


Figure 6.14: The plot shows the frequency-rank distribution of drawn colors when $\nu > \rho$, putting together elements still in urn with elements disappeared ("alive" and "dead"). The blue line is the regression line for the elements from 10th position in the rank and with at least 4 draws: the estimated exponent is -0.5687, close to $-\frac{\rho}{\nu} = 0.6$.

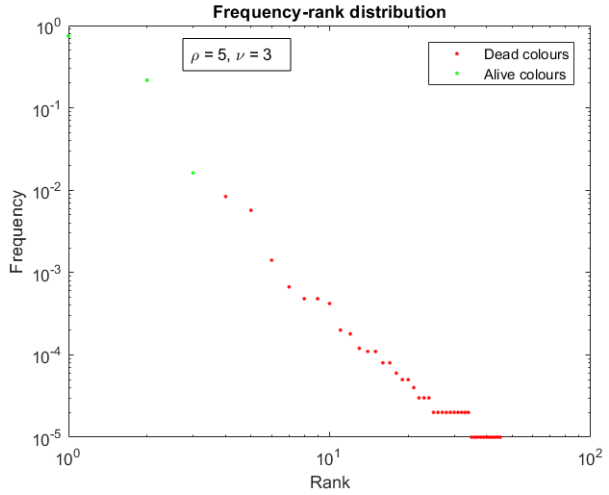


Figure 6.15: The plot shows the frequency-rank distribution of drawn colors, putting together elements still in urn with elements disappeared ("alive" and "dead"). In this case $\nu < \rho$ and the most frequently drawn elements are the ones still alive, while all the dead ones have a very fast decay, which was not estimated due to the low number of distinct colors ever drawn, only 45 after 10^5 draws.

Final results and discussion

The PUT model, modified such that each ball in the urn has a limited random life, was intended to represent a urn with limited capacity. We can divide the results focusing first on the urn and then on the draws.
In the urn we have the following behavior when n is large:

(i) when $\nu < \rho$, U_n fluctuates around $\frac{1-p}{p}\rho$ and V_n goes to zero, making zero the probability of drawing new colors;

(ii) when $\nu > \rho$, U_n fluctuates around $\frac{1-p}{p}\nu$ and V_n is around a constant fraction of U_n, $\frac{\nu-\rho}{a}$, giving a positive almost constant probability of drawing new colors.

Regarding the process $(D_n)_{n\geq 0}$:

(i) when $\nu < \rho$, after an initial increase, it remains constant on the same value due to the disappearing of new colors in the urn;

(ii) when $\nu > \rho$ it has the same linear behavior observed in the original model:

$$D_n \sim \frac{\nu - \rho}{a} n.$$

Regarding the frequency-rank distribution, the simulations highlighted this kind of behavior:

(i) when $\nu < \rho$ the most frequent colors are the only few still alive, while the others have a very fast decay;

(ii) when $\nu > \rho$, as in the PUT, the colors seem to follow a power law with $\alpha = \frac{\rho}{\nu}$, without a clear different pattern between colors still alive or not anymore in the urn.

Chapter 7

Conclusions

In this work a mathematical model for the emergence of innovations has been presented and developed. The necessity for a model has been discussed in Chapter 2: whether a single innovation gets more success than others generally depends on a huge number of factors, but, globally observing many heterogeneous systems, they seem to reproduce Heaps' and Zipf's law in the emerging of new items, ideas or whatever can be considered as novelty.

Many models in history tried to create mechanisms that could reproduce those laws but they somehow failed in representing both together. Reviewing those ones, the only model until now that have been able to reproduce those laws has been the Polya urn model with triggering (PUT), empowering the *richer-get-richer* mechanism and giving the abstract concept of *adjacent possible* a concrete application. Starting from this model in the thesis the following points have been completed.

- It has been defined the process describing the PUT and of the statistics $(D_n)_{n\geq 0}$ and $(K_n^j)_{n\geq 0}$: the former indicates the number of distinct colors drawn after n draws, while the latter indicates the number balls drawn for the j-th distinct color ever drawn.

- It has been used the stochastic approximation in order to obtain an ODE that describes asymptotically the normalized processes $\left(\frac{D_n}{n}\right)_{n\geq 0}$ and $\left(\frac{K_n^j}{n}\right)_{n\geq 0}$. This is a remarkable result, since starting from a stochastic process it has been possible to determine some of its characteristics with a deterministic approach.

