Maynooth University
National University of Ireland Maynooth

# On the average generation of a population

A dissertation submitted for the degree of
Doctor of Philosophy


By:

**Gianfelice Meli**

Under the supervision of:

Prof. Ken Duffy

Hamilton Institute
National University of Ireland Maynooth
Ollscoil na hÉireann, Má Nuad


20 May 2019

# Contents

# Abstract

Estimating the average generation of a collection of cells is helpful in understanding complex cellular differentiation processes, identifying carcinogenic cellular activities, and quantifying the ageing of the immune system. Different techniques based on both direct observations and indirect inference have been proposed, with benefits and limitations varying in the two categories.

In this thesis we enhance the mathematical results underpinning one of these inference methods, firstly proposed by Weber et al. in 2016 [116] and based on a DNA coded randomised algorithm. Assuming some sort of structure in the growth of a cell population, with the use of Branching Processes and Renewal Theory, we establish improved convergence properties of the proposed estimator to the average generation. Expanding and homeostatic populations are studied, allowing the method to be used for more complex patterns of population dynamics that includes the succession of these two phases. Furthermore, we establish the possibility of using the same method in a two-type branching process, obtaining a possible criterion to distinguish among some differentiation models in hemapotoiesis. A quality study of the model allows also us to establish values of the parameters which improve the performance of the estimator. Computer simulations, with parametrisations coming from the immunology field, are along the results with both a validation and exploratory purpose.

CHAPTER 1

# Introduction

## 1.1 Motivation

Every day our body produces millions of cells in order to extend, repair, and renew all its tissues. In order to do that and produce the wide variety of cells we need, processes of cellular division, differentiation and death are continuously in place. Fig. 1.1 describes a generic process of this type. Here, with the use of a tree structure, the relationships between a cell and its descendants are highlighted, showing not only the divisions occurred, but also the deaths and the changes of cell type. In mathematics, this is not the typical way to represent a family history, where usually the ancestor is at the top of the tree and the descendants at the bottom, but Fig. 1.1 provides also a temporal context that gives us more an idea of movement, thanks to a horizontal time-line. The situation illustrated by Fig. 1.1 occurs throughout our bodies, and now we will give three specific examples that help contextualise the motivation for the mathematical work undertaken in this thesis.

Let's think of the processes that allow the creation of a tissue starting from tissue-specific stem cells. The latter are cells that have two special features: following a differentiation process, they are able to generate all the different cells a tissue is made of; they have the ability to indefinitely give birth to cells of their own type (i.e. they are said to be self-renewing). In particular, let's focus on the hematopoiesis, the process by which $10^{11} - 10^{12}$ blood cells are formed each day in a typical human [19]. Compared to other tissues, the stem

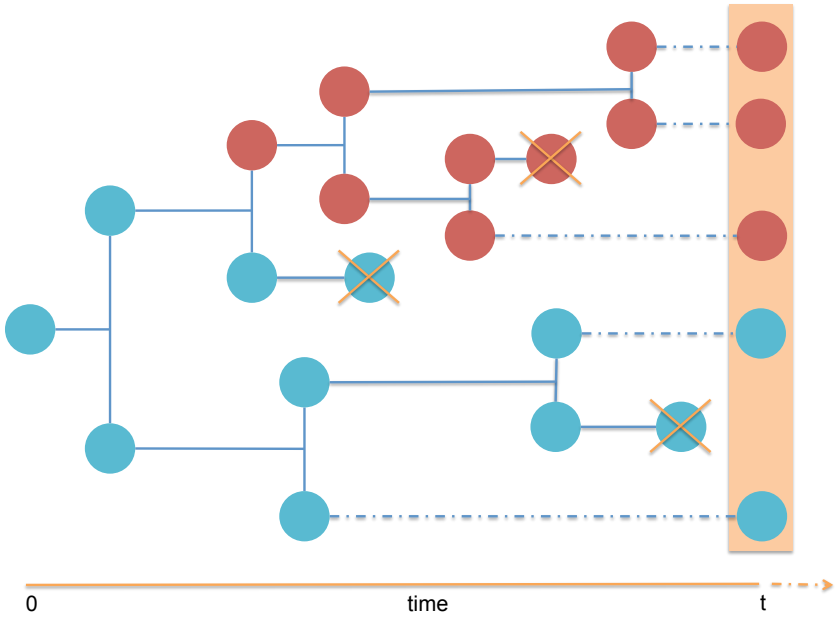



Figure 1.1: **Cell population growth.** Under appropriate stimulation, a cell undergoes a process of division that leads to the birth of 2 new cells. The plot in the figure describes the growth of a population that starts with one cell at time 0 and, after consecutive divisions, is made up of 5 cells at time $t$. In this time frame, cells can be also subject to death (crossed circles) or differentiation, i.e. change of type (change of colour).

cells of the blood system are easier to locate, collect, and study using in vitro cultures, and this makes them popular among both experimental and theoretical researchers [87]. Red blood cells, neutrophils, lymphocytes, and all the other cells that constitute the blood are ultimately products of Hemapotietic Stem Cells (HSC) as a result of a differentiation process. For many years this differentiation process has been seen as the directed tree in Fig. 1.2, where every movement in it was corresponding to a major degree of differentiation, i.e. to a more restricted lineage potential. From that viewpoint, myeloid (erythrocytes, neutrophils, etc.) and lymphoid (T cell, B cells, etc.) lineages are separated early on at the stage in which are produced the Common Myeloid and the Common Lymphoid Progenitors (CMP and CLP). Recent work has challenged this model, questioning in particular the uniqueness of the path that brings to each cell type [87] or the fact that the CMP-CLP split is the first step of commitment [88, 52]. Understanding the differentiation stages in which biological cues can determine the fate of the differentiation process is

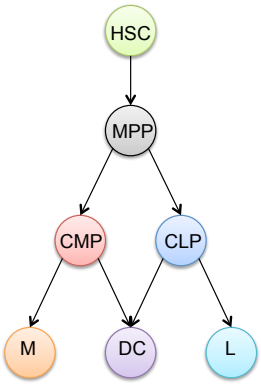



Figure 1.2: **Classical hematopoietic tree.** The prevailing hematopoietic model of the last 3 decades can be drawn with various level of details [94, 37]. Here, for illustration, we detail only few layers. Hematopoietic Stem Cells (HSCs) are self-renewing and capable of generating all the types of blood cells. After division, they can either give birth to HSCs or Multipotent Progenitors (MPPs), cells that retain HSCs potential but are unable to self-renew. So, they can only differentiate in either a Common Myeloid or Common Lymphoid Progenitors (CMPs and CLPs). At this stage we have the separation between the myeloid and the lymphoid branches of the lineage. In particular, Myeloid cells (M), such as erythrocytes and neutrophils, come from CMPs, while Lymphoid cells (L), such as T and B cells, come from CLPs. According to this model, Dendritic Cells (DCs) are the only blood cells that derive from both CMP and CLP.

still a matter of research.

Another example of a process characterised by division and differentiation is an adaptive immune response to an infection. Adaptive immunity is the part of the immune system that provides a tailored response to infections. It does so by integrating information from direct observation of the pathogen, from signals coming from other cells that have recognised the presence of the pathogen, and from previous exposures to the same threats. The two most important components of this group are B and T cells, both parts of the lymphocytic population. The adaptive immune system starts to operate when the organism recognises an antigen, the general term used to describe any substance that can trigger an adaptive immune response (Fig 1.3). Each B and T cell constitutively presents on its surface a specific antigen receptor, as a result of a random process that makes it unlikely for two members of the population to have the same receptors. At this stage these cells are called

naïve for the fact that they have not been exposed to any antigen yet. Once an antigen is encountered for the first time, we have the so called primary response, where B and T cells that bind to the antigen receive signals to activate a clonal expansion. In this way, from a pool of few specific fighters (5-500 cells [61]) it is possible to construct an army against the antigen. Because of this specific antigenic process, usually it takes 1-2 days before the clonal expansion starts, during which time a series of checks and balances takes place in order to be sure of the presence of an infection [84, Chapter 11]. Different is the situation if the body has already seen the antigen before, where in this case we will talk about secondary response. After the end of a primary response, some B and T lymphocytes, kin of the ones that fought the antigen during the first encounter, are kept alive allowing a reaction to a second antigen exposure very fast, most of the time without any symptoms. These cells, called memory cells, are more sensitive to stimulation than naïve cells, having less stringent requirements for activation, and this reduces the time required for them to start a new clonal expansion. A description of the clonal expansion B and T cells are subject to can be seen in Fig. 1.4. Memory cells are able to live even for all our life, and are distinguished from the cells that actively fight the infection called effector cells, which are killed once the threat is disappeared. Both types are produced for the first time during the primary response from the naïve ones, but the correct order is not still clear [16, 64]. Stimulate the production of antigen-specific memory cells with small and controlled amount of antigen is the idea behind any sort of vaccinations.

Studying processes of division, differentiation, and death is not only important to understand the daily processes that determine the correct functioning of our body, but also to have insights on possible criticalities that can put us in danger. This is the case for example for the process that brings to the formation of a cancer. During its life, a cell can be subject to DNA damage. There can be an error during the replication of the DNA (e.g. mismatched bases), some gene alterations caused by environmental agents (UV light, alkylating agents, etc.), or even a spontaneous DNA damage [76]. These are not rare events and the reason why we normally don't have consequences is that cells have also mechanisms that allows them to repair these errors. However, the majority of cancer cells arise from the accumulations of mutations in the normal DNA sequence, that cannot be repaired by the cell, and cause an increase of the

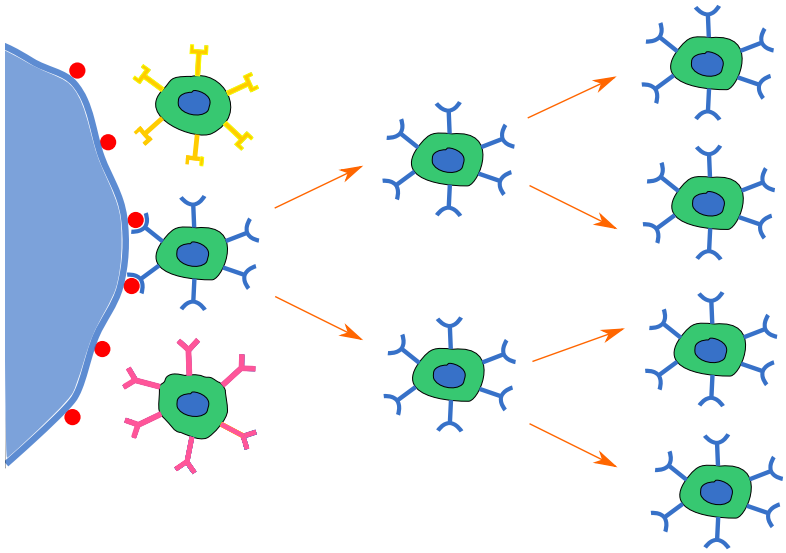


Figure 1.3: **Clonal selection.** Illustration that explains the process that brings an antigen to trigger the expansion of pathogen-specific B cells. Red dots represent antigens expressed on the surface of the blue cell on the left. The green cells are B cells, each of which has its own B Cell Receptor (BCR). The cell whose BCR is complementary to the antigen and so forms a good bond, is activated and starts dividing, creating a clonal army that can fight the infection [84].

replicative capability of the cell. So, Fig. 1.1 can also describe the process that gives birth to a cancer cell and lead to the formation of a tumour. Even if all the cancer populations have in common the faster growth rate compared to healthy ones, not all the cancer are characterised by an increase in the frequency of division. Indeed, it has been seen that some of them obtain the same result by only escaping from death. New drugs, as Venetoclax, target proteins central for the survival of the these cells [96].

So, divisions, differentiations and deaths are presents in a lot of regulatory processes inside our body. In generality, let's consider a growing cell population. For each descendant, let's call generation the number of divisions that led to that cell. The members of the initial population will be defined to have generation 0. Fig. 1.5 describes the generation counting for the descendants of a cell in a time frame $t$. In a cell population, generation dependent behaviour has been implicated in the risk of cancer and its evolution [33, 77, 110], as well as being a determiner in the complex differentiation dynamics of proliferating

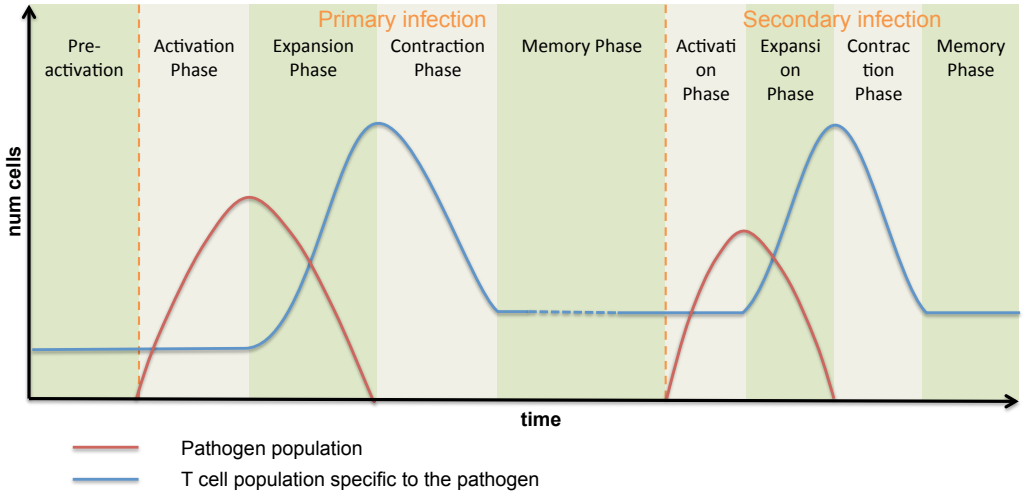



Figure 1.4: **Non-data description of T cell reaction during immune response.** The plot in the figure is adapted from [62]. Before antigen first exposure, the population of naïve T cells specific to the pathogen is around 5-500 units. Once they encounter the pathogen, we have to wait 1-2 days before they start the clonal expansion [84, Chapter 11]. During the expansion phase, the number of T cell specific to the pathogen grows exponentially (sometimes over 10,000 times the original number [102]) allowing them to fight and extinguish the pathogen population (7-12 days post infection). After the resolution of the infection, it starts a contraction phase, in which around 90-95% of the T-cells produced die out [102]. The T cell population comes out from the infection enlarged (100-1000 as many as prior to infection) and enriched from a pool of memory cells, that are able to respond with greater effectiveness in case of re-infection. In fact, in case of a second encounter with the same pathogen, the secondary response will be faster thanks to the fact that there are more T cell specific to the pathogen, and that the memory cell are quicker in respond upon antigen detection. The secondary response will go through the same steps (activation, expansion, contraction, and memory phase) but will be faster than the primary one.

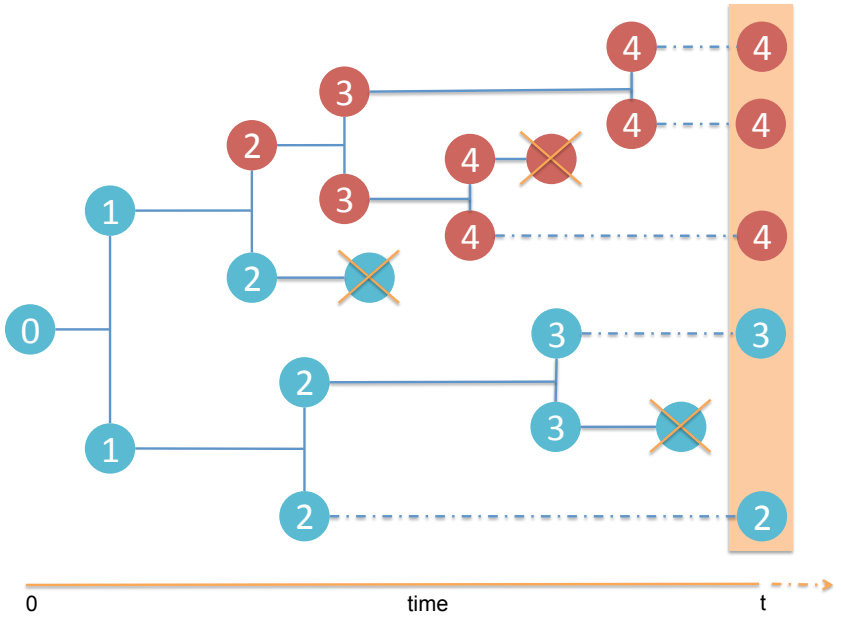



Figure 1.5: **Generation counting in a cell population.** The growth of the cell population described in Fig. 1.1 is reobserved. This time, every cell is also equipped with the information concerning its generation (white number inside every circle).

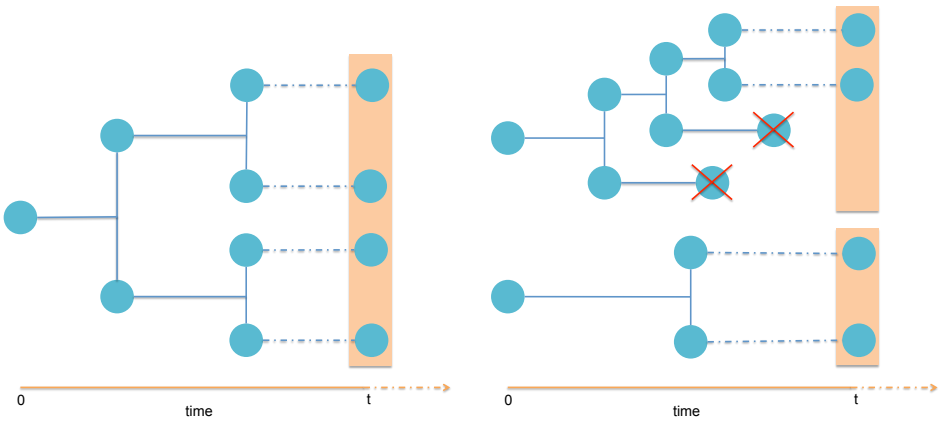
cell systems [47, 108, 112, 46, 26, 118, 24, 74].

If we dispose of information concerning the sizes of starting and final populations, and assuming that cells with same generation divide simultaneously, giving always birth to 2 cells (i.e. with no death), we can estimate the generation of the cells without problems (Fig. 1.6(a)). In all the other cases, the number of members of the population is not enough to infer information concerning the generation of it, not even their average generation (Fig. 1.6(b)).

In the next section we will describe some of the experimental techniques that have been developed to accomplish this task.

## 1.2 Techniques for the estimation of the average generation

A range of experimental techniques have been developed that allow evaluation or estimation of the generations of cells. Due to the possibility to have a greater control on the variables of the experiment, most of these techniques




(a) (b)

Figure 1.6: **Population size doesn't typically determine the average generation.** (a) When a population of cells grows synchronously and is not subject to death, it is possible to find the generation $g$ of the living cells using only the size of the population at a single time. In fact, if the population was originated by $n$ cells, the size of the living population would have to be $2^g n$ (orange box, $n = 1$, time $t$, size $= 2$, $g = 2$). (b) In presence of asynchronism or death, knowledge the number of cells alive at a single time does not uniquely determine the average number of divisions that lead to to the living cells (i.e. the depth of the family tree).

have been developed for in vitro studies, i.e. for situations where cells are let grow in a controlled environment, such as a culture dish. However, recreating in a laboratory the exact same conditions in which a cell population normally grows is not possible, because of the myriad of interactions a cell can be influenced by. So, the in vivo setting, i.e. the scenario where a cell population is studied inside the organism it belongs, remains the biggest challenge and the preferable case for the researchers. Adapting in vitro techniques to in vivo experiments is not always possible, because we are not always able to obtain the required information outside a culture dish. Even when this is possible, the measurements are usually subject to greater noise in the in vivo experiments, making the method useless.

Technological progresses bring to the development of new techniques to observe cells in vivo, but methods based on inference rather than on direct observations are the most informative at the moment. With the use of limited amount of information and probabilistic techniques, these techniques can infer

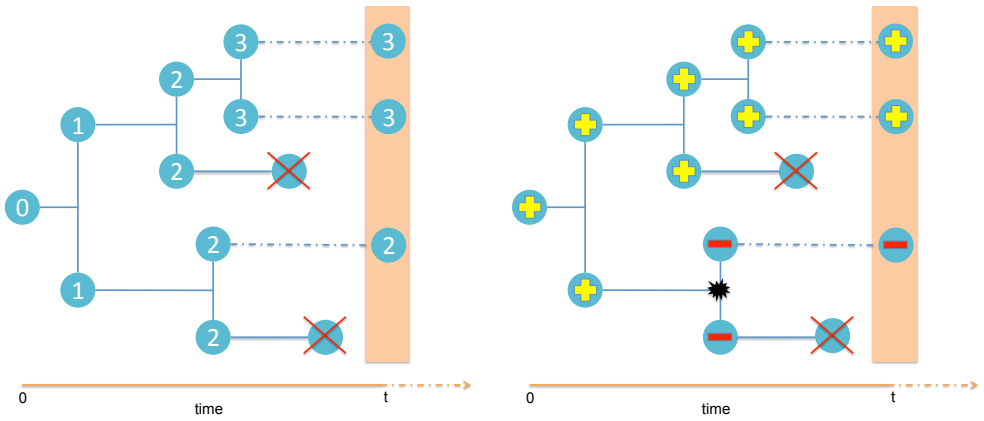



Figure 1.7: **Average generation.** (a) With the progenitor being defined to be in generation 0, the total generation of the process at any time is the sum of the generations, i.e. the number of edges back to the root of the tree, of living cells (orange box, $G(t) = 3 + 3 + 2 = 8$). If $Z(t)$ is the size of the population at time $t$, the average generation is the total generation divided by the number of living cells, $G(t)/Z(t) = 8/3 = 2.67$. (b) The randomised algorithm proposed in [116] for inferring $G(t)/Z(t)$ is based on having a neutral label in the initial cell that is independently lost with probability $p$ during each cell's lifetime (indicated by a black cloud) and is not regained by further offspring once lost. If the proportion of label-positive cells can be measured and the probability of label loss, $p$, is small, then the following relationship holds $G(t)/Z(t) \approx -1/p \log(Z^+(t)/Z(t))$ in two approximate senses more fully explained in [116].

the generation of a cell. If on the one hand the results so obtained are not deterministic, on the other hand they provide answers in situations where the other methods give a poor description. This is the case, for example, when we work with big populations or when we follow the growth of a population for long times.

In the following, we will give a short description of the most used methods, dividing them in two classes: techniques based on direct observations and on inference methods.

### 1.2.1 Direct observations based techniques

#### Time lapse microscopy

One of the oldest and most popular way to track entire lineages of cells is the use of a time lapse microscopy [89, 106, 107, 44, 36, 35, 95]. This technology consists of a microscope that is also able to record images of the field of view at discrete times for a fixed time [29]. The photos thus obtained, are then played consecutively at a faster rate than the one at which they have been captured, creating a movie in which the time grows faster than normal. In this way, if a cell population is let grow in a culture dish, we are able to follow its growth and recreate entire cellular lineages.

The first time-lapse films of cultural cells were made in France back in 1907, but only in 1914 the first purpose-built microcinematographic apparatuses become commercially available in Europe [68]. Since then, technological improvements in the photography field and in the techniques to highlight the cell components, have allowed an enhancement of the quality of the images and so of the utility of the instrument.

This technique provides the best performances for in vitro studies, where most of the factors that obstruct the view can be eliminated or controlled. However, it is also used for in vivo experiments where the region of interest is easily accessible and not in deep [70, 13]. We need special microscopy for these experiments, time-frames are typically limited and we cannot track cells that leave the immediate, unconstrained region being observed.

Another significant limitation of time-lapse microscopy is the fact that cells cannot accurately be followed for a long time. After a while, it becomes impossible to establish the relationship between cells across frames, both because cells become numerous in the field of view and because they move in the medium and start to gather in tridimensional structures [29]. This forces us to study small size population, and even in this case the time lapse observation provide useful information only up to 10 generations.

#### Fluorescent dye dilution

Another popular methodology to estimate the generation of a cell population is the fluorescent dye dilution. It consists of staining initial cells with a

fluorescent dye such that with each division cells inherit approximately half of the molecules from their parent and thus fluoresce with half their intensity [73, 71, 43, 90]. A cell's generation can thus be inferred from its luminous intensity via a technology called flow cytometry. Fig. 1.8 gives an illustration of how this technique works.

This method was first introduced in 1994 [73], and since then it has been the most used one for analysing in vitro and in vivo division of lymphocytes and other hematopoietic cells [72]. In general, this high throughput approach is suitable for adherent cells that cannot be tracked optically, and can be used for in vivo adoptive transfer experiments, where cells are transplanted from one animal into another. In most applications division tracking dyes are used to determine the distribution of a population across generations, but recent developments have created an experiment design where the offspring of individual clones can be identified via colour multiplexes of distinct division diluting dyes [75, 49]. Genetically modified mice also exist that enable an inducible equivalent of a division diluting dye *in vivo* without the need for adoptive transfer of *ex-vivo* stained cells, e.g. [32].

Although these methods allow us to work with cell population of any size, it gives us the possibility to follow them only for 7-8 generations. Indeed, after that the fluorescent signal-to-noise ratio is too low for a cell's generation to be reliably determined.

### 1.2.2 Inference based techniques

In the previous section, we have seen that there are situations, where it is not possible find the generation of a cell through direct observation. This happens for example when we are interested in following the cell population for more than 10 divisions, or when we are not able to find the required information from the in vivo experiment. A solution in these situations is given by methods that rely on inference rather than on direct determination. Let's describe two of the most known techniques.

#### Telomere length

In all eukaryotic cells, at the end of each chromosome, we find special nucleotide sequences that prevent chromosome end-to-end fusions and enable it

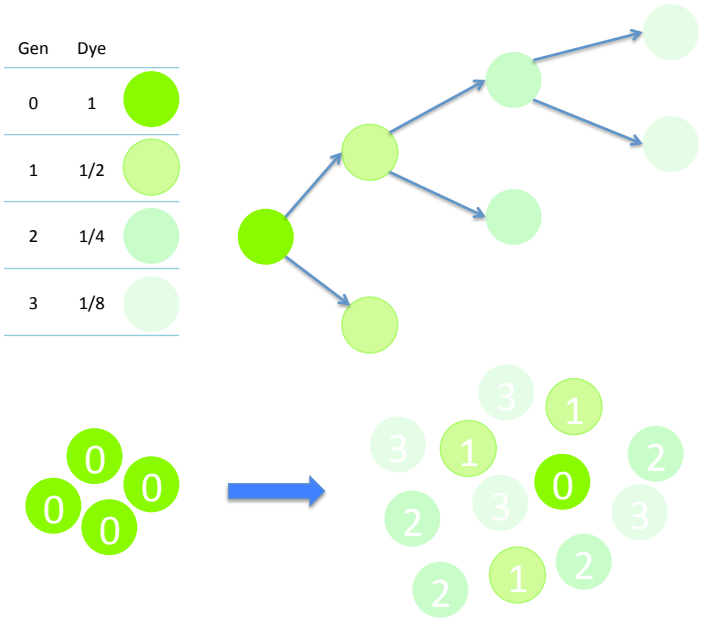



Figure 1.8: **Average generation estimation through fluorescent dye dilution.** The plot describes the growth of a population where cells have been incubated with a fluorescent dye, such as CarboxyFluorescein Succinimidyl Ester (CFSE). Each time a cell divides, the amount of CFSE present inside it is perfectly divided between the two daughters, causing them to fluoresce with half of the intensity. The generations of the cells can then be analysed by flow cytometry, which will provide us with an histogram with the CFSE fluorescence. Through this histogram it is possible to know the size of the populations at each generation. This assay is able to detect up to 7-8 cell divisions, after which CFSE fluorescence can no longer be measured [81]. While this first dye is fluorescent green, more recent dyes are available in a range of colours.

to be efficiently replicated without loss of important genetic material [3, pg. 263]. These regions are called telomeres.

In most somatic cells, i.e. the cells that form the body of the an organism, telomeres lose 50 to 100 base pair (bp) at every cell division [50] due to incomplete end-replication [83]. When the critical length is reached, they lose the capacity to divide, and cells either die by apoptosis or enter senescence [45], i.e. a status of permanent cell cycle arrest.

Telomere length is usually reconstructed thanks to the telomerase, an enzyme that adds to the telomere new nucleotide sequences repeats. Despite the

existence of a mechanism to repair the part of telomere lost, it seems that this is not enough to keep constant its length [99]. So, the telomere length provides indirect information about the replicative history of a cell that can be used to estimate the generation of the cell.

Studies that make use of measurement of average telomere length in order to estimate replicative tree depth in vivo are [41, 4, 113, 117, 99, 46].

### Somatic mutations

During its lifetime, a cell can collect a series of changes in the DNA sequences that may or may not have an impact in the phenotype or the functioning of the cell. When a cell divides, these mutations, together with the ones that rise from errors during DNA duplication, are transmitted to the daughter cells, which in turn will start to collect its own. So, from the sequencing of the genome of a cell, in theory should be possible to infer its ancestry.

As examining the whole genome of a population of cells is not currently feasible, one can focus on studying the mutations in particular regions of the DNA. To make sure that these regions are informative, it is better to choose areas of DNA characterised by a higher mutation rate. Microsatellites, also called Short Tandem Repeats (STRs), are some of these regions and are constituted by the repetition of short (1-6 base pairs) DNA motifs. A mutation in microsatellites causes the deletion or addition of 1-2 repeats of the DNA motif in the sequences and so the length of it can be used to establish the degree of relationship existing among individuals. The high level of polymorphism makes microsatellites useful for different applications. Very famous is the use for forensic identification, where since November 1997 microsatellites have been used by the FBI to link criminals with the scene of crime.

Examples of work that use the number of mutations in microsatellite regions to establish the generation of the cells are [104, 111, 105, 114, 91, 17].

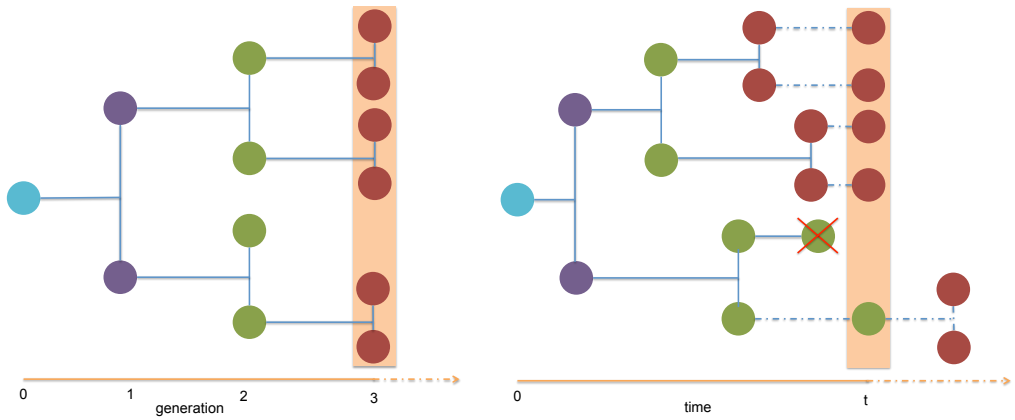
## 1.3 Branching processes as mathematical models of population dynamics

In Section 1.2, we have briefly described some of the most popular experimental techniques to study the average generation of a cell population, highlight-

ing pros and cons in both in vitro and in vivo settings. In this section we introduce the most commonly used class of mathematical models that allows the description of the growth of a population of replicating objects via a tree structure that develops in time: the branching processes. Our intent is to give an introduction of the mathematical framework we work in starting from Section 1.4, where we recapitulate and explain a recently proposed inference method that is the motivation for the work in this thesis. More information on the processes that we briefly describe in this section can be found in books such as [42, 78, 10, 55, 63]. For an historical reconstruction of the different mathematical models used for characterising population dynamics, and in particular on the evolution of branching processes, we also refer to [58, 11].

The first branching process model, which started the whole theory, is called Galton-Watson branching process. Named after Francis Galton (1822-1911) and Henry William Watson (1827-1903), British polyhistor and mathematician respectively, this process was introduced in a paper in 1875 to study the probability of extinction of family names, a cause of concern among aristocrats of the Victorian age [115]. In the simplest form of a Galton-Watson process every member of the population, independently of each other, is substituted by a random number of offspring, according to a common probability distribution, $N$, at integer lifetimes that correspond to generations. That is, this process is synchronous, with all members of the same generation counted as being alive at the same time. We denote by $Z_n$ the number of individuals in the $n$-th generation and we assume $Z_0 = 1$. Due to the independence assumption, if we can deduce properties of the system with $Z_0 = 1$, the ones with $Z_0$ generic will automatically follow. An example of growth population modelled with a Galton-Watson branching process can be seen in Fig. 1.9(a).

Given the number of offspring of an individual does not depend on its generation or others members of the population, the Galton-Watson branching process is characterised by the fact that its status at each generation depends on the past only through the number of members of the previous generation. This property will later be called Markov property (e.g. [93, Chapter 2]). This recursive structure enables computation of important quantities of the model, such as the Probability Generating Function (PGF) and the moments of $Z_n$, in an iterative way. This was noticed by Galton and Watson, who also discovered that the probability of extinction of the family was a fixed point of




(a) Galton-Watson (b) Bellman-Harris

Figure 1.9: **Galton-Watson and Bellman-Harris branching trees.** In the figure we provide a tree representation example of the growth of a population according to two models of branching processes. (a) When a population grows according to a Galton-Watson branching process, the cell inter-division times are disregarded and a study of the population among generations is conducted. Members of the same generation are coloured with the same tint. In the example showed, we have that $\{Z_0, Z_1, Z_2, Z_3, \ldots\} = \{1, 2, 4, 6, \ldots\}$, where $Z_n$ is the number of individuals in generation $n$. (b) When the lifetimes of the individuals are taken into account, and the dynamic of the population is modelled using a Bellman-Harris branching process, the family described in plot (a) can be represented in function of time. If we use the same colour system of the precedent panel to mark individuals in the same generation and assume that the lengths of the horizontal solid arrows are the lifetimes of the individuals, one of the possible situations is the one in the plot. In the example considered, if with $Z(t)$ we denote the size of the population at time $t$, we have $Z(t) = 5$. In view of the foregoing, the plot in (a) can also be regarded as the result of a Bellman-Harris branching growth with constant lifetimes. In this way the divisions of the members of the population occur simultaneously and at any time the individuals are always in the same generation.

the PGF of the offspring number distribution $N$, i.e. $\mathbb{P}\left(\lim_{n\to\infty} Z_n = 0\right)$ is solution of the equation $s = \mathbb{E}(s^N)$. Unfortunately, they didn't account for multiple solutions, and so incorrectly concluded that extinction was a certain fate. Subsequent work showed that the mean number offspring, i.e. $\mathbb{E}(N)$, determines the behaviour of the population at large generations: if $\mathbb{E}(N) > 1$ extinction or infinite growth are the only two possible events almost surely, both with strict positive probability, whereas if $\mathbb{E}(N) = 1$ or $\mathbb{E}(N) < 1$ the extinction of the population happens with probability one. In all three cases, called respectively supercritical, critical, and subcritical cases, we have that $\mathbb{E}(Z_n) = \mathbb{E}(N)^n$, where $\mathbb{E}(N)$ is the expected number of offspring after a cell division.

To reasonably model many phenomena, including those of cell division, it is necessary to move away from the synchronism that characterises the divisions of a Galton-Watson branching process. The American mathematicians Richard Ernest Bellman (1920-1984) and Theodore Edward Harris (1919-2004) were the first one to consider a continuous time structure inside the simple reproductive branching process [12], introducing a model that now is known by their names. Unlike the Galton-Watson model, in a basic Bellman-Harris process the individuals have random continuous valued lifetimes, independently of each other, but following a common probability distribution $L$, at the end of which they are substituted with their offspring. Thus, the size of the population can be studied at a given time $t \geq 0$ and denoted with $Z(t)$. The Markov property that characterises the Galton-Watson branching process is not generally retained in this generalisation necessitating a distinct approach for analysis. In their ground-breaking work, Bellman and Harris realised that techniques coming from Renewal Theory, an independent branch of probability that models the occurrence of events with random i.i.d. inter-arrival times, form the core of any analysis.

The asynchronism in the divisions brings also surprising important phenomenological consequences: the expected size of the population, $\mathbb{E}(Z(t))$, is greater than the expected size of a population with deterministic lifetimes of length $\mathbb{E}(L)$, i.e. $\mathbb{E}(Z(t)) \geq \mathbb{E}(N)^{\lfloor t/\mathbb{E}(L)\rfloor}$, with the equality only if $\mathrm{Var}(L) = 0$. For the same reasons, the doubling time of a population with $\mathbb{P}(N = 2) = 1$ is smaller than the expected cell lifetime, $\mathbb{E}(L)$ (e.g. [39, pg. 159]). But, similarly to the Galton-Watson branching process, the value of $\mathbb{E}(N)$ still determines

if the process is supercritical ($\mathbb{E}(N) > 1$), critical ($\mathbb{E}(N) = 1$), and subcritical ($\mathbb{E}(N) < 1$).

Even if the theory of branching process started in the social demographic context, its results had impact also outside that field. At the beginning of the 20th century branching processes were applied to genetics and biology more generally, with Ronald A. Fisher (1890-1962) and John B. S. Haldane (1892-1964) as major representatives. During World War II and the beginning of the nuclear age a huge drive in the developing of the theory arrived. At the same time Theodore E. Harris (1919-2005) and Boris A. Sevastyanov (1923-2013), American and Russian mathematicians respectively, were funded by military agencies of their respective countries to conduct research on branching processes trying to understand nuclear fission. In this period, formulas for the probability of extinction of a population played a role in the calculation of the critical mass of fissionable material needed for a sustained chain reaction. Indeed, even the proliferation of free neutrons in a nuclear fission reaction can be modelled using a branching process.

In this thesis, our focus is on the relationship in a closed population, i.e. one where immigration and emigration do not occur, between the population size and the sum of the generations of all living cells. As a result, we employ the original Bellman-Harris branching process framework where cells only give rise to offspring at the end of their lives, and the key quantities recorded are population number and, for our study, total generation. We sometimes start by describing the problem in a Galton-Watson setting in order to give intuition for the nature of the results we prove afterwards. For the applications we have in mind, the Bellman-Harris setting suffices. Furthermore, given the non-extinction of the population is the most interesting case for us, and given this is a fate made possible only in the supercritical case, i.e. $\mathbb{E}(N) > 1$, in the following chapters we work within this framework.

Since their introduction, however, branching processes have been subject to extensive mathematical study, and naturally arise in diverse applications from the life-sciences to queuing theory and beyond. Those studies have resulted in substantial generalisations of the framework that include, for example, populations with exogenous immigrants, populations where individuals can give rise to offspring during their lives, multi-type populations that consist of individu-

als of more than one type, each with distinct proliferation and differentiation parameterisations. Other important mathematical developments include the treatment of branching random walks and generalisations that allow the study of general functionals of the population [10, 55, 7, 56, 57, 39, 8]. In Chapter 3, some of the results proved for a Bellman-Harris branching process in Chapter 2 are extended to a one-directional two-type Bellman-Harris branching process, but we won't explore more general models than that.

## 1.4 A DNA coded randomised algorithm

The method that we are going to describe will motivate the study in this thesis. In this section we introduce it and describe benefits that make this technique a promising way to estimate the average generation of a cell population. The mathematical results that are at the base of this method will be improved and generalised throughout the thesis, strengthening the quality of the expected inference results.

In [116], Weber, Perié, and Duffy proposed a new design for in vivo inference of average generation that relies on a DNA coded randomised algorithm. For illustration, consider a single initial cell at time $t = 0$. As in Fig. 1.7(a), let $Z(t)$ be the number of offspring alive at time $t$ and $G(t)$ be the sum of the generations of all living cells at that time. The proposal to infer $G(t)/Z(t)$ was to equip the initial cell with a neutral label, i.e. one whose presence or absence has no ramifications for population dynamics, such that during each cell's lifetime, immediately prior to cell division, with a small probability $p$ the label is irrevocably and heritably lost (Fig. 1.10). Thus either all the offspring of a label-positive cell have the label, which occurs with probability $1 - p$, or all do not, which occurs with probability $p$. With $Z^+(t)$ denoting the number of label positive cells at time $t$, as in Fig. 1.7(b), the suggested estimator is

$$\frac{G(t)}{Z(t)} \approx -\frac{1}{p} \log\left(\frac{Z^+(t)}{Z(t)}\right). \tag{1.1}$$

This surprising formula is desirable for a number of reasons: 1) it allows for cell death; 2) it does not require knowledge of cell cycle times; and 3) for inference it requires only a proportional measurement rather than absolute numbers. Moreover, to infer the relative developmental depth of two populations whose ancestors are equipped with such neutral label, one does not need to know $p$,

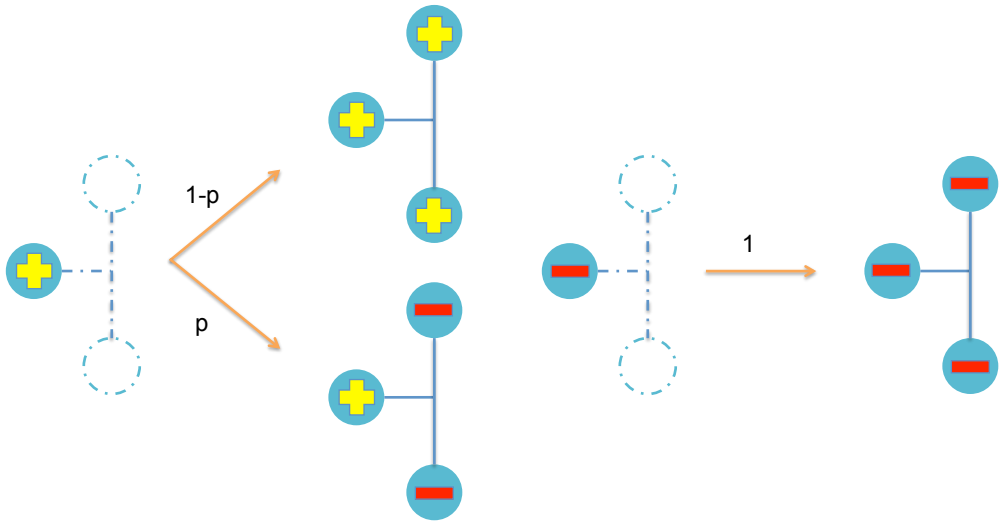



Figure 1.10: **Label functioning.** Weber, Perié, and Duffy described in [116] a method to estimate the average generation of a growing cell population based on the presence, in the initial pool of cells, of a neutral label that follows the dynamic showed in the figure. (Left side) Every cell that has the label (circles with yellow plus), given it divides, will give or not the label to the offspring with probability $1-p$ and $p$, respectively. (Right side) Cells that are not equipped with the label (circles with red minus) will always generate offspring without labels.

the probability of label loss per cell lifetime, if it is the same for both. A DNA coded randomised algorithm, based on the existing FUCCI cell cycle reporter [100], to realise the design is proposed in [116].