- The study of equilibria of the ODE has determined the asymptotic points of convergence of the solutions of the ODE, based on the respective values of parameters ρ and ν, and on the initial condition in some cases.

- The results of the ODE analysis have been used in order to obtain expressions for D_n and K_n^j: mainly it has been proved that if $\nu < \rho$ the sub-linear behavior is reproduced as Heaps' law states, while if $\nu > \rho$ the number of distinct elements grows linearly.

Instead it has been always obtained, considering j large enough, $K_n^j \propto j^{-\frac{\rho}{\nu}}$, where j is considered as the order of draw for the color. This fact has shown that the PUT model favors the success of early drawn colors, and the order j can approximate the rank of the elements.

- All the results have been also observed in the analysis of the numerical simulation of PUT, in a large finite time.

It has been assumed that the number of draws of a color could determine a measure of success for the color itself, however the study has been focused on a global system and it has not given any suggestions for making an element (a song, an article or a start-up) more successful than another, since it depends on intrinsic characteristic of the idea considered.

However, if there is the possibility of managing the whole system, an idea of optimization could be the one of having the the most possible elements drawn more than a specific number k. For example in Quatrini thesis [19] the model was applied to fit the frequency-rank distribution of start-ups in some regional ecosystem, where the frequency was represented by the equity fundings received by the start-up, showing that in most cases it reproduced the Zipf's law. In a more abstract analysis the author supposed that ρ represented the ability of the ecosystem to exploit and develop the start-ups already present, while ν represented the ability to invest in order to explore and find new ideas for new start-ups.

Given this idea the analysis has been extended and developed in the two following points:

- It has been determined a recursive formula to calculate the number of distinct colors drawn at least k times, given the ones drawn at least $k-1$ times, with specific attention on the case $\nu > \rho$.

- Imagining a defined budget $C = \nu + \rho$ to invest in exploration and exploitation, it has been defined a trade-off between the two parameters such that the number of elements drawn at least twice or three times was maximized. It was considered the case $\nu > \rho$, which may be the case of an initial phase of life of the ecosystem in which there is the need of exploring more than exploiting, in order to find good ideas as fast as possible and exploit them later.

At the end the work has focused on applying some changes to the model in order to improve some weaker points. The changes applied have been two, one independent from the other.

- The first modification has been intended to give each color its own reinforcement parameter to represent the intrinsic quality of the innovation. Since in the PUT model it has been observed that the success of a color depended only on the order

of the first draw, in this way it has been hoped to observe also a dependence on the reinforcement parameter. After a rough analysis, numerical simulations have been run, sampling reinforcement parameters from three different distributions: they have confirmed that with this model even later but "better" elements could reach high positions in the rank, even though much larger parameters with respect to others are necessary for very late elements to get popularity. Rate of appearance of distinct elements and frequency-rank distribution have been observed to be similar to the PUT model, but with different coefficients.

- The second modification has been intended to represent the obsolescence of innovations with time and, in the aftermath, a urn with limited capacity. The model suggested has been the PUT with the difference that after each draw every ball in the urn has a small constant probability of disappearing: therefore each ball in urn has a geometric distributed lifetime. Again after a rough analysis numerical simulations have been run, observing a fluctuation of the number of balls in urn around a defined value. Regarding rate of appearance of distinct elements and frequency-rank distribution, when $\nu > \rho$ it has been observed a similar behavior as the PUT model, on the other hand when $\nu < \rho$ at some time the innovation process stops, keeping "alive" only some few elements.

Future work

The future work could go in two different directions: application of models to real life systems and improving of existing ones.

- The applications should focus on understanding what the parameters ρ and ν may concretely represent in a real system, through some statistical analysis on the attributes of the system; once determined it should be possible to apply optimization policies for the improvement of a start-ups ecosystem for example.
  The first modification of the model could be more specific, analysing how the intrinsic value of an innovation could be estimated and creating a specific model that could reproduce the history of a system.
  The second modification could be interesting in order to understand why some innovations disappear and the stagnation of some systems.

- In Chapter 2 it was presented the state of the art showing that in the last years there has been a great development of models for the emergence of innovations. The PUT model, together with its extensions from Tria et. al [5] and Monechi et al. [6], seems to be one of the best to reproduce Heaps' and Zipf's laws in the same model; however it would be interesting to create an expanding network model, more complex but capable of concretely represent the *adjacent possible*.