Two distinct derivations of the approximation (1.1) are provided in [116]. One, based on properties of cumulant generating functions, establishes that for a fixed time $t$ and an arbitrary lineage relationship between the cells constituting $Z(t)$, the expected number of label-positive cells, $\mathbb{E}(Z^+(t))$, over all possible delabellings recovers the correct value as the probability of label loss goes to 0:

$$\frac{G(t)}{Z(t)} = \lim_{p\to 0} -\frac{1}{p}\log\left(\frac{\mathbb{E}(Z^+(t))}{Z(t)}\right).$$

For a single realisation of the delabelling process, as would occur experimentally, however, this provides no assurance. To establish a result that does so, some structure is needed on the family tree. Consequently, a complementary result is also established in [116] within the context of the standard model of an asynchronously developing tree, the Bellman-Harris branching process.

That is, a growing tree model where cells have i.i.d. lifetimes and independent i.i.d. numbers of offspring numbers at the end of their lives [42, 63]. Despite the fact that Bellman-Harris branching processes do not perfectly describe the biological processes, mainly for the lack of independence among real cells, they provide a conceptually simple but powerful tool that allows quantitative predictions, beyond metaphorical representations. With $Z(t)$ being the number of cells alive at time $t$ in a super-critical Bellman-Harris branching process, for each $p \in (0,1)$ such that the label-positive sub-tree with $Z^+(t)$ cells living at time $t$ is super-critical, it is established in [116] that there exists a constant $\pi(p)$ such that

$$\lim_{t\to\infty} -\frac{1}{pt}\log\left(\frac{Z^+(t)}{Z(t)}\right) = \pi(p), \text{almost surely if } \lim_{t\to\infty}\inf Z^+(t) > 0 \tag{1.2}$$

and

$$\lim_{t\to\infty}\frac{\mathbb{E}(G(t))}{t\mathbb{E}(Z(t))} = \lim_{p\to 0}\pi(p). \tag{1.3}$$

The right hand side of equation (1.3), together with (1.2), says that as long as the label-positive sub-tree continues to exist, ultimately the estimate of average generation converges on each single path of the process. The left hand side, however, is not entirely satisfactory. It is an average quantity over realisations of the branching process and it forms the ratio of expectations, $\mathbb{E}(G(t))/\mathbb{E}(Z(t))$, rather than the expectation of the ratio $\mathbb{E}(G(t)/Z(t))$.

## 1.5 Contributions of this thesis and overview of results

The aim of this thesis is to provide additional mathematical support to the method that allows the estimation of the average generation of a cell population, proposed in [116] and briefly described in Section 1.4. The contributions of this thesis are mainly in three directions.

**Enhancement of the mathematical results in [116] in the context of an expanding cell population.** As in Weber, Perié, and Duffy [116] we model the growth of the population using a Bellman-Harris branching process [42, 63]. This allows us to see the population as a process developing in time, but also introduce some structure that is useful for predicting the expected behaviour of the moments of the total generation $G(t)$. With the help

of techniques from Renewal Theory [93], it is also possible to study its limit behaviour and growth rate. Under super-critical conditions, where large times assure big populations (if they survive), each population behaves similarly to its average path, leading us to conclude strong almost sure results. According to [116], the proportion of label positive cells of a population allows us to predict only the quantity $\mathbb{E}(G(t))/\mathbb{E}(Z(t))$, i.e. a sort of average behaviour of the average generation, but the only proof that this can work with the actual average generation assumes no relations of any type among the cells in the collection. Our results allow us to say that this is true under the Bellman-Harris branching process assumptions, where we can estimate the average generation of a cell population using only one realisation of the delabelling process. Assuming a population behaves as a super-critical Bellman-Harris branching process, consequences of this result are clear from an experimental point of view: for every population experiment, (1.1) allows us to estimate the real average generation of the considered population, not only an expected behaviour. We conduct this analysis in Chapter 2. More reassurances on the robustness of the estimation method proposed in [116] come from the analysis we make in Chapter 4. There, considering the special case of Bellman-Harris branching process known in the literature as Pure Birth process (cells divide into two offspring and have exponentially distributed lifetimes) we study how the variance of the average generation behaves in time. Studies on the generation of a random selected living cell [101] and the fact that $\mathbb{E}(G(t)/Z(t))$ grows linearly in time [116], would suggest that $\mathrm{Var}(G(t)/Z(t))$ grows to infinity too. We prove that this supposition is incorrect and that $\mathrm{Var}(G(t)/Z(t))$ converges to the constant 7 independently from the rate of division. This result is a consequence of the fact that $Z(t)$ and $G(t)$ are two processes strongly correlated at the level of sample paths.

**Enlargement of the range of applicability of the random delabelling average generation estimator.** In [116], limit theorems are only proved for an expanding single type Bellman-Harris branching process. Despite the interest associated with this case, even if a cell population is not subject to differentiation, we know that in normal conditions even cancer doesn't grow indefinitely [2]. For example in Fig. 1.4 we can see how during an infection, the T cell immune response is a combination of expansion, contraction and homeostasis. We extend the strong results found for the expansion phase of

a population to the latter of these cases, using a renewal process to model the population dynamics. This motivates the use of estimator (1.1) also for a homeostatic population, highlighting its potential in contexts outside the mere expansion and opening to more complex dynamics that interlace different growing phases in a particular time frame. However, as we have seen at the beginning of this introductory chapter, differentiation is something embedded in cellular development. For this reason a method that allows the estimation of the average generation of a population only when it doesn't change type could be extremely restrictive. We also overcome this problem, proving that the method still works for a two-type population. Fundamental will be in this context the result found by Jagers for two-type Bellman-Harris branching processes [53]. All together, these generalisations have an important impact from an experimental confidence point of view: even if a cell population is changing growth behaviour or is subject to differentiation during the experiment, the proportion of the label positive cell will still provide an estimate of its average generation, proving the robustness and flexibility of the method. This result opens the method to a wide range of applications, that until now were not formally supported. These generalisations can be found in Chapter 3 (two-type) and Chapter 6 (homeostasis).

**Finding the best parameterisation.** The results described until now, including the ones in [116], are dependent on the probability of label loss $p$ being "close" to 0. However, a careful examination of the optimal $p$ is warranted. When we work with finite times, as biologists do when they question the cell populations, the parameterisation plays an important role. For example, when the $p$ is too small, there is the risk that no cell loses the label in the time frame considered, implicating the impossibility of an estimation of the average generation. But, if we try to use a "large" parameter $p$ to avoid this scenario, we take the risk to completely lose the label from the population or, when this is not happening, that the estimator does not describe the quantity wanted. Our third major contribution concerns this point: using 3 different models of increasing complexity to describe the growth of a cell population, included the Bellman-Harris branching process, we try to understand if the requirement of working with a $p$ small is always necessary. When this is the case, we give a suggestion on how the parameter $p$ should scale with the size of the population in order to maximise the performance of the estimator and have the probabil-

ity of not getting any estimate below a certain threshold. This study, that we conduct in Chapter 5, resorts to the use of several approximations, which find justification in the two following working environments: we either increase the size of the initial sample, unchanging the behaviour of the average generation, or we let develop the population for large time frames so that, if it survives, it increases its size along with the average generation. Both frameworks allow to smoothen the behaviour of the population by increasing the number of cells in it and make the approximations reasonable. Despite that the two settings are non-commensurable and we choose one or the other depending of the model.

Throughout the thesis we provide simulations that illustrate the mathematical results found. For concreteness, we use as example a model of population dynamics of B cells during an adaptive immune response. B cells exhibit a nice and clear change of conduct before, during, and after an infection is found (Fig. 1.4). The parameterisation that we use is obtained from [44], where B cells are activated by CpG DNA and studied via time lapse microscopy. We don't consider intra-family correlation and heterogeneity in lifetime distribution among generations as made in [44], but we take from that the class of lifetime distribution (lognormal) of B cells and a parameterisation of that. In our study we also sometimes provide examples for lifetimes that independently follow an exponential distribution, one of the assumptions commonly made in the field for mathematical convenience even though no experimentally measured cell system has been reported that is consistent with this hypothesis.

CHAPTER 2

# Average generation of a super-critical Bellman-Harris process

## 2.1 Introduction

In Section 1.4, we have described a technique for the estimation of the average generation of a cell population based on a DNA coded randomised algorithm, recently proposed by Weber, Perié, and Duffy in [116]. In this chapter we want to improve the mathematical results used to justify that method, increasing in this way its potentialities. Before doing that, we describe and formalise the problem, using the same notation introduced in Chapter 1.

The object of our study is a cell population where each cell, after a random lifetime $L$, gives birth to a random number, $N$, of offspring upon death. The lifetimes and the number of offspring of each cell are assumed to be independent of each other and also i.i.d. with the rest of the cells in the population. Under these assumptions the dynamics of the population size is described by a Bellman-Harris branching process [42, 63, 10]. Within this chapter, we suppose that the process is super-critical, i.e. the expected number of offspring after each division $h := \mathbb{E}(N)$ is greater than 1, a condition that shows the attitude of the population to expand increasing its size. For each descendant,

we define generation as the number of divisions that led to that cell, assigning generation 0 to cells which constitute the initial population. We denote with $G(t)$ the sum of the generations of the living cells at time $t$, such that $G(0) = 0$ represents the initial setting (Fig. 1.5). In order for the estimator of the average generation in Section 1.4 to work, each cell in the initial population has to be equipped with a neutral label, i.e. a label that does not alter the dynamic of the cell. This label, independently from the lifetime of the cell, can be lost right before the division of the cell with a probability $p$, otherwise, it is inherited by its offspring and same rules will apply to them. Let's denote with $Z(t)$ and $Z^+(t)$ the size of the entire and of the label-positive populations at time $t$, respectively.

According to [116], for each $p \in (0, 1)$ such that the label-positive sub-tree with $Z^+(t)$ cells living at time $t$ is a super-critical Bellman-Harris branching process, we have almost surely that

$$\lim_{t\to\infty} -\frac{1}{pt} \log\left(\frac{Z^+(t)}{Z(t)}\right) = \pi(p), \quad \text{if} \liminf_{t\to\infty} Z^+(t) > 0 \tag{2.1}$$

and

$$\lim_{t\to\infty} \frac{\mathbb{E}(G(t))}{t\mathbb{E}(Z(t))} = \bar{\pi} = \lim_{p\to 0} \pi(p), \tag{2.2}$$

for a constant $\bar{\pi} > 0$. As highlighted in Section 1.4, these results say that a function of the proportion of label positive cells and a quantity somewhat similar to the average generation of the population converge to the same quantity. Furthermore, (2.1) is a strong result because it says that it is true for every single realisation of the delabelling process, a desired property from an experimental point of view. Instead, what is not completely satisfying is the Left-Hand Side (LHS) of (2.2), where in lieu of $\mathbb{E}(G(t))/\mathbb{E}(Z(t))$, we would like to have at least the expected average generation $\mathbb{E}(G(t)/Z(t))$.

In the present chapter we rectify this shortcoming by proving a substantially stronger result: that for a Bellman-Harris branching process the average generation divided by time converges almost surely to $\bar{\pi} > 0$, i.e.

$$\lim_{t\to\infty} \frac{G(t)}{tZ(t)} = \bar{\pi} = \lim_{p\to 0} \pi(p), \quad \text{if } \liminf_{t\to\infty} Z(t) > 0. \tag{2.3}$$

The result in (2.3) greatly strengthens the only previous result we are aware of, that proved in [101] where convergence in probability of average generation

is established for processes in which there is no death. Given the ubiquity of Bellman-Harris processes, it is likely to be of interest for other reasons, but for our purposes it is most significant in providing extra support for merits of the proposed average generation inference methodology.

In order to establish this fact we prove a collection of surprising results for the paired processes $(Z(t), G(t))$ of a super-critical Bellman-Harris process. In particular, with $L$ being a lifetime distribution, $h > 1$ being the average number of offspring of a cell at the end of its life and $\alpha$ being the Malthusian parameter, i.e. the solution to

$$h\mathbb{E}(e^{-\alpha L}) = 1, \tag{2.4}$$

then

$$\lim_{t\to\infty} \left(e^{-\alpha t} Z(t), t^{-1} e^{-\alpha t} G(t)\right) = (c_1 \mathcal{Z}, c_2 \mathcal{Z}), \tag{2.5}$$

where $\mathcal{Z}$ is a random variable and $c_1, c_2$ are constants. Namely, even though the total generation advances at a different rate to the population size, the random element of the prefactor is the same for both, and properties of the ratio $G(t)/Z(t)$ follow.

To establish those results we use a combination of both old and novel arguments, essentially following the methodology described by Harris [42], but relying on a peculiar renewal theorem, inspired by results of Asmussen [6], for what is known as defective probability measures, which are measures whose total mass is smaller than 1 [93, Chapter 3]. The Malthusian parameter can be thought of as determining an exponential tilt that identifies a measure with density $h\exp(-\alpha t)dP(L \leq t)$. That is a probability measure as it integrates to 1 thanks to equation (2.4). Defective probability measures, however, naturally arise in the study of the higher moments of branching processes as one encounters renewal equations with more extreme exponential tilts, $\exp(-k\alpha t)$ for $k > 1$, resulting in measures that integrate to less than 1. The new results allow us to obtain an integral formulation for the probability generating functions of the prefactors described above. To clinch the result, we essentially insert the guess that the randomness in the prefactors of the two processes is the same.

## 2.2 Motivation for the mathematical result

A time-dependent model of a family tree is necessary to investigate the temporal dyamics of average generation. Analysis is trivial in the simplest such stochastic model, the Galton-Watson branching process [115, 42, 63]. It assumes that all cells of a given generation share a common lifetime at the end of which they produce i.i.d. numbers of offspring for the next generation. If $t_n$ is the time of birth of the $n^{\text{th}}$ generation, then the total generation is simply $G(t_n) = nZ(t_n)$. Consequently, the well known result for the limit behaviour of $Z(t_n)$ as $n$ becomes large in the super-critical case [42, Chapter 1] also describes the prefactor on front of the distribution of $G(t_n)$,

$$\lim_{n\to\infty} \frac{Z(t_n)}{h^n} = \mathcal{Z} \implies \lim_{n\to\infty} \frac{G(t_n)}{nh^n} = \mathcal{Z} \tag{2.6}$$

where $h > 1$ is the average number of offspring, $\mathcal{Z}$ is a non-negative random variable such that $\mathbb{E}(\mathcal{Z}) = 1$, and the equalities in (2.6) are meant in distribution.

On relaxing the constraint that all lifetimes are equal, however, there seems to be little *a priori* reason why the analogous quantity to $\mathcal{Z}$ in (2.6), which is $\mathcal{Z}$ in (2.5), should be shared by both $Z(t)$ and $G(t)$. Moving away from synchronicity, if the lifetimes of cells are i.i.d. positive and non-lattice random variables, the development forms a Bellman-Harris branching process [42, 63]. In that setting, cells are spread across generations and the ratio $G(t)/Z(t)$ is no longer deterministic. As $\mathbb{E}(G(t))/(t\mathbb{E}(Z(t)))$ converges to a constant [116], it is reasonable to suspect that the average generation will still grow linearly in time. That possibility is also suggested by Fig. 2.1, where, for independent simulations of a super-critical Bellman-Harris process with Malthusian parameter $\alpha$ defined in (2.4), $Z(t)e^{-\alpha t}$ and $G(t)e^{-\alpha t}$ are plotted, illustrating the factor $t$ in the ratio between them.

Collating observations across multiple simulations, however, Fig. 2.2 suggests something analogous to (2.6) is taking place. Fig. 2.2(a) plots the empirical cumulative distribution function of the renormalised total cell numbers and total generation at a large time, suggesting equality in distribution. Fig. 2.2(b) displays a scatter plot of the per-simulation prefactors of those quantities for large $t$. There is a strong positive correlation in these values, hinting at their relatedness. Finally Fig. 2.2(c) shows sample paths of the difference between

the renormalised total cell numbers less renormalised total generation, which appears to be converging to zero. This further suggests convergence in probability of the sample-path average generation of a Bellman-Harris process, conditional on survival. Thus, even though $G(t)/Z(t)$ is no longer deterministic, the randomness in $G(t)/Z(t)$ does not reside in the linear term, but in something smaller, which is one result that is formally established in this chapter.

## 2.3 Total generation convergence in a super-critical B-H branching process

### 2.3.1 Assumptions, notation and previous results

The following notation and assumptions are in force throughout Section 2.3. We work within a Bellman-Harris branching processes with strictly positive non-lattice lifetime random variable $L$ and non-negative offspring random variable $N$. We define $h := \mathbb{E}(N)$ and $v := \mathbb{E}(N(N-1))$, and assume that both are finite. We work within the super-critical case, $h > 1$, so that the population has a positive probability of escaping extinction [42].

We will make use of the Malthusian parameter $\alpha$ defined in (2.4). As $h > 1$, $\alpha > 0$ exists and is unique. For $h > 1$, it is established in Proposition 1 of [116] that the Malthusian parameter $\alpha$ is a real analytic function of $h$. For our purposes, we don't need to consider $\alpha$ as a function of $h$, but we will sometimes use the notation $\alpha'$ to indicate the value $d\alpha(x)/dx|_{x=h}$. To study the limit behaviour of scaled versions of the process $(Z(t), G(t))$ we use standard notions of convergence in distribution (D), in mean square ($L^2$), and almost surely (a.s.) [97, 30]. Convolution between functions will be denoted by the operator $*$. Occasionally in the text we will refer to the underlying measurable space or the probability space, which we denote respectively as $(\Omega, \mathcal{B}(\Omega))$ and $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$, with $\mathcal{B}(\Omega)$ Borel $\sigma$-algebra of $\Omega$. Example constructions of such spaces can be found in [42, Chapter VI.2].

A brief summary of known results concerning $Z(t)$ and $G(t)$ will follow. According to [42, 54], under the above assumptions, the limit behaviour of $Z(t)$ satisfies

$$\frac{Z(t)}{e^{\alpha t}} \xrightarrow{a.s., L^2} c\mathcal{Z}, \tag{2.7}$$

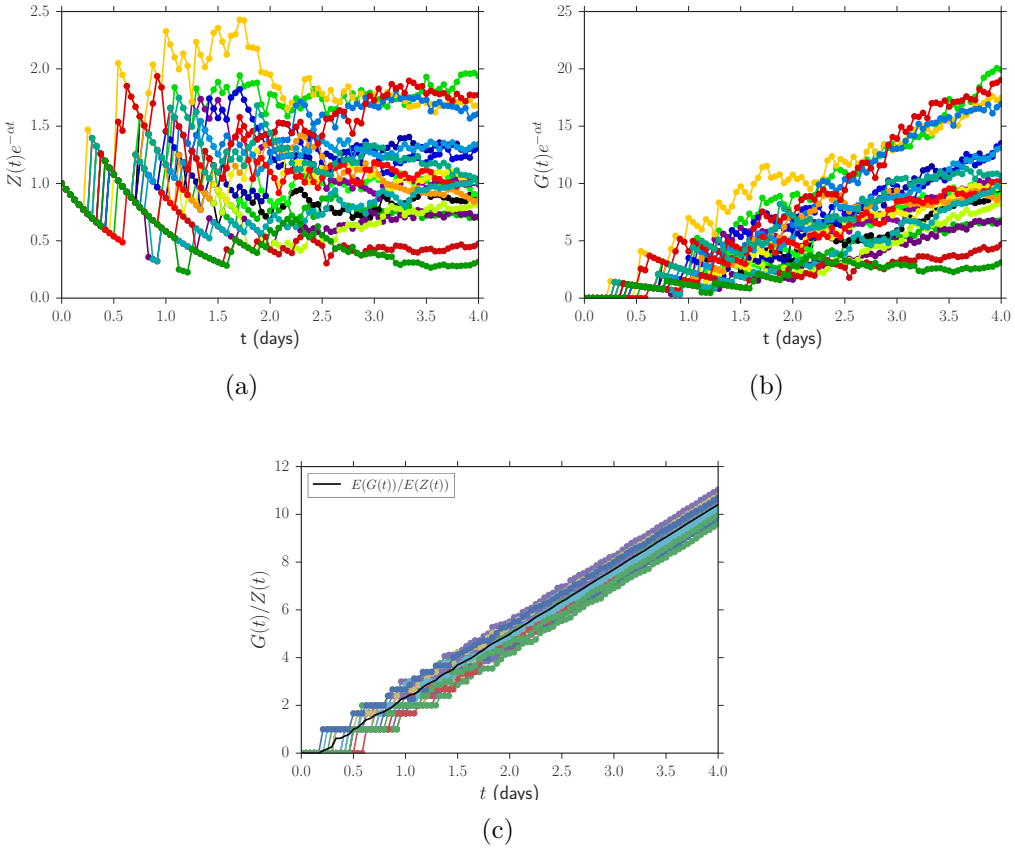



Figure 2.1: **Growth rates of population size, $Z(t)$, and total generation, $G(t)$ of a super-critical Bellman-Harris process.** Each plots present 20 Monte Carlo simulations of a Bellman-Harris branching process starting at $t = 0$ with a single cell, where paths are conditioned to have living cells at the final time-point of the simulation. Lifetimes are lognormal with mean 9.3 hours and standard deviation 2.54, which coincide with those measured for murine B cells stimulated *in vitro* with CpG DNA [44]. At the end of each cell's life it gives rise to no cells with probability 1/5 and two with probability 4/5. (a) With $Z(t)$ being the population size at time $t$ and $\alpha > 0$ being the Malthusian parameter defined in equation (2.4), this figure plots the evolution of $Z(t)/e^{\alpha t}$, which is known to converge almost surely and in mean square to a random variable $A$, e.g. [42]. (b) With $G(t)$ denoting the total generation of the process (see Fig. 1.7) at time $t$, for the same paths this plot shows $G(t)/e^{\alpha t}$, which grows linearly over time with a random slope $B$. Results in Section 2.3.4 establish that $A$ and $B$ are almost surely the same, up to a multiplicative constant, on a path-by-path basis. Thus the average generation process, $G(t)/Z(t)$, grows linearly in time, but with the same slope for every path. This can be seen empirically in panel (c) where 20 instances of this process are plotted (solid lines with markers), as well as $\mathbb{E}(G(t))/\mathbb{E}(Z(t))$ (solid black line).

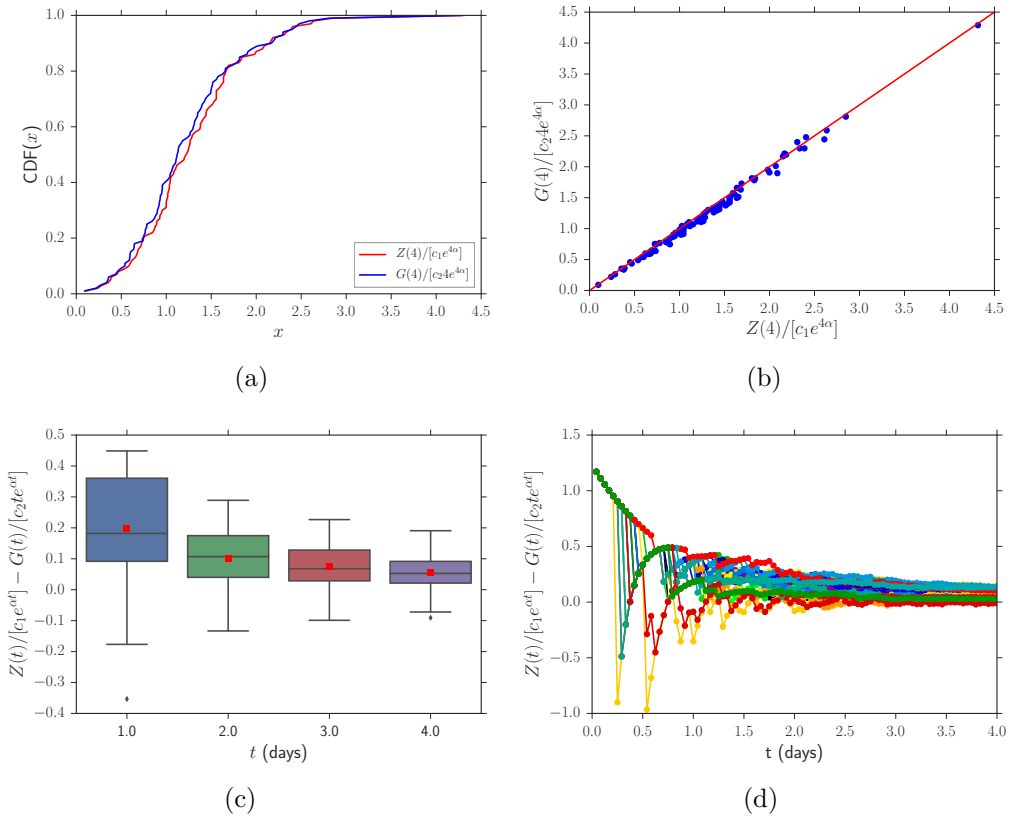



Figure 2.2: **Comparison between simulations of $Z(t)/e^{\alpha t}$ and $G(t)/te^{\alpha t}$.** These show results from 100 Monte Carlo simulations of a Bellman-Harris process with paramaterization as in Fig. 2.1. (a) At $t = 4$ days, empirical cumulative distribution function (eCDF) of $Z(t)/(c_1e^{\alpha t})$ and $G(t)/(c_2te^{\alpha t})$ are shown, where $c_1$ and $c_2$ are constants that normalise the limit behaviour of means of the two processes and are computed numerically. The eCDFs of the prefactor on the population size and the slope of the total generation process are similar suggesting that they follow the same distribution. (b) Also at $t = 4$ days, the scatter plot of $Z(t)/(c_1e^{\alpha t})$ versus $G(t)/(c_2te^{\alpha t})$ for the same path suggests a stronger result, that there is equality almost surely. This impression is further informed by plot (c) where a boxplot of the distribution of $Z(t)/(c_1e^{\alpha t}) - G(t)/(c_2te^{\alpha t})$ is shown after each day for 4 days. Here it seems the difference between the two random variables is shrinking to 0, feeling confirmed by plot (d) where 20 paths describing the same quantity are shown.

where $\mathcal{Z}$ is a non-negative random variable such that $\mathbb{E}(\mathcal{Z}) = 1$, and

$$c = \lim_{t\to\infty} \frac{\mathbb{E}(Z(t))}{e^{\alpha t}} = \frac{\int_0^\infty \mathbb{P}(L > t)e^{-\alpha t}dt}{h\int_0^\infty ue^{-\alpha u}d\mathbb{P}(L \le u)} = \frac{h-1}{h^2\alpha\int_0^\infty ue^{-\alpha u}d\mathbb{P}(L \le u)}.$$

For the expected value of $G(t)$, the following is proven in Theorem 2 of [116]

$$\lim_{t\to\infty} \frac{\mathbb{E}(G(t))}{te^{\alpha t}} = h\alpha' c, \quad \text{where } \alpha' = \frac{1}{h^2\int_0^{+\infty} ue^{-\alpha u}d\mathbb{P}(L \le u)}. \tag{2.8}$$

There, we find also information concerning the asymptotic covariance of $G(t)$ and $Z(t)$ and the ratio of their expectations,

$$\lim_{t\to\infty} \frac{\mathbb{E}(G(t)Z(t))}{te^{2\alpha t}} = c^2 h\alpha' k \quad \text{and} \quad \lim_{t\to\infty} \frac{\mathbb{E}(G(t))}{t\mathbb{E}(Z(t))} = h\alpha',$$

where

$$k = \frac{v\int_0^\infty e^{-2\alpha u}d\mathbb{P}(L \le u)}{1 - h\int_0^\infty e^{-2\alpha u}d\mathbb{P}(L \le u)} \tag{2.9}$$

and $v = \mathbb{E}(N(N-1))$. The scaling of means in equations (2.7) and (2.8) suggests the definition of normalised versions of the processes $Z(t)$ and $G(t)$,

$$\mathcal{Z}_t := \frac{Z(t)}{ce^{\alpha t}} \quad \text{and} \quad \mathcal{G}_t := \frac{G(t)}{ch\alpha' te^{\alpha t}}, \tag{2.10}$$

whose use will simplify notation in the proofs.

In order to establish the main result, equation (2.3), stated in Corollary 2.3.12 of Section 2.3.5, we study the limit behaviour of the process $\{\mathcal{G}_t\}$. We do that in two steps: first, in Section 2.3.3 we consider $\{\mathcal{G}_t\}$ as an $L^2$ process and determine its mean square limit; then, in Section 2.3.5 we reinforce that result by proving that the convergence is also valid with probability 1 under a condition on the speed of $L^2$ convergence. In Section 2.3.3, we make extensive use of a particular version of Key Renewal Theorem for defective measures that we establish in Section 2.3.2. Once we prove in Section 2.3.4 that $\mathcal{G}_t$ and $\mathcal{Z}_t$ share the same random prefactor on front of their dominant term for large $t$, we are finally able to characterise the limit behaviour of $G(t)/(tZ(t))$.

### 2.3.2 A new Renewal Theorem for Defective Measures

In order to prove (2.8) in [116], a version of the Renewal Theorem due to Asmussen, Theorem 6.2(b) of [6], is used in a fundamental way. In this section we generalise that theorem to make it applicable for defective measures, i.e.

measures with total mass less than one. Before going to the main result of the section, Theorem 2.3.3, we first state a non-standard version of the classical Dominated Convergence Theorem (DCT), which can be applied to a collection of sequences of functions $\{(f_{t,\tau})_{t\in\mathbb{R}_{\geq 0}} : \tau \in \mathbb{R}_{\geq 0}\}$, each one converging pointwise, when $t \to \infty$, to a same function $f$, uniformly for $\tau \geq 0$. This can be proved essentially repeating the same steps of the classical DCT, including the use of Fatou's lemma, but this time the hypothesis of the uniformity in $\tau$ allows a stronger conclusion. This proposition is followed by a lemma that depends on it.

**Proposition 2.3.1** (Non-standard DCT)**.** *Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ be a measure space, and for every $\tau \geq 0$ let $(f_{t,\tau})_{t\geq 0}$ be a sequence of functions in $L^1(\mu)$ that converges pointwise to $f$ uniformly for $\tau \in [0, \infty)$, i.e. given $\epsilon > 0$ and $u \in \mathbb{R}$ there exists a $t_{\epsilon,u} > 0$ s.t. for every $t \geq t_{\epsilon,u}$ and $\tau \geq 0$ we have $|f_{t,\tau}(u) - f(u)| < \epsilon$. Assume there is $g \in L^1(\mu)$ s.t. $|f_{t,\tau}(u)| \leq g(u)$ for every $t, \tau$, and $u$. Then, $f \in L^1(\mu)$ and*

$$\lim_{t\to\infty} \int_{\mathbb{R}} f_{t,\tau}(u)d\mu(u) = \int_{\mathbb{R}} f(u)d\mu(u) \qquad \text{uniformly for } \tau \geq 0,$$

*i.e. given $\epsilon > 0$ there exists a $t^*_\epsilon > 0$ s.t. for every $t \geq t^*_\epsilon$ and $\tau \geq 0$ we have $|\int_{\mathbb{R}} f_{t,\tau}(u) - f(u)d\mu(u)| < \epsilon$.*

We are now going to use this version of the DCT to study the limit behaviour of convolutions between functions and probability measures. We are interested in these particular structures because we will show that the moments of $G(t)$ can be written in that form.

**Lemma 2.3.2** (Convolution with a finite measure doesn't change convergence rates)**.** *Consider $f = f(t, \tau) : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ locally bounded in $t$ and s.t., for every $\tau \geq 0$, $f(t, \tau)/[t^p(t + \tau)^q] \to c_1$ when $t \to \infty$, with $c_1 < \infty$, $p, q \geq 0$, and let $\mu$ be a finite measure on $(\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}))$. Then, for every $\tau \geq 0$*

$$\lim_{t\to\infty} \frac{1}{t^p(t + \tau)^q} \int_0^t f(t - u, \tau)\mu(du) = c_1\mu([0, \infty)). \tag{2.11}$$

*Furthermore, if $|f(t, \tau)| \leq f_1(t)f_2(t + \tau)$, with $f_i(s) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ positive locally bounded functions for $i \in \{1, 2\}$, $f_1(s)/s^p \to a_1$, $f_2(s)/s^q \to a_2$, and $f(t, \tau)/[t^p(t+\tau)^q] \xrightarrow[t\to\infty]{} c_1$ uniformly for $\tau \geq 0$ with $a_1, a_2, c_1 < \infty$ and $p, q \geq 0$, then (2.11) is true uniformly for $\tau \in [0, \infty)$.*

**Proof:** We only prove the second part of the lemma, as the first part follows from the same rationale with the use of the classical Dominated Convergence Theorem instead of Proposition 2.3.1.

For the following, we extend the functions $f$, $f_1$, and $f_2$ to $\mathbb{R}\times\mathbb{R}_{\geq 0}$, $\mathbb{R}$, and $\mathbb{R}$, respectively, by defining $f(t,\tau)=f_1(t)=f_2(t)=0$ when $t<0$. If we can establish that $(|f(t-u,\tau)|/[t^p(t+\tau)^q])\mathbb{1}_{[0,t)}(u)$ is bounded by a constant $M$, for every $u\in\mathbb{R}$, $\tau\geq 0$, and $t$ sufficiently large, we can apply the DCT in Proposition 2.3.1 and conclude that equation (2.11) holds uniformly for $\tau\in[0,\infty)$.

Given $\epsilon>0$, from the hypotheses made, we know that there exists $u_\epsilon>0$ s.t. for every $u\geq u_\epsilon$ we have $f_1(u)/u^p\leq a_1+\epsilon$ and $f_2(u)/u^q\leq a_2+\epsilon$. Without loss of generality we can suppose $t\geq t_{\epsilon,u}:=\max\{u_\epsilon,1\}$. So, for every $u\in\mathbb{R}$, we have

$$\begin{aligned}0\leq g_t(u):=\frac{f_1(u)}{t^p}\mathbb{1}_{[0,t)}(u)&=\frac{f_1(u)}{t^p}\mathbb{1}_{[0,u_\epsilon)}(u)+\frac{f_1(u)}{t^p}\mathbb{1}_{[u_\epsilon,t)}(u)\\&\leq f_1(u)\mathbb{1}_{[0,u_\epsilon)}(u)+\frac{f_1(u)}{u^p}\mathbb{1}_{[u_\epsilon,\infty)}(u)\\&\leq \sup_{[0,u_\epsilon)} f_1(u)+a_1+\epsilon=M_1<\infty,\end{aligned}\tag{2.12}$$

where in the last equality we have used the fact that $f_1$ is a locally bounded function. From (2.12), we have that $g_t(u)$ is dominated by $M_1$ for every $u\in\mathbb{R}$ and $t\geq t_{\epsilon,u}$. So, the same will be true for $g_t(-u)$, and for its translation $g_t(t-u)$. Similar reasoning applies to $f_2$, obtaining

$$\frac{f_1(t-u)}{t^p}\mathbb{1}_{[0,t)}(u)\leq M_1,\qquad \frac{f_2(t-u)}{t^q}\mathbb{1}_{[0,t)}(u)\leq M_2,$$

for every $u\in\mathbb{R}$ and $t\geq t_{\epsilon,u}$. Remembering that by hypothesis $|f(t,\tau)|\leq f_1(t)f_2(t+\tau)$, for every $u\in\mathbb{R}$, $t\geq t_{\epsilon,u}$, and $\tau\geq 0$ we have

$$\begin{aligned}\frac{|f(t-u,\tau)|}{t^p(t+\tau)^q}\mathbb{1}_{[0,t)}(u)&\leq\frac{f_1(t-u)}{t^p}\mathbb{1}_{[0,t)}(u)\frac{f_2(t+\tau-u)}{(t+\tau)^q}\mathbb{1}_{[0,t+\tau)}(u)\\&\leq M_1M_2=:M\end{aligned}$$

That concludes the proof. □

Armed with that Lemma, we can now prove the main result of this section.

**Theorem 2.3.3** (A defective measure version of Theorem 6.2(b) [6])**.** *Consider the integral equation*

$$K(t,\tau) = f(t,\tau) + \int_0^t K(t-u,\tau)\rho(du), \tag{2.13}$$

*where* $K, f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$*, and* $\rho$ *is a positive defective measure on* $(\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}))$*, i.e.* $\rho([0,\infty)) < 1$*. If* $f(t,\tau)$ *is locally bounded in* $t$ *and s.t.* $f(t,\tau)/[t^p(t+\tau)^q] \to c_1$ *when* $t \to \infty$*, with* $c_1 < \infty$*,* $p, q \geq 0$*, then for every* $\tau \geq 0$

$$\lim_{t\to\infty} \frac{K(t,\tau)}{t^p(t+\tau)^q} = \frac{c_1}{1-\rho([0,\infty))}. \tag{2.14}$$

*Furthermore, if* $f$ *is s. t.* $|f(t,\tau)| \leq f_1(t)f_2(t+\tau)$*, with* $f_i(s) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *locally bounded functions,* $i \in \{1,2\}$*, s.t.* $f_1(s)/s^p \to a_1$*,* $f_2(s)/s^q \to a_2$*, and* $f(t,\tau)/[t^p(t+\tau)^q] \xrightarrow[t\to\infty]{} c_1$ *uniformly for* $\tau \geq 0$ *with* $a_1, a_2, c_1 < \infty$ *and* $p, q \geq 0$*, then* (2.14) *is true uniformly for* $\tau \geq 0$*.*

**Proof:** From [93, Theorem 3.5.1], the only solution of (2.13) that is bounded on every finite interval of $t$ has the form

$$K(t,\tau) = (U * f)_\tau(t) = \int_0^t f(t-u,\tau)U(du), \tag{2.15}$$

where we have $U([0,t)) = \sum_{n=0}^{\infty} \rho^{*n}([0,t))$, $\rho^{*n}([0,t)) = (\rho * \rho^{*(n-1)})([0,t))$, and $\rho^{*0}([0,t)) = \mathbb{1}_{[0,\infty)}(t)$. Using Lemma 2.3.2 and the fact that $U([0,\infty)) = 1/(1-\rho([0,\infty)))$ [93, Section 3.11], we obtain (2.14). □

Thanks to the linearity of integration, we have the following mild generalisation.

**Corollary 2.3.4.** *If in Theorem 2.3.3 we substitute the condition* $|f(t,\tau)| \leq f_1(t)f_2(t+\tau)$ *with* $|f(t,\tau)| \leq \sum_{i=1}^{n} f_{2i-1}(t)f_{2i}(t+\tau)$*, where* $f_i$ *are locally bounded functions s.t.* $f_{2i-1}(t)/t^p \to a_{2i-1}$*,* $f_{2i}(t)/t^q \to a_{2i}$*,* $a_{2i-1}, a_{2i} < \infty$ *for every* $1 \leq i \leq n$*, then the conclusions of Theorem 2.3.3 hold.*

### 2.3.3 Mean square convergence

Equation (2.8) states that $\mathbb{E}(\mathcal{G}_t) \to 1$, with $\mathcal{G}_t$ defined in (2.10). A natural question that this result raises is whether there exists a non-negative random variable $\mathcal{G}$, s.t. $\mathbb{E}(\mathcal{G}) = 1$, to which $\mathcal{G}_t$ converges in mean. Studying the behaviour of the second moment of $\mathcal{G}_t$, in Theorem 2.3.9, the main result of the

section, we will prove something stronger than that: the convergence is true also in $L^2$. To achieve that we will need a version, stated in Proposition 2.3.5, of one of the results presented in [116] concerning the Probability Generating Function (PGF) of $(G(t), Z(t))$, that better fits our purpose. We use it in Lemmas 2.3.6 and 2.3.7 where a study of the covariance between $G(t)$ and $Z(t)$, and of the relation between different terms of the total generation process is made. This will lead us to Corollary 2.3.8, which allows us to finally prove Theorem 2.3.9.