Appendix A

Stochastic approximation fundamentals

Let us suppose to have a generic stochastic process $(A_n)_{n\geq 0}$ which counts some quantity and may or may not increase by one at each time n of the process with probability that depends only on $B_n = \frac{A_n}{n}$:

$$A_{n+1} = A_n + \xi_{n+1}, \qquad \xi_{n+1} \sim \text{Bernoulli}(p(B_n)),\ n = 0,1,\dots \tag{A.1}$$

where $p(B_n) : \mathbb{R} \to [0,1]$, $\forall n \geq 0$.
Borkar, in Chapter 2 of his book [20], introduces a lemma which is useful in order to study these kind of processes. It relates some kinds of stochastic processes, specifically for urn models, to the study of ordinary differential equations (ODE), which are deterministic. The following part of this appendix will present all the assumptions needed in order to prove the lemma, enunciated at the end.
At first let us rewrite the expression in (A.1) in order to use B_n instead of A_n, starting start from

$$\frac{A_{n+1}}{n+1} = \frac{n}{n+1} \cdot \frac{A_n}{n} + \frac{\xi_{n+1}}{n+1}$$

that becomes

$$B_{n+1} = \frac{n}{n+1} B_n + \frac{\xi_{n+1}}{n+1}$$

and then

$$B_{n+1} = B_n + \frac{\xi_{n+1}}{n+1} - \frac{B_n}{n+1},$$

which finally, adding and subtracting $\frac{p(B_n)}{n+1}$ leads to

$$B_{n+1} = B_n + \frac{1}{n+1} \left[p(B_n) - B_n + \xi_{n+1} - p(B_n) \right]. \tag{A.2}$$

In a more general context it is useful to consider multidimensional processes instead of scalars, therefore we consider $A_n, B_n \in \mathbb{R}^d$ $\forall n \geq 0$ while $\xi_{n+1} \in \{0,1\}^d$ $\forall n \geq 0$ and the

map $p : \mathbb{R}^d \to [0,1]^d$, $d \geq 1$.
Now it possible to define $h(B_n) = p(B_n) - B_n$ and $M_{n+1} = \xi_{n+1} - p(B_n)$ and rewrite expression (A.2), considering a general step-size $a(n)$ instead of $n+1$:

$$B_{n+1} = B_n + a(n)\left[h(B_n) + M_{n+1}\right]. \tag{A.3}$$

This last expression can be considered as more general form of the (A.2), but, in order to prove Lemma 1, it should satisfy the following assumptions.

Assumption 1. *1. The map h is a Lipschitz function: $||h(x) - h(y)|| \leq L||x - y||$ for some Lipschitz constant $0 < L < \infty$.*

2. The step-sizes $\{a(n)\}$ are positive scalars such that

$$\sum_{n=0}^{\infty} a(n) = \infty, \quad \sum_{n=0}^{\infty} a(n)^2 < \infty.$$

3. $\{M_n\}$ is a martingale difference sequence with respect to the increasing family of σ-fields

$$\mathcal{F}_n = \sigma(B_m, M_m : m \leq n) = \sigma(B_0, M_1, \ldots, M_n), \quad n \geq 0,$$

i.e.

$$E[M_{n+1}|\mathcal{F}_n] = 0 \text{ a.s., } n \geq 0.$$

Moreover $\{M_n\}$ are square-integrable:

$$E[||M_{n+1}||^2|\mathcal{F}_n] \leq K(1 + ||B_n||^2) \text{ a.s., } n \geq 0,$$

for some constant $K > 0$.

4. It holds $\sup_n ||B_n|| < \infty$ a.s..

5. This last condition is a generalization of the expression (A.3) in the case it presents in the following form:

$$B_{n+1} = B_n + a(n)\left[h(B_n) + M_{n+1} + \epsilon(n)\right], \tag{A.4}$$

where $\epsilon(n)$ is either random or deterministic. It should satisfy then $\lim_{n\to\infty} \epsilon(n) = 0$.

Next it is possible to define time instants $t(0) = 0$ and $t(n) = \sum_{k=0}^{n-1} a(k)$, $n \geq 1$, and the intervals $I_n = [t(n), t(n+1)]$, giving the following definitions.

Definition 3. *$\bar{b}(t)$ is a piecewise linear function interpolated on the values of the process $(B_n)_{n\geq 0}$, such that $\bar{b}(t(n)) = B_n$, $\forall n \geq 0$. The linear function on each interval I_n is*

$$\bar{b}(t) = B_n + (B_{n+1} - B_n)\frac{t - t(n)}{t(n+1) - t(n)}, \quad t \in I_n.$$

Definition 4. *The functions $b^s(t)$ and $b_s(t)$ ($s \in \mathbb{R}$) are the solutions of Cauchy problem*

$$\dot{b} = h(b) \tag{A.5}$$

respectively with initial condition $b(s) = \bar{b}(s)$, defined on $[s, +\infty]$, and with ending condition $b(s) = \bar{b}(s)$, defined on $[-\infty, s]$.