**Proposition 2.3.5** (A reformulation of Theorem 2 of [116])**.** *For $s_1, s_2, r_1, r_2, t, \tau \in \mathbb{R}_{\geq 0}$, define $F(s_1, s_2, r_1, r_2, t, \tau) := \mathbb{E}(s_1^{G(t)} s_2^{G(t+\tau)} r_1^{Z(t)} r_2^{Z(t+\tau)})$. Then, we have*

$$\begin{aligned} F(s_1, s_2, r_1, r_2, t, \tau) &= r_1 r_2 \mathbb{P}(L > t + \tau) \\ &\quad + r_1 \int_t^{t+\tau} \rho_N \left( \mathbb{E}\left( s_2^{G(t+\tau-u)} (s_2 r_2)^{Z(t+\tau-u)} \right) \right) d\mathbb{P}(L \leq u) \\ &\quad + \int_0^t \rho_N \Big( F(s_1, s_2, s_1 r_1, s_2 r_2, t-u, \tau) \Big) d\mathbb{P}(L \leq u), \quad (2.16) \end{aligned}$$

*where $\rho_N(s) = \mathbb{E}(s^N)$, the probability generating function of the offspring number, $N$.*

Using Proposition 2.3.5, we analyse the limiting behaviour of the covariance between $Z(t)$ and $G(t)$.

**Lemma 2.3.6** (Limit behaviour of the covariance of $\mathcal{G}_t$ and $\mathcal{Z}_t$)**.** *Using the previous notation, we have*

$$\lim_{t\to\infty} \mathbb{E}(\mathcal{G}_t \mathcal{Z}_{t+\tau}) = k = \lim_{t\to\infty} \mathbb{E}(\mathcal{G}_{t+\tau} \mathcal{Z}_t) \qquad \text{uniformly in } \tau \geq 0, \qquad (2.17)$$

*where $k$ is defined in* (2.9)*.*

**Proof:** We prove only the first of the equalities in (2.17) as the other can be obtained in a similar way.

Consider the integral equation (2.16) and take the derivative first for $s_1$, sec-

ondly for $r_2$, and then evaluate it at $(1,1,1,1,t,\tau)$. We obtain that

$$\begin{aligned}\mathbb{E}(G(t)Z(t+\tau)) =& v\int_0^t \mathbb{E}(G(t-u))\mathbb{E}(Z(t+\tau-u))d\mathbb{P}(L\leq u)\\ &+ v\int_0^t \mathbb{E}(Z(t-u))\mathbb{E}(Z(t+\tau-u))d\mathbb{P}(L\leq u)\\ &+ h\int_0^t \mathbb{E}(Z(t-u)Z(t+\tau-u))d\mathbb{P}(L\leq u)\\ &+ h\int_0^t \mathbb{E}(G(t-u)Z(t+\tau-u))d\mathbb{P}(L\leq u), \end{aligned} \tag{2.18}$$

where we recall that $h=\mathbb{E}(N)$ and $v=\mathbb{E}(N(N-1))$. Multiplying both sides of this equation by $e^{-\alpha t}e^{-\alpha(t+\tau)}$ and denoting

$$K(t,\tau) := \frac{\mathbb{E}(G(t)Z(t+\tau))}{e^{\alpha t}e^{\alpha(t+\tau)}},$$

$$d\overline{\mathbb{P}}(L\leq u) := he^{-2\alpha u}d\mathbb{P}(L\leq u), \qquad d\mathbb{P}'(L\leq u) := ve^{-2\alpha u}d\mathbb{P}(L\leq u),$$

$$\begin{aligned} f(t,\tau) :=& \int_0^t \frac{\mathbb{E}(G(t-u))}{e^{\alpha(t-u)}}\frac{\mathbb{E}(Z(t+\tau-u))}{e^{\alpha(t+\tau-u)}}d\mathbb{P}'(L\leq u)\\ &+ \int_0^t \frac{\mathbb{E}(Z(t-u))}{e^{\alpha(t-u)}}\frac{\mathbb{E}(Z(t+\tau-u))}{e^{\alpha(t+\tau-u)}}d\mathbb{P}'(L\leq u)\\ &+ \int_0^t \frac{\mathbb{E}(Z(t-u)Z(t+\tau-u))}{e^{\alpha(t-u)}e^{\alpha(t+\tau-u)}}d\overline{\mathbb{P}}(L\leq u) \end{aligned} \tag{2.19}$$

we have that

$$K(t,\tau) = f(t,\tau) + \int_0^t K(t-u,\tau)d\overline{\mathbb{P}}(L\leq u). \tag{2.20}$$

Observe that $\overline{\mathbb{P}}$ is a defective measure. In fact,

$$\int_0^{+\infty} d\overline{\mathbb{P}}(L\leq u) = h\int_0^{+\infty} e^{-2\alpha u}d\mathbb{P}(L\leq u) < h\int_0^{+\infty} e^{-\alpha u}d\mathbb{P}(L\leq u) \overset{(2.4)}{=} 1. \tag{2.21}$$

As $\mathbb{E}(\mathcal{G}_t\mathcal{Z}_{t+\tau}) = \mathbb{E}(G(t)Z(t+\tau))/[h\alpha' c^2 t e^{\alpha t}e^{\alpha(t+\tau)}]$, in order to conclude the proof, we would like to apply Theorem 2.3.3 at (2.20) with $p=1$ and $q=0$. So, we need to prove that the hypotheses on $f(t,\tau)$ are verified.

Note that $f(t,\tau)$ is the sum of three integrals, where each integrand, divided by $t$, converges to a constant (which is 0 for the last two integrands) when $t\to\infty$, uniformly for $\tau\geq 0$ (see (2.7), (2.8), and [42, pg. 145]). Furthermore, each of these integrands is dominated by the product of two locally bounded

functions (the moments of $Z(t)$ and $G(t)$ are locally bounded solutions of integral equations of the type in equation (2.20), see [42, pg. 142] and [116, Theorem 2]), one depending on $t$ and another one depending on $t+\tau$ (for the last integrand, use the Cauchy-Schwarz inequality to see it). As these dominant functions satisfy the hypotheses of Lemma 2.3.2 with $p=1$ and $q=0$ (see (2.8) and (2.7)), we can conclude that

$$\lim_{t\to\infty}\frac{f(t,\tau)}{t}=h\alpha' c^2\int_0^\infty d\mathbb{P}'(L\le u)=h\alpha' c^2 v\int_0^\infty e^{-2\alpha u}d\mathbb{P}(L\le u).$$

Moreover, if we consider the first of the integrals in (2.19) and apply the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}&\int_0^t \frac{\mathbb{E}(G(t-u))}{e^{\alpha(t-u)}}\frac{\mathbb{E}(Z(t+\tau-u))}{e^{\alpha(t+\tau-u)}}d\mathbb{P}'(L\le u)\\ \le&\left(\int_0^t\left|\frac{\mathbb{E}(G(t-u))}{e^{\alpha(t-u)}}\right|^2 d\mathbb{P}'(L\le u)\right)^{1/2}\left(\int_0^{t+\tau}\left|\frac{\mathbb{E}(Z(t+\tau-u))}{e^{\alpha(t+\tau-u)}}\right|^2 d\mathbb{P}'(L\le u)\right)^{1/2}\\ =:&\ f_1(t)f_2(t+\tau),\end{aligned}\tag{2.22}$$

with $f_1(t)$ and $f_2(t)$ satisfying the hypotheses of Theorem 2.3.3. As the same reasoning holds for the other integrals in (2.19) (for the last integral we use Cauchy-Schwarz inequality twice), thanks to Theorem 2.3.3, with $p=1$ and $q=0$, and Corollary 2.3.4 we obtain

$$\lim_{t\to\infty}\frac{K(t,\tau)}{t}=\frac{h\alpha' c^2 v\int_0^\infty e^{-2\alpha u}d\mathbb{P}(L\le u)}{1-h\int_0^\infty e^{-2\alpha u}d\mathbb{P}(L\le u)}\quad\text{uniformly for }\tau\ge 0.$$

Recalling the definition of $k$, $\mathcal{G}_t$, and $\mathcal{Z}_{t+\tau}$ at (2.9) and (2.10), we have completed the proof of the first inequality in (2.17). □

We now study the covariance between the total generation process at two distinct times, for which we will need to use Lemma 2.3.6.

**Lemma 2.3.7** (Limit behaviour of the covariance of $\mathcal{G}_t$ and $\mathcal{G}_{t+\tau}$)**.** *Using the previous notation, we have*

$$\lim_{t\to\infty}\mathbb{E}(\mathcal{G}_{t+\tau}\mathcal{G}_t)=k\qquad\textit{uniformly for }\tau\ge 0,$$

*where $k$ is defined in* (2.9).

**Proof:** The proof is similar to that in Lemma 2.3.6, so some details are omitted.

If we take the derivative of equation (2.16) first for $s_1$, secondly for $s_2$, and then evaluate it at $(1,1,1,1,t,\tau)$, we obtain

$$\begin{aligned}\mathbb{E}(G(t+\tau)G(t)) &= v\int_0^t \mathbb{E}(G(t+\tau-u))\mathbb{E}(G(t-u))d\mathbb{P}(L\le u)\\ &+ v\int_0^t \Big[\mathbb{E}(Z(t+\tau-u))\mathbb{E}(Z(t-u)) + \mathbb{E}(G(t+\tau-u))\mathbb{E}(Z(t-u))\\ &+ \mathbb{E}(Z(t+\tau-u))\mathbb{E}(G(t-u))\Big]d\mathbb{P}(L\le u)\\ &+ h\int_0^t \Big[\mathbb{E}\Big(G(t+\tau-u)Z(t-u)\Big) + \mathbb{E}\Big(Z(t+\tau-u)G(t-u)\Big)\\ &+ \mathbb{E}\Big(Z(t+\tau-u)Z(t-u)\Big)\Big]d\mathbb{P}(L\le u)\\ &+ h\int_0^t \mathbb{E}(G(t+\tau-u)G(t-u))d\mathbb{P}(L\le u). \end{aligned} \tag{2.23}$$

Multiplying both sides of this equation by $e^{-\alpha t}e^{-\alpha(t+\tau)}$ and denoting

$$K(t,\tau) := \frac{\mathbb{E}(G(t+\tau)G(t))}{e^{\alpha t}e^{\alpha(t+\tau)}},$$

$$d\overline{\mathbb{P}}(L\le u) := he^{-2\alpha u}d\mathbb{P}(L\le u), \qquad d\mathbb{P}'(L\le u) := ve^{-2\alpha u}d\mathbb{P}(L\le u),$$

$$\begin{aligned} f(t,\tau) := &\int_0^t \frac{\mathbb{E}(G(t+\tau-u))}{e^{\alpha(t+\tau-u)}}\frac{\mathbb{E}(G(t-u))}{e^{\alpha(t-u)}}d\mathbb{P}'(L\le u)\\ &+ \int_0^t \Big[\frac{\mathbb{E}(Z(t+\tau-u))}{e^{\alpha(t+\tau-u)}}\frac{\mathbb{E}(Z(t-u))}{e^{\alpha(t-u)}} + \frac{\mathbb{E}(G(t+\tau-u))}{e^{\alpha(t+\tau-u)}}\frac{\mathbb{E}(Z(t-u))}{e^{\alpha(t-u)}}\\ &+ \frac{\mathbb{E}(Z(t+\tau-u))}{e^{\alpha(t+\tau-u)}}\frac{\mathbb{E}(G(t-u))}{e^{\alpha(t-u)}}\Big]d\mathbb{P}'(L\le u)\\ &+ \int_0^t \Big[\frac{\mathbb{E}\Big(G(t+\tau-u)Z(t-u)\Big)}{e^{\alpha(t+\tau-u)}e^{\alpha(t-u)}} + \frac{\mathbb{E}\Big(Z(t+\tau-u)G(t-u)\Big)}{e^{\alpha(t+\tau-u)}e^{\alpha(t-u)}}\\ &+ \frac{\mathbb{E}\Big(Z(t+\tau-u)Z(t-u)\Big)}{e^{\alpha(t+\tau-u)}e^{\alpha(t-u)}}\Big]d\overline{\mathbb{P}}(L\le u), \end{aligned} \tag{2.24}$$

we have that

$$K(t,\tau) = f(t,\tau) + \int_0^t K(t-u,\tau)d\overline{\mathbb{P}}(L\le u). \tag{2.25}$$

As already observed in (2.21), $\overline{\mathbb{P}}$ is a defective measure. In order to conclude the proof, we would like to apply Theorem 2.3.3 to (2.25), and so we need to prove that the hypotheses on $f(t,\tau)$ are verified. This will be easier by proving a weaker version of Lemma 2.3.7 which states that $\lim_{t\to\infty}\mathbb{E}(G(t)^2)/[t^2e^{2\alpha t}] =$

$(h\alpha' c)^2 k$. This result, that we now prove, is obtained applying the first part of Theorem 2.3.3 to (2.25), when $\tau = 0$.

For $\tau = 0$, we have that $K(t,0) = \mathbb{E}(G(t)^2)/e^{2\alpha t}$ and

$$\begin{aligned} f(t,0) = &\int_0^t \left( \frac{\mathbb{E}(G(t-u))^2}{e^{2\alpha(t-u)}} + \frac{\mathbb{E}(Z(t-u))^2}{e^{2\alpha(t-u)}} \right) d\mathbb{P}'(L \leq u) \\ &+ 2\int_0^t \frac{\mathbb{E}(G(t-u))}{e^{\alpha(t-u)}} \frac{\mathbb{E}(Z(t-u))}{e^{\alpha(t-u)}} d\mathbb{P}'(L \leq u) \\ &+ \int_0^t \left[ 2\frac{\mathbb{E}\Big(G(t-u)Z(t-u)\Big)}{e^{2\alpha(t-u)}} + \frac{\mathbb{E}(Z(t-u)^2)}{e^{2\alpha(t-u)}} \right] d\overline{\mathbb{P}}(L \leq u). \quad (2.26) \end{aligned}$$

Notice that all five terms inside the integrals in (2.26) are locally bounded in $t$ (the moments and the covariance of $Z(t)$ and $G(t)$ are locally bounded solutions of integral equations of the type (2.25), see [116, Theorem 2]) and, divided by $t^2$, the integrands converge to constants, being 0 for all but the first term in the first integrand (see (2.8)). So, we can use Lemma 2.3.2 with $p = 2$ and $q = 0$, obtaining

$$\lim_{t\to\infty} \frac{f(t,0)}{t^2} = (h\alpha' c)^2 \int_0^\infty d\mathbb{P}'(L \leq u) = (h\alpha' c)^2 v \int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \leq u). \quad (2.27)$$

As $f(t,0)$ is locally bounded in $t$ (it is a finite sum of convolutions of locally bounded functions), equation (2.27) allows us to apply Theorem 2.3.3 obtaining

$$\begin{aligned} \lim_{t\to\infty} \frac{K(t,0)}{t^2} = \lim_{t\to\infty} \frac{\mathbb{E}(G(t)^2)}{t^2 e^{2\alpha t}} &= \frac{(h\alpha' c)^2 v \int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \leq u)}{1 - h\int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \leq u)} \\ &= (h\alpha' c)^2 k. \quad (2.28) \end{aligned}$$

Let's go back to the proof of Lemma 2.3.7 and see that $f(t,\tau)$ satisfies the hypotheses of Theorem 2.3.3. In (2.24), each of the seven integrands, when divided by $t(t+\tau)$, converges to a constant when $t \to \infty$, uniformly for $\tau \geq 0$ (see (2.8),(2.7),(2.17), and [42, pg. 145]). Furthermore, each of these integrands is dominated by the product of two locally bounded functions, one depending on $t$ and another one depending on $t+\tau$ (use the Cauchy-Schwarz inequality for the last three integrands to see it). As these functions satisfy the hypotheses of Lemma 2.3.2 (see (2.8),(2.7), and (2.28)), we can conclude that, uniformly for $\tau \geq 0$,

$$\lim_{t\to\infty} \frac{f(t,\tau)}{t(t+\tau)} = (h\alpha' c)^2 \int_0^\infty d\mathbb{P}'(L \leq u) = (h\alpha' c)^2 v \int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \leq u).$$

Moreover, using the Cauchy-Schwarz inequality (for the last three integrals we have to use it twice), each of the integrals in (2.24) are dominated by the product of two functions, one depending on $t$ and the other one on $t+\tau$, which satisfy the hypotheses of Theorem 2.3.3. So, Corollary 2.3.4 implies

$$\lim_{t\to\infty} \frac{K(t,\tau)}{t(t+\tau)} = \frac{(h\alpha' c)^2 v \int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \le u)}{1 - h\int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \le u)} \quad \text{uniformly for } \tau \ge 0.$$

The definitions of $k$ and $\mathcal{G}_t$ at (2.9) and (2.10), respectively, allow to conclude the proof. □

An immediate consequence of this lemma is the following.

**Corollary 2.3.8** ($\mathcal{G}_t$ is a Cauchy sequence in $L^2$.)**.** *Using the previous notation, we have*

$$\lim_{t\to\infty} \mathbb{E}((\mathcal{G}_{t+\tau} - \mathcal{G}_t)^2) \to 0 \qquad \textit{uniformly for } \tau \ge 0.$$

**Proof:** From Lemma 2.3.7 and (2.28), uniformly for $\tau \ge 0$, we have that

$$\lim_{t\to\infty} \mathbb{E}((\mathcal{G}_{t+\tau} - \mathcal{G}_t)^2) = \lim_{t\to\infty} \Big[\mathbb{E}(\mathcal{G}_{t+\tau}^2) + \mathbb{E}(\mathcal{G}_t^2) - 2\mathbb{E}(\mathcal{G}_{t+\tau}\mathcal{G}_t)\Big] = k + k - 2k = 0.$$

□

We have just proved that $\mathcal{G}_t$ is a Cauchy sequence in $L^2$, i.e. for every $\epsilon > 0$ there exists a $t_\epsilon > 0$ s.t. for every $t > t_\epsilon$ and $\tau \ge 0$ we have $\mathbb{E}((\mathcal{G}_{t+\tau} - \mathcal{G}_t)^2) < \epsilon$. Thanks to the completeness of the $L^2$ space, we can now easily prove Theorem 2.3.9.

**Theorem 2.3.9** (Mean square convergence of $G(t)$)**.** *There exists a non-negative random variable $\mathcal{G} \in L^2$ such that*

$$\lim_{t\to\infty} \mathbb{E}((\mathcal{G}_t - \mathcal{G})^2) = 0,$$

*with $\mathbb{E}(\mathcal{G}) = 1$ and $Var(\mathcal{G}) = k - 1 = [(v+h)\int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \le u) - 1]/[1 - h\int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \le u)] > 0$.*

**Proof:** The existence of a such $\mathcal{G}$ follows from Corollary 2.3.8, the fact that the $L^2$ space is complete, and that $\mathcal{G}_t$ satisfies the Cauchy criterion for convergence in $L^2$. Using (2.8) and the fact that $L^2 \subset L^1$, we know that

$\mathbb{E}(\mathcal{G}) = \lim_{t\to\infty} \mathbb{E}(\mathcal{G}_t) = 1$, so it remains only to compute the variance. From the $L^2$ convergence we have that $\mathbb{E}(\mathcal{G}^2) = \lim_{t\to\infty} \mathbb{E}(\mathcal{G}_t^2)$. Then,

$$\begin{aligned}\mathrm{Var}(\mathcal{G}) = \mathbb{E}(\mathcal{G}^2) - \mathbb{E}(\mathcal{G})^2 = \lim_{t\to\infty} \mathbb{E}(\mathcal{G}_t^2) - 1 &\overset{(2.28)}{=} k - 1 \\ &\overset{(2.9)}{=} \frac{(v+h)\int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \le u) - 1}{1 - h\int_0^\infty e^{-2\alpha u} d\mathbb{P}(L \le u)}. \end{aligned} \tag{2.29}$$

The positivity of (2.29) follows from the same argument used by Harris in [42, pg. 146]. Indeed, there he proved that the process $\mathcal{Z}_t$ converges a.s. to a random variable $\mathcal{Z}$ with the same mean and variance as $\mathcal{G}$. □

Theorem 2.3.9 gives us the mean square convergence of $\mathcal{G}_t$, which implies also the convergence in probability and in mean. In Section 2.3.5 we will see that the convergence is also true with probability one.

### 2.3.4 Functional equation for the MGF of $(\mathcal{G}, \mathcal{Z})$

A surprising consequence of Theorem 2.3.9 and [42, Theorem 19.1] is that the processes $\mathcal{G}$ and $\mathcal{Z}$ share the same mean and variance. In this section, using the Moment Generating Function (MGF) of the pair $(\mathcal{G}, \mathcal{Z})$, we prove that these two variables are actually almost surely equal. That is, on a path-by-path basis, the prefactor for the normalised population size and for the normalised total generation is the same with probability one.

**Theorem 2.3.10** ($Z(t)$ and $G(t)$ have same randomness in their dominant terms)**.** *Given*

$$\frac{G(t)}{ch\alpha' t e^{\alpha t}} = \mathcal{G}_t \overset{L^2}{\to} \mathcal{G} \quad \textit{and} \quad \frac{Z(t)}{ce^{\alpha t}} = \mathcal{Z}_t \overset{a.s.}{\to} \mathcal{Z}$$

*we have that*

$$\mathcal{G} = \mathcal{Z} \quad a.s.$$

**Proof:** The proof is divided in two parts: first, we prove that $\mathcal{G}$ and $\mathcal{Z}$ are equally distributed, then that they coincide with probability one.

Theorem 2.3.9, together with (2.7), imply that $(\mathcal{G}_t, \mathcal{Z}_t) \overset{D}{\to} (\mathcal{G}, \mathcal{Z})$ in distribution. So, we can characterise the distribution of the pair $(\mathcal{G}, \mathcal{Z})$ studying the MGF of $(\mathcal{G}_t, \mathcal{Z}_t)$ when $t \to \infty$.

Proposition 2.3.5 gives us an equation solved by the Probability Generating Function (PGF) of the vector $(G(t), G(t+\tau), Z(t), Z(t+\tau))$. Evaluating this

equation in $(s_1, 1, r_1, 1, t, 0)$, we obtain the following expression solved by the PGF $F(s_1, r_1, t) := \mathbb{E}(s_1^{G(t)} r_1^{Z(t)})$ of $(G(t), Z(t))$

$$F(s_1, r_1, t) = r_1 \mathbb{P}(L > t) + \int_0^t \rho_N\Big(F(s_1, s_1 r_1, t-u)\Big) d\mathbb{P}(L \leq u). \tag{2.30}$$

Replacing $s_1$ with $\exp(-s/[hc\alpha' t e^{\alpha t}])$ and $r_1$ with $\exp(-r/[ce^{\alpha t}])$, for $s, r \geq 0$, we obtain an expression solved by the MGF $\phi(s, r, t)$ of $(\mathcal{G}_t, \mathcal{Z}_t)$:

$$\begin{aligned}
\phi(s, r, t) &= \mathbb{E}\Big(e^{-\frac{sG(t)}{hc\alpha' t e^{\alpha t}}} e^{-\frac{rZ(t)}{ce^{\alpha t}}}\Big) \\
&= e^{-\frac{r}{ce^{\alpha t}}} \mathbb{P}(L > t) + \int_0^t \rho_N\Big(\mathbb{E}\Big(e^{-\frac{(t-u)se^{-\alpha u}}{t} \mathcal{G}_{t-u}} e^{-\frac{(s+hr\alpha' t)e^{-\alpha u}}{h\alpha' t} \mathcal{Z}_{t-u}}\Big)\Big) d\mathbb{P}(L \leq u) \\
&= e^{-\frac{r}{ce^{\alpha t}}} \mathbb{P}(L > t) + \int_0^t \rho_N\Big(\phi\Big(\frac{(t-u)}{t} s e^{-\alpha u}, \frac{(s + hr\alpha' t)}{h\alpha' t} e^{-\alpha u}, t-u\Big)\Big) d\mathbb{P}(L \leq u).
\end{aligned}$$

Taking the limit for $t \to \infty$ of $\phi(s, r, t)$, we obtain that the function $\phi(s, r) := \mathbb{E}(\exp(-s\mathcal{G}) \exp(-r\mathcal{Z}))$ solves the integral equation

$$\phi(s, r) = \int_0^\infty \rho_N\Big(\phi\Big(se^{-\alpha u}, re^{-\alpha u}\Big)\Big) d\mathbb{P}(L \leq u) \quad s, r \geq 0. \tag{2.31}$$

This means that if we consider $r = 0$, the function $\psi(s) := \mathbb{E}[\exp(-s\mathcal{G})]$, that represents the MGF of $\mathcal{G}$, solves the integral equation

$$\psi(s) = \int_0^\infty \rho_N\Big(\psi\Big(se^{-\alpha u}\Big)\Big) d\mathbb{P}(L \leq u), \quad s \geq 0 \tag{2.32}$$

with $\psi(0) = 1$ and $\psi'(0) = -1$. The uniqueness of the solution of this problem [69, pg. 122] and the fact that the MGF of the variable $\mathcal{Z}$ solves (2.32) too [42, pg. 146], give us that the MGFs of $\mathcal{Z}$ and $\mathcal{G}$ coincide for $s \geq 0$. Using a result proved by Mukherjea et al. [80, Theorem 2], we can conclude that $\mathcal{Z}$ has the same distribution as $\mathcal{G}$.

Now, if we consider $r = s$ in (2.31), we can see that the function $\mathbb{E}[\exp(-s(\mathcal{G} + \mathcal{Z}))]$, that represents the MGF of $\mathcal{G} + \mathcal{Z}$, solves (2.32) but with the initial conditions $\psi(0) = 1$ and $\psi'(0) = -2$. Another solution of (2.32) with the same initial conditions is given by $2\mathcal{Z}$. Also in this case, the uniqueness of the solution and [80, Theorem 2] allows us to conclude that $2\mathcal{Z} \overset{D}{=} \mathcal{Z} + \mathcal{G}$.

These last two results give us that $\mathcal{Z} \overset{a.s.}{=} \mathcal{G}$. In fact, given both $\mathcal{Z}$ and $\mathcal{G}$ are in $L^2$, we have

$$\begin{aligned}
2\mathcal{Z} \overset{D}{=} \mathcal{Z} + \mathcal{G} \quad &\Longrightarrow \quad 4\mathrm{Var}(\mathcal{Z}) = \mathrm{Var}(\mathcal{Z}) + \mathrm{Var}(\mathcal{G}) + 2\mathrm{Cov}(\mathcal{Z}, \mathcal{G}) \\
\Longrightarrow \quad \mathrm{Var}(\mathcal{Z}) = \mathrm{Cov}(\mathcal{Z}, \mathcal{G}) \quad &\Longrightarrow \quad \mathrm{Corr}_{\mathcal{Z},\mathcal{G}} := \frac{\mathrm{Cov}(\mathcal{Z}, \mathcal{G})}{\sqrt{\mathrm{Var}(\mathcal{Z})}\sqrt{\mathrm{Var}(\mathcal{G})}} = 1,
\end{aligned}$$

where in the last inequality we have used the definition of Pearson's correlation coefficient. The correlation coefficient equal to 1 implies that $\mathcal{G} = a\mathcal{Z} + b$ a.s., for $a \geq 0$, $b \in \mathbb{R}$ [18, Theorem 4.5.7]. Since $\mathcal{Z}$ and $\mathcal{G}$ are in $L^2$ and $\mathcal{Z} \overset{D}{=} \mathcal{G}$, we have that $\text{Var}(a\mathcal{Z} + b) = \text{Var}(\mathcal{Z})$ and $\mathbb{E}(a\mathcal{Z} + b) = \mathbb{E}(\mathcal{Z})$, from which we obtain $a^2 = 1$, so $a = 1$, and $b = 0$. This conclude the proof. □

Thus, from Theorem 2.3.10, $\mathcal{Z}$ can be used in lieu of $\mathcal{G}$ from here on.

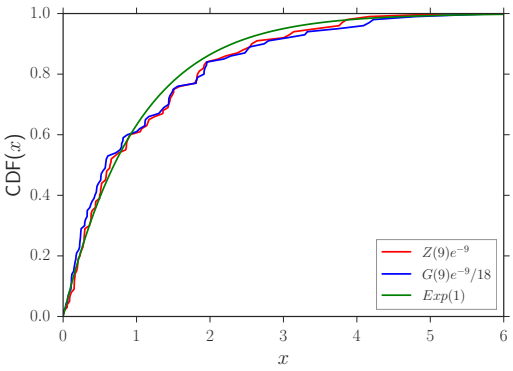
**Example 2.3.1** (Pure birth process)**.** When the lifetimes of cells are exponential distributed, i.e. $L \overset{D}{\sim}\text{Exp}(\lambda)$, and the number of offspring for each cell is $N = 2$ a.s., the branching process obtained is called a pure birth process [93, pg. 370]. In this case, we can fully characterise $\mathcal{Z}$ by finding an expression for its MGF. In fact, the PGF of $Z(t)$, $F_Z(s,t) := \mathbb{E}(s^{Z(t)})$, solves the differential equation $F'_Z(s,t) = \lambda[F_Z(s,t)^2 - F_Z(s,t)]$ [42, pg. 106], and knowing that for exponential lifetimes, the Malthusian parameter, $\alpha(h)$, is given by $\alpha(h) = \lambda(h-1)$, we find the closed-form expression

$$F_Z(s,t) = \frac{s}{s - (s-1)e^{\lambda t}}. \tag{2.33}$$

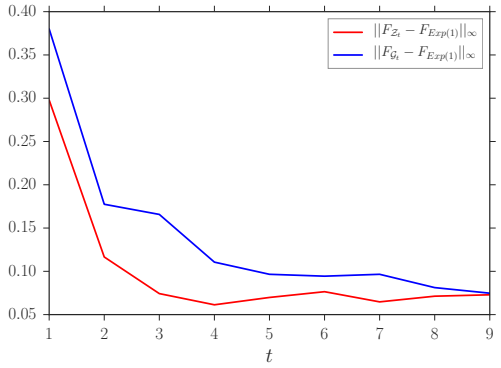
We can now use the expression in (2.33) to characterise $\mathcal{Z}$ as

$$\phi_{\mathcal{Z}}(s) := \mathbb{E}(e^{-s\mathcal{Z}}) = \lim_{t\to\infty} F_Z(e^{-s/e^{\lambda t}}, t) = \frac{1}{1+s}, \tag{2.34}$$

which is the Laplace transform of an Exp(1) distribution (e.g. see [38, Example 7.45]), and where we have used the fact that $\mathbb{E}(Z(t))/e^{\lambda t} = 1$. So, for a pure birth process, $\mathcal{Z}$ and $\mathcal{G}$ are exponential random variables with parameter 1 that, according to Theorem 2.3.10, are equal almost everywhere. We can also see that in Fig. 2.3, where in the first panel a comparison between $\mathcal{Z}_t$, $\mathcal{G}_t$, and a random variable exponentially distributed with parameter 1, for $t = 9$, is shown. Given the empirical Cumulative Distribution Functions (eCDFs) of $\mathcal{Z}_t$ and $\mathcal{G}_t$ represent two approximations of the CDF of $\mathcal{Z}$ for large $t$, Fig. 2.3(a) suggests that a convergence in distribution of $\mathcal{Z}_t$ and $\mathcal{G}_t$ to $\mathcal{Z}$ is taking place. This is more evident in Fig. 2.3(b) where, using the concept of distance coming from the uniform norm, $d(f,g) := ||f - g||_\infty$ (e.g. see [97, pg. 150]), the distance between the eCDF of $\mathcal{Z}_t$ and $\mathcal{G}_t$, and the CDF of $\mathcal{Z}$ are described for different values of $t$.




(a)




(b)

Figure 2.3: **Comparison between the distributions of $\mathcal{Z}_t$, $\mathcal{G}_t$, and Exp(1).** The figure uses 100 Monte Carlo simulations of a pure birth process, starting with one cell at time $t = 0$. (a) The first panel shows the empirical Cumulative Distribution Function (eCDF) of $\mathcal{Z}_t$ (red line), $\mathcal{G}_t$ (blue line), and the CDF of an exponential distribution with parameter 1 (green line) for $t = 9$. (b) Using the notion of distance coming from the uniform norm, $d(f, g) := ||f - g||_\infty$ (e.g. see [97, pg. 150]), the plot shows how close the eCDFs of $\mathcal{Z}_t$ and $\mathcal{G}_t$ (resp. $F_{\mathcal{Z}_t}$ and $F_{\mathcal{Z}_t}$) are to the CDF of Exp(1) ($F_{Exp(1)}$) for different value of $t$.

### 2.3.5 Almost sure convergence of $G(t)$

We have gathered the results needed to establish one of the significant results of the thesis: the almost sure convergence of a normalised version of the process $\{G(t)\}$. In order to prove that, we will assume something concerning the speed of convergence of $\mathcal{G}_t$ to $\mathcal{Z}$ as $L^2$ functions. This assumption is equivalent to the one made by Harris in [42, Chapter VI, Theorem 21.1] concerning the size of the population, which - for the population size - was later established by Jagers [54] to be unnecessary.

**Theorem 2.3.11** (Almost sure convergence of $G(t)$)**.** *If* $\int_0^\infty \mathbb{E}((\mathcal{G}_t - \mathcal{Z})^2)dt < \infty$*, we have that*

$$\frac{G(t)}{hc\alpha' te^{\alpha t}} = \mathcal{G}_t \xrightarrow[t\to\infty]{a.s} \mathcal{Z}.$$

**Proof:** We start with the additional hypothesis $p_0 = \mathbb{P}(N = 0) = 0$ in order to have $G(t)$ as a finite, non-decreasing step function of $t$. Using Fubini's theorem on $\int_0^\infty \mathbb{E}((\mathcal{G}_t - \mathcal{Z})^2)dt < \infty$, we obtain that $\mathbb{P}(\int_0^\infty (\mathcal{G}_t - \mathcal{Z})^2 dt < \infty) = 1$. Since $G(t)$ is non-decreasing in $t$, we have

$$\mathcal{G}_{t+\tau} = \frac{G(t+\tau)}{hc\alpha'(t+\tau)e^{\alpha(t+\tau)}} \geq \frac{t}{(t+\tau)e^{\alpha\tau}} \frac{G(t)}{hc\alpha' te^{\alpha t}} = \frac{t}{(t+\tau)e^{\alpha\tau}} \mathcal{G}_t, \qquad (2.35)$$

where the inequalities are true for every realisation of the random variables.

Let's suppose that $\mathcal{G}_t \xrightarrow[t\to\infty]{a.s.} \mathcal{Z}$ not true . If $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ is the probability space where $\mathcal{G}_t$ and $\mathcal{Z}_t$ are defined, then there exists a set $A \subseteq \{\omega \in \Omega | \lim_{t\to\infty} \mathcal{G}_t(\omega) \neq \mathcal{Z}(\omega)\}$ that is measurable and such that $\mathbb{P}(A) > 0$. Since $\mathcal{Z} > 0$ a.s. [42, Remark 1, Section 20], we can also suppose that $\mathcal{Z}(\omega) > 0$ for every $\omega \in A$.

For every $\omega \in A$ we have that at least one of the possibilities $\limsup_{t\to\infty} \mathcal{G}_t(\omega) > \mathcal{Z}(\omega)$ and $\liminf_{t\to\infty} \mathcal{G}_t(\omega) < \mathcal{Z}(\omega)$ is true. We will see that in both cases we will have $\int_0^\infty (\mathcal{G}_t(\omega) - \mathcal{Z}(\omega))^2 dt = +\infty$, leading to the contradiction $\mathbb{E}(\int_0^\infty (\mathcal{G}_t - \mathcal{Z})^2 dt) = +\infty$.

Let us start fixing $\omega \in A$ and assuming $\limsup_{t\to\infty} \mathcal{G}_t(\omega) > \mathcal{Z}(\omega)$. This implies that there exist a $\delta > 0$ and a sequence $(t_i)_{i\in\mathbb{N}}$, with $\lim_{i\to\infty} t_i = \infty$, such that $\mathcal{G}_{t_i}(\omega) > (1+\delta)\mathcal{Z}(\omega)$, $i \in \mathbb{N}$. If we consider $0 < \epsilon < \delta$, without loss of generality we can choose this sequence such that

$$t_{i+1} - t_i > \frac{(\delta - \epsilon)t_i}{1 + \epsilon + \alpha t_i(1+\delta)} := b_i.$$

Note that $\delta, \epsilon$, and $t_i$ depend on $\omega$ and that $(b_i)_{i\in\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}}$ are monotonically increasing.

Using (2.35) and the relation $e^{-\alpha\tau} \geq 1 - \alpha\tau$, we obtain for every $i \in \mathbb{N}$

$$\begin{aligned}\mathcal{G}_{t_i+\tau}(\omega) &\overset{(2.35)}{\geq} \frac{t_i}{(t_i+\tau)e^{\alpha\tau}}\mathcal{G}_{t_i}(\omega) > \frac{t_i}{t_i+\tau}(1-\alpha\tau)(1+\delta)\mathcal{Z}(\omega), \quad \tau \in (0,\infty)\\ &\geq (1+\epsilon)\mathcal{Z}(\omega) \qquad\qquad \tau \in (0, b_i)\\ &\geq (1+\epsilon)\mathcal{Z}(\omega) \qquad\qquad \tau \in (0, b_1),\end{aligned} \tag{2.36}$$

where we have used the fact that the function $t_i(t_i+\tau)^{-1}(1-\alpha\tau)(1+\delta)$ is decreasing in $\tau$, that for $\tau = b_i$ it is equal to $(1+\epsilon)$, and that $(b_i)_{i\in\mathbb{N}}$ is an increasing sequence.

Hence, using (2.36), we have for every $i$ that

$$\begin{aligned}&\int_{t_i}^{t_{i+1}} (\mathcal{G}_t(\omega) - \mathcal{Z}(\omega))^2 dt \geq \int_{t_i}^{t_i+b_1} (\mathcal{G}_t(\omega) - \mathcal{Z}(\omega))^2 dt\\ &= \int_0^{b_1} (\mathcal{G}_{t_i+\tau}(\omega) - \mathcal{Z}(\omega))^2 d\tau \geq (\epsilon\mathcal{Z}(\omega))^2 b_1 > 0.\end{aligned}$$

This allows us to say that $\int_0^\infty (\mathcal{G}_t(\omega) - \mathcal{Z}(\omega))^2 dt = +\infty$.

The same conclusion can be obtained assuming $\liminf_{t\to\infty} \mathcal{G}_t(\omega) < \mathcal{Z}(\omega)$. Indeed, for the definition of lim inf we have that there exist $\delta \in (0,1)$ and a sequence $(t_i)_{i\in\mathbb{N}}$, with $t_i > 1$ and $\lim_{i\to\infty} t_i = \infty$, such that $\mathcal{G}_{t_i} < (1-\delta)\mathcal{Z}$. We can also pretend that $t_{i+1} - t_i > a > 0$, where $a$ is chosen in order to satisfy the following inequalities for $i$ big enough

$$\begin{aligned}0 < \mathcal{G}_{t_i-\tau} &\overset{(2.35)}{\leq} \frac{t_i}{t_i-\tau}e^{\alpha\tau}\mathcal{G}_{t_i} < (1-\delta)\frac{t_i}{t_i-\tau}e^{\alpha\tau}\mathcal{Z} \qquad \tau \in (0, t_1)\\ &\leq (1-\epsilon)\mathcal{Z} \qquad\qquad \tau \in (0, a),\end{aligned}$$

where $\epsilon$ is a constant s.t. $0 < \epsilon < \delta$. The existence of such $a$ is a consequence of the fact that $\psi(t,\tau) := (1-\delta)e^{\alpha\tau}t/(t-\tau)$, as long as $\tau < t$, is increasing in $\tau$ and decreasing in $t$. Indeed, this implies that there exists $a > 0$ s.t. for $\tau \in [0, a]$, $(1-\delta) = \psi(1,0) \leq \psi(1,\tau) \leq (1-\epsilon)$, from which we can conclude that for $\tau \in [0,a]$ and $t \geq 1$, we have $(1-\delta) = \psi(t,0) \leq \psi(t,\tau) \leq (1-\epsilon)$ .

Then, we have

$$\int_{t_{i-1}}^{t_i} (\mathcal{Z}(\omega) - \mathcal{G}_t(\omega))^2 dt \geq \int_{t_i-a}^{t_i} (\mathcal{Z}(\omega) - \mathcal{G}_t(\omega))^2 dt$$
$$\geq \int_0^a (\mathcal{Z}(\omega) - \mathcal{G}_{t_i-\tau}(\omega))^2 d\tau \geq (\epsilon \mathcal{Z}(\omega))^2 a.$$

As before, this implies that $\int_0^\infty (\mathcal{G}_t(\omega) - \mathcal{Z}(\omega))^2 dt = +\infty$.

So, for every $\omega \in A$ we have $\int_0^\infty (\mathcal{G}_t(\omega) - \mathcal{Z}(\omega))^2 dt = +\infty$ and, because $\mathbb{P}(A) > 0$, we have $\mathbb{E}(\int_0^\infty (\mathcal{G}_t - \mathcal{Z})^2 dt) = +\infty$. This contradicts the hypothesis of the theorem and so we have proved that $\lim_{t\to\infty} \mathcal{G}_t = \mathcal{Z}$ with probability 1 under the condition $p_0 = 0$.