After presenting the assumptions and definitions needed we can enunciate the Lemma.

Lemma 1 (See Borkar, Lemma 1 [20])**.** *Given a stochastic process in the form* (A.3) *or* (A.4)*, satisfying all the assumptions from Assumption 1, and defined $\bar{b}(t)$, $b^s(t)$ and $b_s(t)$ as in definitions 3 and 4, for any $T > 0$ it holds*

$$\lim_{s\to\infty} \sup_{t\in[s,s+T]} ||\bar{b}(t) - b^s(t)|| = 0 \quad a.s. \tag{A.6}$$

$$\lim_{s\to\infty} \sup_{t\in[s-T,s]} ||\bar{b}(t) - b_s(t)|| = 0 \quad a.s. \tag{A.7}$$

Concretely Lemma 1 states that asymptotically the ODE solution $b(t)$, given an initial (or ending) condition at a very large time s such that has the same value as $\bar{b}(s)$, have the trajectory that is tracked by the values of the process B_n almost surely; then, by solving the Cauchy problem and through a careful analysis, it is possible to give an approximation of the values of the process for large times.

Bibliography

[1] *Mathematical Model Reveals the Patterns of How Innovations Arise*. https://www.technologyreview.com/s/603366/mathematical-model-reveals-the-patterns-of-how-innovations-arise/. 2017-01-13.

[2] Stuart A Kauffman and Richard C Strohman. *The Origins of Order: self organization and selection in evolution*. Vol. 993. Oxford university press New York, 1994.

[3] Stuart A Kauffman. *Investigations*. Oxford University Press, 2000.

[4] Vittorio Loreto et al. "Dynamics on expanding spaces: modeling the emergence of novelties". In: *Creativity and universality in language*. Springer, 2016, pp. 59–83. URL: https://arxiv.org/abs/1701.00994.

[5] Francesca Tria et al. "The dynamics of correlated novelties". In: *Scientific reports* 4 (2014), p. 5890. URL: https://www.nature.com/articles/srep05890.pdf.

[6] Bernardo Monechi et al. "Waves of novelties in the expansion into the adjacent possible". In: *PloS one* 12.6 (2017), e0179303. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179303.

[7] Iacopo Iacopini, Staša Milojević, and Vito Latora. "Network dynamics of innovation processes". In: *Physical review letters* 120.4 (2018), p. 048301.

[8] HS Heaps. "Information Retrieval-Computational Aspects Academic Press". In: *New York* (1978).

[9] George K Zipf. *Human behavior and the principle of least effort*. 1950.

[10] Sandy L Zabell. "Predicting the unpredictable". In: *Synthese* 90.2 (1992), pp. 205–232.

[11] HA Simon. "HA Simon, Biometrika 42, 425 (1955)." In: *Biometrika* 42 (1955), p. 425.

[12] Fred M Hoppe. "Pólya-like urns and the Ewens' sampling formula". In: *Journal of Mathematical Biology* 20.1 (1984), pp. 91–94.

[13] George Pólya. "Sur quelques points de la théorie des probabilités". In: *Ann. Inst. H. Poincaré* 1.2 (1930), pp. 117–161.

[14] Norman Lloyd Johnson and Samuel Kotz. “Urn models and their application; an approach to modern discrete probability theory”. In: (1977).

[15] H Mahmoud. *Pólya urn models. Texts in Statistical Science*. 2008.

[16] Bernat Corominas-Murtra, Rudolf Hanel, and Stefan Thurner. “Understanding scaling through history-dependent processes with collapsing sample space”. In: *Proceedings of the National Academy of Sciences* (2015), p. 201420946.

[17] François Jacob, Franpcois Jacob, and Fran Jacob. “The possible and the actual”. In: (1982).

[18] Steven Johnson. “Where Good Ideas Come From: The Natural History of Innovation”. In: (2010).

[19] Francesco Quatrini. “Gli ecosistemi dell’innovazione: alcune proposte per la costruzione di un modello matematico”. MA thesis. Politecnico di Torino, 2018.

[20] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Vol. 48. Springer, 2009.