When $p_0 \neq 0$, we can observe that $G(t) = G_B(t) - G_D(t)$, where $G_B(t)$ and $G_D(t)$ are the sum of the generation of the cells born and dead before or at time $t$, respectively. Also for these processes we can find integral equations for the probability generating function similar to the one found for $G(t)$ and repeat all the previous steps. Thanks to the monotonicity of $G_B(t)$ and $G_D(t)$, this time we don't need the assumption $p_0 = 0$, obtaining the almost sure convergence of $G_B(t)/n_1 te^{\alpha t}$ and $G_D(t)/n_2 te^{\alpha t}$ to the random variables $\mathcal{Z}_B$ and $\mathcal{Z}_D$ respectively, where $n_1, n_2$ are positive constants. This allows us to conclude that $\mathcal{G}_t$ converges to $\mathcal{Z}_B - \mathcal{Z}_D$. □

Having established the almost sure result for the limiting behaviour of the total generation process $G(t)$, we are in a position to make the final deduction of the section that leads to equation (2.3). Thanks to equation (2.7), Theorem 2.3.11, and the Continuous Mapping Theorem (e.g. [103, pg. 24]), we have the following corollary.

**Corollary 2.3.12** (Almost sure average generation inference)**.** *If* $\mathbb{E}(N^2) < \infty$*,* $lim\ inf_{t\to\infty} Z^+(t) > 0$*, and* $\int_0^\infty \mathbb{E}((\mathcal{G}_t - \mathcal{Z})^2) dt < \infty$*, we have almost surely that*

$$\lim_{t\to\infty} \frac{G(t)}{tZ(t)} = h\alpha', \quad \lim_{t\to\infty} -\frac{1}{pt} \log\left(\frac{Z^+(t)}{Z(t)}\right) = \frac{\alpha(h) - \alpha(h(1-p))}{p}, \quad (2.37)$$

*and*

$$\lim_{p\to 0} \frac{\alpha(h) - \alpha(h(1-p))}{p} = h\alpha', \quad (2.38)$$

*where the second limit in* (2.37) *and* (2.38) *are established in [116, Theorem 1].*

Thus the average estimation scheme, firstly proposed in [116], is almost surely correct on a path-by-path basis for a Bellman-Harris branching process. The principle behind this result is that for almost all paths, $G(t)/Z(t) = h\alpha' t + o(t)$ for large $t$. That is, the randomness in the average generation is not contained in the linear factor but in something asymptotically smaller (see Fig. 2.1(c)). On the other hand, so long as the $Z^+$ population persists, $-1/t \log(Z^+(t)/Z(t)) \approx \alpha(h(1-p)) - \alpha(h)$ for large $t$. However, as the Malthus parameter is real analytic [116, Proposition 1], $\alpha(h(1-p))$ coincides with its Taylor expansion around $p = 0$, $\alpha(h(1-p)) = \alpha(h) - h\alpha'(h)pt + O(p^2)$. Thus we have that $-1/(pt) \log(Z^+(t)/Z(t)) \approx h\alpha'(h)$, the same constant as appears for the time-rescaled average generation.

**Example 2.3.1** (Continued)**.** Continuing the example in the Markovian setting started in the previous section, we can see that the quantity $h\alpha'$ that appears in equation (2.37) and (2.38) is given by $2\lambda$, with $\lambda$ parameter of the exponential lifetime distribution.

Corollary 2.3.12 is the core result of the chapter and provides stronger mathematical support for using the proportion of label positive cell to estimate the average generation of a population for a single realisation of a growing tree. From what we have seen, the results of this chapter involves only homogeneous populations, i.e. where each cell divides independently from the others but according to the same probability laws. In the next chapter we extend Theorem 2.3.11 and Corollary 2.3.12 to a two-type setting, where cells belonging to the same type share the same division laws. This extends the range of application of the method to heterogeneous populations that can rise for example as result of a differentiation process or as a consequence of a genetic mutation.

CHAPTER 3

# Average generation in a two-type Branching Process

## 3.1 Introduction

In Chapter 2, we studied a growing cell population in which cells are subject to division and death, modelling it with a super-critical Bellman-Harris branching process. That allows us to obtain insights on the limit behaviour of the sum of the generations of the living cell at time $t$, $G(t)$ (Theorem 2.3.11). Joining this result with what is already known about the size of the population [42, 54], $Z(t)$, and the proportion of label-positive cells [116], $Z^+(t)/Z(t)$, we obtained an enhancement of the mathematical results underpinning the DNA coded randomised algorithm for the estimation of the average generation introduced in [116] (Corollary 2.3.12).

In addition to division and death, cells often undergo changes in cell-type. For example, many tissues of the human body are formed through progressive stages of proliferation and change in cell-type, called cellular differentiation, from stem cells [66, 1], while cancer cells arise as mutants with aberrant DNA from healthy cells [76, 48]. Changes in cell-type are often accompanied by changes in population kinetics, e.g. [2], and to better understand these differ-

entiation processes, it is desirable to obtain information on the average generation of each population as they are often reported as being division-linked [47, 25, 26, 86].

As a basic model of changes in cell type, in the present chapter we extend the results in Chapter 2 to a two-type Bellman-Harris branching process subject to one-way differentiation, a model first considered by Jagers in [53] where cells of one type can give rise to another but not vice-versa. These results significantly extend the remit and utility of the inference of average generation by random delabelling. In particular, if the initial cell is equipped with a neutral label that is heritably lost with a fixed probability per division, we prove that the average generation of each cell-type can be inferred from knowledge of that probability and the proportion of label positive cells.

Before presenting those results, in the next section we explain the model we are going to use to describe the growth of the two-type cell population in more detail. The notation and the assumptions we employ throughout this chapter are consistent with those used, for the single-type case, in Chapter 2 and with those employed in [53], where sample path results for the population size were first established in this two-type setting.

## 3.2 Model and notation

As in Fig. 1.5, consider a cell population whose members are one of two types, type-1 and type-2. Each cell lives a random type-dependent lifetime $L_i$, $i \in \{1, 2\}$, after which it dies or divides generating $N_i$ offspring. We assume $L_i$ and $N_i$ are independent for each cell, and amongst all cells. Furthermore, we suppose that type-1 cells can generate cells of both types, i.e. $N_1$ takes values in $\mathbb{N}^2$ and $\mathbb{P}(N_1 = (k, j))$ is the probability that $k$ type-1 and $j$ type-2 cells are generated from a type-1 cell after division. We also denote with $\rho_1(x_1, x_2) := \sum_{k,j\in\mathbb{N}} \mathbb{P}(N_1 = (k, j))\, x_1^k x_2^j$ the Probability Generating Function (PGF) of $N_1$. Differently, we assume that the offspring of type-2 cells are exclusively type-2 cells, so that $N_2$ takes value in $\mathbb{N}$ and has PGF $\rho_2$. We denote by $h_i := (\partial/\partial x_i)\rho_1(1, 1)$ the average number offspring of type-$i$ generated from a type-1 cell and, with $\mu := (d/dx)\rho_2(1)$, the average number of offspring obtained from a type-2 cell.

What we have just described is the typical setting of a two-type cell population

in which each sub-population behaves according to a Bellman-Harris branching process with immigration from type-1 to type-2. Given part of the offspring of a type-1 cell is changing type, the growth rate of the type-1 population is a function of $h_1$ and, as in the single-type case studied in Chapter 2, we suppose that $h_1$ and $\mu$ are greater than 1 so that both populations are super-critical.

To each cell we assign a generation, the integer that records how many divisions led to that cell (Fig. 1.5). We define cells at time zero as being in generation zero. Furthermore, we suppose the cells in the initial population are equipped with a neutral label (i.e. one that does not influence population dynamics) that, independently for each cell, is heritably lost immediately prior to a cell's division with probability $p$. For $i \in \{1, 2\}$, we denote by $Z_i(t)$ the total number of type-$i$ cells in the population at time $t$, by $G_i(t)$ the total generation of type-$i$ cells at time $t$, and by $Z_i^+(t)$ the size of type-$i$ label-positive at time $t$. To describe the growth rates of these processes, we need the concept of Malthusian parameter that we have already introduced in Section 2.1. In particular, we define $\alpha_1$ and $\alpha_2$ as the solutions of the equations

$$h_1 \mathbb{E}\left(e^{-\alpha_1 L_1 t}\right) = 1 \quad \text{and} \quad \mu \mathbb{E}\left(e^{-\alpha_2 L_2 t}\right) = 1. \tag{3.1}$$

The existence and the uniqueness of the solutions of these equations are guaranteed by the hypotheses $h_1 > 1$ and $\mu > 1$, which also lead to $\alpha_1, \alpha_2 > 0$. As said in Section 2.3.1 and proved in [116], $\alpha_1$ and $\alpha_2$ can be seen as analytic functions of $h_1$ and $\mu$, respectively. This allows us to define their first derivatives at the points $h_1$ and $\mu$, using the expressions

$$\alpha_1' = \frac{1}{h_1^2 \int_0^{+\infty} t e^{-\alpha_1 t} d\mathbb{P}(L_1 \leq u)} \quad \text{and} \quad \alpha_2' = \frac{1}{\mu^2 \int_0^{+\infty} t e^{-\alpha_2 t} d\mathbb{P}(L_2 \leq u)}.$$

In the next section, we will see that whether $\alpha_1$ is greater or smaller than $\alpha_2$ has a significant impact on the behaviour of the type-2 population, influencing the growth rate of both $Z_2(t)$ and $G_2(t)$. Fortunately, this doesn't influence the estimator's ability to infer, at large time and for small delabelling probability, the average generation of the type-2 population with the use of the proportion of label-positive cells of the same type.

## 3.3 Results

According to the model described in Section 3.2, the population dynamics of type-1 cells are unaffected by type-2 cells. In fact, treating differentiation as

death, the type-1 population behaves as a single type process. If the starting population only has type-2 cells, the system is again in the single type setting, given no changes from type-2 to type-1 cells are allowed. Thus the interesting setup is when the system is initiated with cells of type-1 and queries are of the population size and average generation of type-2 cells.

Let $\mathbb{P}_i$ and $\mathbb{E}_i$ denote the probability and the expectation conditional on the total population starting with a single cell of type $i \in \{1, 2\}$. The growth of the type-2 population size given one initial type-1 cell, $Z_2(t)$ under $\mathbb{P}_1$, is studied in [53]. Those results can be immediately applied to study $Z_2^+(t)$, given the first cell is type-1 and label-positive. Analogous results for $G_2(t)$ can be obtained by repeating the steps made in the single-type case, starting from the generalisation of Proposition 2.3.5 to the two-type problem.

**Proposition 3.3.1.** *Let us denote with* $F_1(s_1, s_2, s_3, s_4, r_1, r_2, r_3, r_4, t, \tau) := \mathbb{E}_1(s_1^{G_1(t)} s_2^{G_2(t)} s_3^{G_1(t+\tau)} s_4^{G_2(t+\tau)} r_1^{Z_1(t)} r_2^{Z_2(t)} r_3^{Z_1(t+\tau)} r_4^{Z_2(t+\tau)})$ *and with the function* $F_2(s_2, s_4, r_2, r_4, t, \tau) := \mathbb{E}_2(s_2^{G_2(t)} s_4^{G_2(t+\tau)} r_2^{Z_2(t)} r_4^{Z_2(t+\tau)})$ *for* $s_i, r_i, t, \tau \in \mathbb{R}_{\geq 0}$, *with* $i = 1, \ldots, 4$. *Under this notation, we have that*

$$
\begin{aligned}
&F_1(s_1, s_2, s_3, s_4, r_1, r_2, r_3, r_4, t, \tau) = r_1 r_3 \mathbb{P}(L_1 > t + \tau) \\
&+ r_1 \int_t^{t+\tau} \rho_{N_1} \left( F_1(1, 1, s_3, s_4, 1, 1, s_3 r_3, s_4 r_4, t - u, \tau), F_2(1, s_4, 1, s_4 r_4, t - u, \tau) \right) \\
&\qquad d\mathbb{P}(L_1 < u) \\
&+ \int_0^t \rho_{N_1} \left( F_1(s_1, \ldots, s_4, s_1 r_1, \ldots, s_4 r_4, t - u, \tau), F_2(s_2, s_4, s_2 r_2, s_4 r_4, t - u, \tau) \right) \\
&\qquad d\mathbb{P}(L_1 < u),
\end{aligned}
$$

*where* $\rho_{N_1}(s) := \mathbb{E}(s^{N_1})$, *and that* $F_2 = F$ *is given by* (2.16).

**Proof:** The recursive equation for $F_1$ can be found using arguments similar to the ones used in [116, Theorem 2] to find an integral expression for the PGF of $(G(t), Z(t))$ in the single-type case. The idea behind it is to use the Law of Total Probability (e.g. [65, Theorem 8.6]), with the partition of the sample space given by $\{\{L_2 > t + \tau\}, \{t < L_2 \leq t + \tau\}, \{L_2 \leq t\}\}$, to decompose $F_1$ as a sum of 3 terms. Furthermore, $F_2$ coincides exactly with the function $F$ defined in Proposition 2.3.5. □

Using Proposition 3.3.1, Lemma 2.3.2 and Theorem 2.3.3 we can establish the growth rates of $\mathbb{E}_1(G_2(t))$, $\mathbb{E}_1(G_2(t)Z_2(t))$, $\mathbb{E}_1(G_2(t)^2)$, $\mathbb{E}_1(Z_2(t)Z_2(t+\tau))$,

$\mathbb{E}_1(G_2(t)Z_2(t+\tau))$, $\mathbb{E}_1(G_2(t+\tau)Z_2(t))$, and $\mathbb{E}_1(G_2(t)G_2(t+\tau))$. In particular, we obtain that

**Lemma 3.3.2.** *Under the assumptions made, when $t \to \infty$, we have that the quantities*

$$\frac{\mathbb{E}_1(Z_2(t))}{e^{\alpha t}}, \qquad \frac{\mathbb{E}_1(G_2(t))}{te^{\alpha t}}, \qquad \frac{\mathbb{E}_1(Z_2(t)Z_2(t+\tau))}{e^{2\alpha t}},$$
$$\frac{\mathbb{E}_1(G_2(t)Z_2(t+\tau))}{te^{\alpha t}e^{\alpha(t+\tau)}}, \qquad \frac{\mathbb{E}_1(Z_2(t)G_2(t+\tau))}{(t+\tau)e^{\alpha t}e^{\alpha(t+\tau)}}, \qquad \frac{\mathbb{E}_1(G_2(t)G_2(t+\tau))}{(t+\tau)^2e^{2\alpha(t+\tau)}}$$

*converge to constants different from 0 for $\tau \in [0,\infty)$, where $\alpha := \max(\alpha_1, \alpha_2)$. Moreover, for the quantities above that depend on $\tau$, it is also true that the convergence is uniform for $\tau \in [0,\infty)$.*

**Proof:** The proof essentially follows the same line of reasoning of Lemmas 2.3.6 and 2.3.7 and so is not repeated. □

From this lemma, starting with one label-positive type-1 cell, the in-expectation result relating the average generation to the proportion of labelled cells follows immediately:

$$\lim_{t\to\infty} \frac{\mathbb{E}_1(G_2(t))}{t\mathbb{E}_1(Z_2(t))} = -\lim_{p\to 0}\lim_{t\to\infty} \frac{1}{pt}\log\left(\frac{\mathbb{E}_1(Z_2^+(t))}{\mathbb{E}_1(Z_2(t))}\right).$$

This equation says that a quantity somewhat similar to the expected average generation of the type-2 population can be determined from averages of the delabelling proportion, where the averages are taken over the multi-type Bellman-Harris construction, and all delabellings. These results do not provide per sample-path guarantees, which would be desirable. To obtain those convergence results, one notes that a combination of [42, Theorems 19.1 and 21.1], Theorem 2.3.9, and Theorem 2.3.11 gives that

$$\lim_{t\to\infty} \frac{Z_i(t)}{c_i e^{\alpha_i t}} \overset{L^2,a.s.}{=} \mathcal{Z}_i \quad \text{and} \quad \lim_{t\to\infty} \frac{G_i(t)}{d_i t e^{\alpha_i t}} \overset{L^2,a.s.}{=} \mathcal{Z}_i \quad \text{under } \mathbb{P}_i, \tag{3.2}$$

where

$$c_1 = \frac{h_1 - 1}{h_1^2\alpha_1 \int_0^\infty te^{-\alpha_1 t}d\mathbb{P}(L_1 \le t)}, \qquad c_2 = \frac{\mu - 1}{\mu^2\alpha_2 \int_0^\infty te^{-\alpha_2 t}d\mathbb{P}(L_2 \le t)},$$

$d_1 = c_1h_1\alpha_1'$, $d_2 = c_2\mu\alpha_2'$, and where for the almost sure results concerning $\{G_i(t)\}$ in (3.2), we have assumed that $\int_0^\infty \mathbb{E}[(G_i(t)/(d_i te^{\alpha_i t}) - \mathcal{Z}_i)^2]dt < \infty$.

On top of that, Lemma 3.3.2 enables us to conclude the mean square limit of $G_2(t)$ and the almost sure one under $\mathbb{P}_1$, using reasonings similar to the ones used in Theorem 2.3.11. Moreover, from [116], if $\liminf_{t\to\infty} Z_i^+(t) > 0$, under $\mathbb{P}_i$ we have that almost surely

$$\pi_i(p) := \lim_{t\to\infty} -\frac{1}{pt}\log\left(\frac{Z_i^+(t)}{Z_i(t)}\right) = \begin{cases} (\alpha_1(h_1) - \alpha_1(h_1(1-p)))/p & \text{if } i = 1 \\ (\alpha_2(\mu) - \alpha_2(\mu(1-p)))/p & \text{if } i = 2 \end{cases},$$

and

$$\lim_{p\to 0} \pi_i(p) = \begin{cases} h_1\alpha_1' & \text{if } i = 1 \\ \mu\alpha_2' & \text{if } i = 2 \end{cases},$$

where we have assumed that the first cell is label positive.

We present two sets of results depending on whether $\alpha_1 > \alpha_2$ or vice versa. If $\alpha_1 < \alpha_2$, which would model, for example, the emergence of fast dividing cancer cells from a population of slowly dividing healthy cells [82, 5], the growth rate of the type-2 cells is greater than the type-1 cells and their average generation is determined by the derivative of the latter Malthus parameter.

**Proposition 3.3.3** ($\alpha_1 < \alpha_2$)**.** *If $(\partial/(\partial x_i \partial x_j))\rho_1(1,1)$, for $1 \le i \le j \le 2$, and $(\partial/(\partial x)^2)\rho_2(1)$ are finite, we have that*

$$\lim_{t\to\infty} \frac{Z_2(t)}{c_{1,2}e^{\alpha_2 t}} \overset{L^2,a.s.}{=} \mathcal{W}, \quad \text{and} \quad \lim_{t\to\infty} \frac{G_2(t)}{d_{1,2}te^{\alpha_2 t}} \overset{L^2}{=} \mathcal{W} \quad \text{under } \mathbb{P}_1, \tag{3.3}$$

*where*

$$c_{1,2} = \frac{h_2 c_2 \int_0^\infty e^{-\alpha_2 t} d\mathbb{P}(L_1 \le t)}{1 - h_1 \int_0^\infty e^{-\alpha_2 t} d\mathbb{P}(L_1 \le t)}, \qquad d_{1,2} = c_{1,2}\mu\alpha_2', \tag{3.4}$$

*and $\mathcal{W}$ is a non-negative random variable such that we have $\mathbb{P}_1(\mathcal{W} = 0) = \mathbb{P}_1(\lim_{t\to\infty} Z_1(t) = 0, \lim_{t\to\infty} Z_2(t) = 0)$ and $\mathbb{E}_1(\mathcal{W}) = 1$.*

*If $\int_0^\infty \mathbb{E}_1[(G_2(t)/(d_{1,2}te^{\alpha_2 t}) - \mathcal{W})^2]dt < \infty$, the second limit in (3.3) is also true almost surely. Assuming the initial cell is of type-1, i.e. $Z_1^+(0) = 1$ and $Z_2(0) = G_1(0) = G_2(0) = 0$, $\mathbb{E}(N^2) < \infty$, and lim $inf_{t\to\infty} Z_2^+(t) > 0$, we have almost surely that*

$$\lim_{t\to\infty} \frac{G_2(t)}{tZ_2(t)} = \mu\alpha_2', \quad \lim_{t\to\infty} -\frac{1}{pt}\log\left(\frac{Z_2^+(t)}{Z_2(t)}\right) = \frac{\alpha_2(\mu) - \alpha_2(\mu(1-p))}{p}, \tag{3.5}$$

*and*

$$\lim_{p\to 0} \frac{\alpha_2(\mu) - \alpha_2(\mu(1-p))}{p} = \mu\alpha_2'.$$

**Proof:** The first affirmation is a consequence of [54], Lemma 3.3.2 through an adaptation of Theorem 2.3.9. The second part of the statement can be proved adapting Theorems 2.3.10 and 2.3.11 to the two-type Bellman-Harris branching process. □

If $\alpha_2 < \alpha_1$, as might occur with the production of terminally differentiated cells from multipotent ones [20, pg. 700] (e.g. the hematopoiesis process described in Section 1.1), the growth rate of the type-1 cells is greater than the type-2 cells and their average generation is determined by the derivative of the former Malthus parameter. That is, in this setting, so long as the type-1 population continues to exist, the average generation of the type-2 cells is dominated by immigrants from the type-1 population.

**Proposition 3.3.4** ($\alpha_2 < \alpha_1$)**.** *If* $(\partial/(\partial x_i \partial x_j))\rho_1(1,1)$*, for* $1 \leq i \leq j \leq 2$*, and* $(\partial/(\partial x)^2)\rho_2(1)$ *are finite, we have that*

$$\lim_{t\to\infty} \frac{Z_2(t)}{c_{2,1}e^{\alpha_1 t}} \overset{L^2,a.s.}{=} \mathcal{Z}_2 \quad \textit{and} \quad \lim_{t\to\infty} \frac{G_2(t)}{d_{2,1}te^{\alpha_1 t}} \overset{L^2}{=} \mathcal{Z}_2 \quad \textit{under } \mathbb{P}_1, \tag{3.6}$$

*where*

$$c_{2,1} = \frac{h_2(1 - \int_0^\infty e^{-\alpha_1 t} d\mathbb{P}(L_2 \leq t))}{h_2^2 \alpha_1 (1 - \mu \int_0^\infty e^{-\alpha_1 t} d\mathbb{P}(L_2 \leq t))}, \qquad d_{2,1} = c_{2,1} h_1 \alpha_1', \tag{3.7}$$

*and* $\mathcal{Z}_2$ *random variable defined in* (3.2) *with* $\mathbb{P}_1(\mathcal{Z}_2 = 0) = \mathbb{P}_1(\lim_{t\to\infty} Z_1(t) = 0)$ *and* $\mathbb{E}_1(\mathcal{Z}_2) = 1$*.*

*If* $\int_0^\infty \mathbb{E}_1[(G_2(t)/(d_{2,1}te^{\alpha_2 t}) - \mathcal{Z}_2)^2]dt < \infty$*, the second limit in* (3.6) *is also true almost surely. Assuming the initial cell is of type-1, i.e.* $Z_1^+(0) = 1$ *and* $Z_2(0) = G_1(0) = G_2(0) = 0$*,* $\mathbb{E}(N^2) < \infty$*, and lim inf*$_{t\to\infty} Z_1^+(t) > 0$*, we have almost surely that*

$$\lim_{t\to\infty} \frac{G_2(t)}{tZ_2(t)} = h_1\alpha_1', \quad \lim_{t\to\infty} -\frac{1}{pt} \log\left(\frac{Z_2^+(t)}{Z_2(t)}\right) = \frac{\alpha_1(h_1) - \alpha_1(h_1(1-p))}{p}, \tag{3.8}$$

*and*

$$\lim_{p\to 0} \frac{\alpha_1(h_1) - \alpha_1(h_1(1-p))}{p} = h_1\alpha_1'.$$

**Proof:** The proof follows the same lines of Proposition 3.3.3 and so is omitted. □

**Example 3.3.1** (Exponential lifetimes)**.** As made in Examples 2.3.1, we compute the constants that appear in Proposition 3.3.3 and Proposition 3.3.4 when the lifetimes of the cells are exponential distributed. Assume $L_1 \overset{D}{\sim}\text{Exp}(\lambda_1)$ and $L_2 \overset{D}{\sim}\text{Exp}(\lambda_2)$. Given that for exponential lifetimes we have $\alpha_1(h_1) = \lambda_1(h_1-1)$ and $\alpha_2(\mu) = \lambda_2(\mu-1)$, we obtain that $\alpha_1' = \lambda_1$ and $\alpha_2' = \lambda_2$. So, the quantity $\mu\alpha_2'$ and $h_1\alpha_1'$ that appear in (3.5) and (3.8) respectively, are given by $\mu\lambda_2$ and $h_1\lambda_1$. Using the fact that in the case considered $c_1 = 1 = c_2$, we have also that

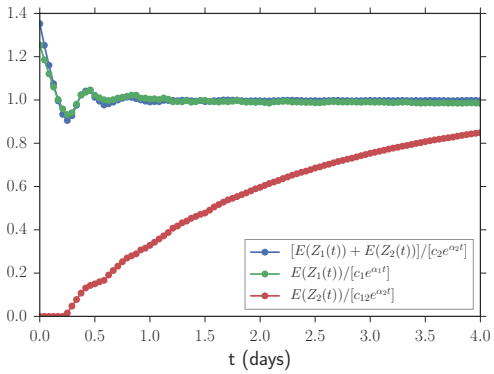
$$c_{1,2} = \frac{h_2\lambda_1}{\lambda_2(\mu-1) - \lambda_1(h_1-1)}, \qquad c_{2,1} = \frac{1}{h_2[\lambda_1(h_1-1) - \lambda_2(\mu-1)]},$$

$d_{1,2} = c_{1,2}\mu\lambda_2$, and $d_{2,1} = c_{2,1}h_1\lambda_1$.

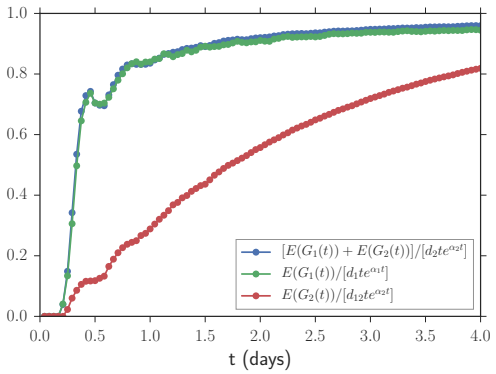
We conclude the chapter by presenting some simulated results that illustrate the features of these two-type results, both for average generation and for its inference. Fig. 3.1 provides average normalised paths of the processes $Z_i(t)$ and $G_i(t)$. In Fig. 3.1(a)-(b), $\alpha_1 < \alpha_2$, but despite the fact the type-2 population is the fastest growing on average, it is the slowest one to converge. This occurs due to the random delay in the production of any type-2 cells. Note also that the total population of both type-1 and type-2 cells behave as a single-type branching process with $N = 2$ and log-normal lifetime distribution. Hence, the growth rates of $Z(t) = Z_1(t) + Z_2(t)$ and $G(t) = G_1(t) + G_2(t)$ are the same as if the type-2 population was started with one type-2 cell.

In Fig. 3.1(c)-(d), $\alpha_1 > \alpha_2$. Here, the second population is dominated by differentiation from the first cell type, with both populations have the growth rate of the type-1 population. The behaviour of $Z(t)$ and $G(t)$ for the entire population is the sum of the corresponding processes for the two types.

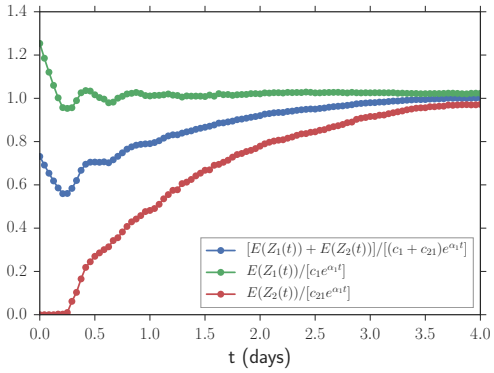
Turning to the relatedness in random prefactors, Fig. 3.2(b) is consistent with the deduction that there is equality almost surely between the rescaled limit of the population size and total generation of the second type. Fig. 3.2(c) shows the prefactor for type-1 and type-2 population sizes. Consistent with results in [53], red dots are suggestive that when $\alpha_2 < \alpha_1$ both normalised processes converge to the same random variable. For $\alpha_1 < \alpha_2$, however, this is not the case for the blue dots and the random variables appear uncorrelated. Fig. 3.2(d) is analogous to Fig. 3.2(c) but for the total generation process, with



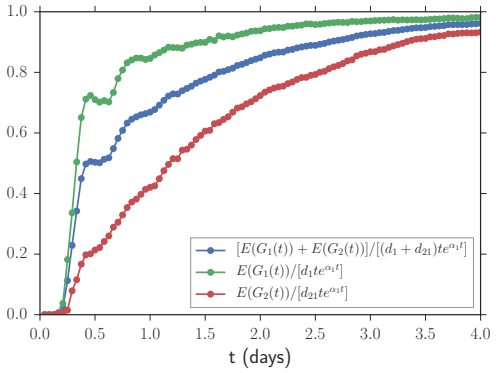

(a) normalised populations size ($\alpha_1 < \alpha_2$)




(b) normalised total generations ($\alpha_1 < \alpha_2$)




(c) normalised populations size ($\alpha_2 < \alpha_1$)




(d) normalised total generations ($\alpha_2 < \alpha_1$)

Figure 3.1: **Average growth rates of population sizes and total generations of each type starting with a single type-1 cell and using the scalings in Propositions 3.3.3 and 3.3.4**. Cells have lognormal lifetime with mean 9.3 hours and standard deviation 2.54 [44]. Type-1 cells give rise to type-1 cells with probability 5/6 and to type-2 cells with probability 1/6. Means are computed averaging the results of 1000 Monte Carlo simulations of populations growing for four days. (a)-(b) These illustrations are in the case $\alpha_1 < \alpha_2$ as both types of cells always have two offspring. (c)-(d) These are in the setting $\alpha_2 < \alpha_1$, obtained by setting $N_1 = 2$ and $\mathbb{P}(N_2 = 0) = 2/5 = 1 - \mathbb{P}(N_2 = 2)$.

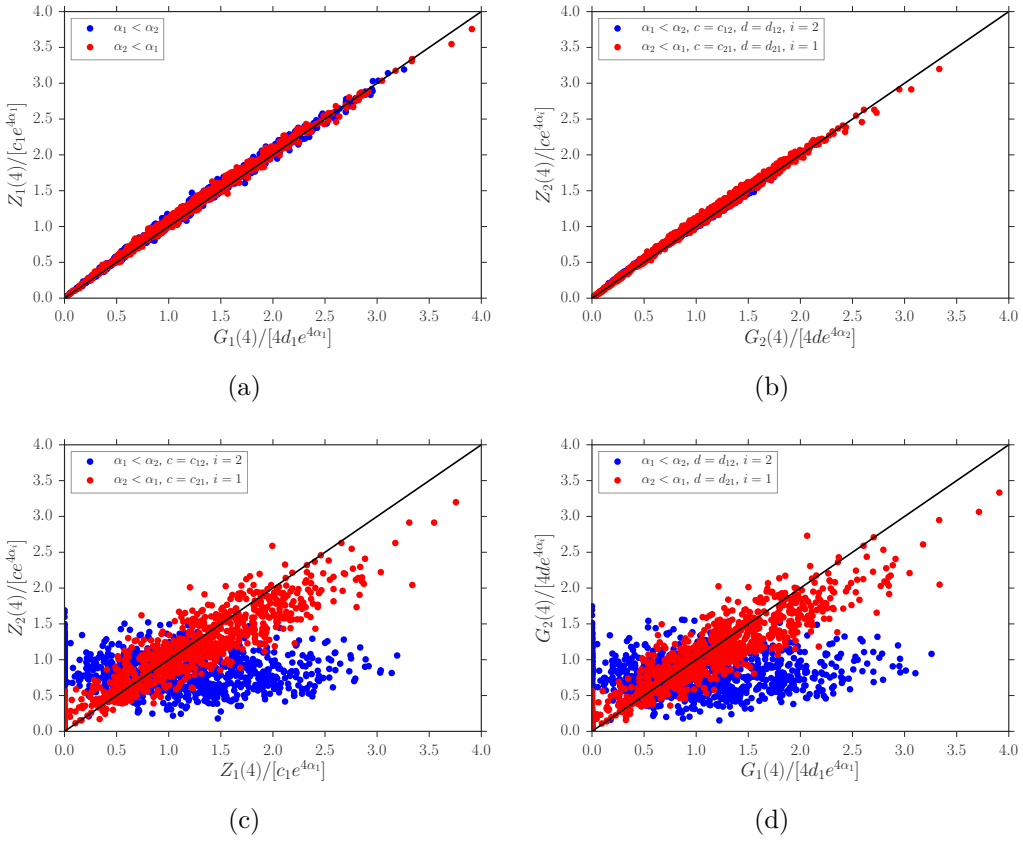
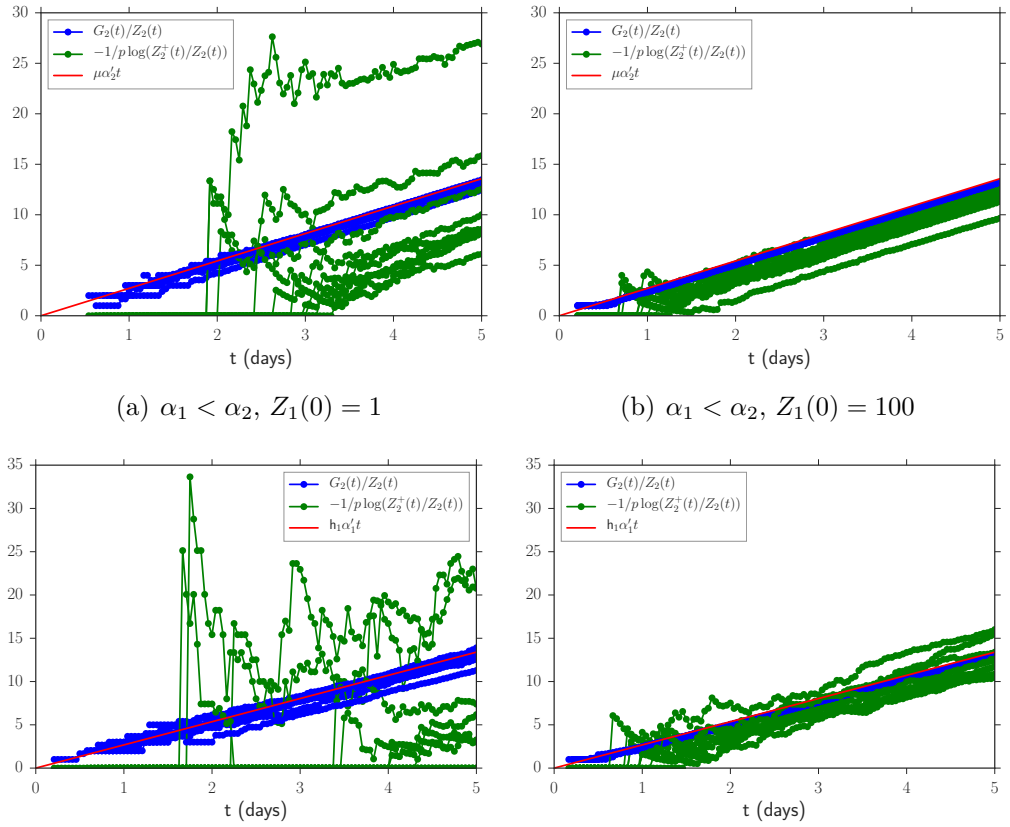



Figure 3.2: **Relationships in per-path randomness.** Plots were created using the same 1000 Monte Carlo simulations used to generate Fig. 3.1. Blue points correspond to $\alpha_1 < \alpha_2$, while red ones to $\alpha_2 < \alpha_1$. (a) Scatter plot of normalised versions of $Z_1(t)$ and $G_1(t)$ is displayed at $t = 4$ days. Pearson correlation coefficient for both blue and red points is 0.99. Similar situation in (b), where normalised versions of $Z_2(t)$ and $G_2(t)$ are plotted at time $t = 4$. Even in this case, the Pearson correlation coefficient for both blue and red points is 0.99. (c) We now compare the randomness in the limit behaviour of the size of the two populations, i.e. we display the normalised versions of $Z_1(t)$ and $Z_2(t)$ at $t = 4$ days in a scatterplot. Pearson correlation coefficient for blue and red points is $-0.19$ and 0.94, respectively. (d) In a similar way, a scatter plot of normalised versions of $G_1(t)$ and $G_2(t)$ is displayed at $t = 4$ days. Pearson correlation coefficient for blue and red points is $-0.09$ and 0.94, respectively.




(a) $\alpha_1 < \alpha_2$, $Z_1(0) = 1$

(b) $\alpha_1 < \alpha_2$, $Z_1(0) = 100$

(c) $\alpha_2 < \alpha_1$, $Z_1(0) = 1$

(d) $\alpha_2 < \alpha_1$, $Z_1(0) = 100$

Figure 3.3: **Sample-path estimation of average generation.** For each sub-panel, ten Monte Carlo simulations of a two-type population are presented. These employ the same parameterisation in Fig. 3.1, with the exception of the initial population size in the two right hand side panels. Each initial cell is equipped with a neutral label that doesn't alter population dynamics, and which is lost irrevocably to all subsequent offspring with probability $p = 10^{-2}$ per cell division. The red line indicates the theoretical prediction of the mean average generation. Blue lines indicate the development of the per-path average generation, while the green lines are the estimates from the delabelling formula (1.1). (a-b) Plots are in the setting $\alpha_1 < \alpha_2$ case, but start with one and 100 type-1 cells at $t = 0$, respectively. (c-d) Equivalent of (a-b) but with $\alpha_2 < \alpha_1$.

the same deduction as for the population size holding where when $\alpha_1 > \alpha_2$, the randomness is common to both types and otherwise it is not.

Part of the significance of Propositions 3.3.3 and 3.3.4 is that they provide an instrument by which one can infer the average generation of each of the populations in a two-type Bellman-Harris branching process, generalising the

results in [116, Proposition 2]. In the presence of cells equipped with a neutral label that is heritably lost with a fixed probability at each division, the average generation and a function of the proportion of label-positive cells of each type share the same dominant term. The mathematical results say that the slope of the average generation and the slope of the estimator are the same when the probabilistic regularity of a large population takes hold. Figs 3.3(a) and 3.3(c) illustrate this relationship for the type-2 population via the use of some Monte Carlo simulations in the presence of a single initial label positive cell of type-1. In this setting the large population regularity only takes hold at later times. Starting with more than one initially labelled cell, illustrated with 100 in Figs 3.3(b) and 3.3(d), results in the desired asymptotic equivalence occurring at a much earlier time. For true cellular systems, the cell numbers are likely to be greater than that. For example, if the two-type branching process is describing the production of effector T-cells from naïve ones during an adaptive immune response in a mouse, we have that the initial population of pathogen-specific naïve T-cells is around 100-1000. This can be 1000 times bigger in a secondary adaptive immune response [62] (see Fig. 1.4).

CHAPTER 4

# A closer look at the variance of the average generation of individual trees

## 4.1 Introduction

In Chapter 2 we have shown that modelling a growing cell population as a super-critical Bellman-Harris branching process allows us to establish how the average generation of the populations behaves over time. In particular, if $Z(t)$ and $G(t)$ denote the size and the total generation of the population at instant $t$ respectively, thanks to the almost sure result contained in Corollary 2.3.12, we have that for large $t$

$$\frac{G(t)}{Z(t)} \approx \alpha'(h)ht + o(t) \tag{4.1}$$

where $h$ is the average number offspring generated after division by each cell, and $\alpha(h)$, defined in (2.4), represents the exponential growth rate of the size of the population.

According to [116, Theorem 1] and Corollary 2.3.12, we also have that the first term in (4.1) can be estimated using the DNA randomised algorithm proposed in [116] and described in Section 1.4, that is

$$\lim_{p\to 0}\lim_{t\to\infty} -\frac{1}{pt}\log\left(\frac{Z^+(t)}{Z(t)}\right) = \alpha'(h)h,$$

where $Z^+(t)$ denotes the number of label-positive cells alive at time $t$, and where $p$ is the probability of label loss for each cell just before its division occurs.

Having information on the $o(t)$ function that appears in (4.1) is important to better understand the estimates provided by estimator (1.1). Studying this term is not easy under the general hypothesis of a Bellman-Harris branching process and for this reason in this chapter we assume a specific scenario, simplifying the model by making restrictions to specific lifetime and offspring number distributions. In particular, let's assume that the lifetime of the cell, $L$, is exponentially distributed, i.e. $L \overset{D}{\sim} Exp(\lambda)$, and that the per cell number of offspring, $N$, is s.t. $\mathbb{P}(N = 2) = 1$. This model, which is a special case of Bellman-Harris branching process, is referred in the literature as Pure Birth process (e.g. [93, pg. 370]).

Pure Birth processes, apart from population dynamics, appear as a fundamental model of study in a large number of applications from data-structures in computer science to likelihood methods in phylogenetics to the study of random walkers on random graphs.

Thanks to the assumption made for the lifetime distribution, the process $Z(t)$ becomes a continuous-time Markov Chain on the discrete state space $\mathbb{N}$ (see [42, Chapter 5] and [65, Section 17.3]). The same is not true for the process $G(t)$, whose value at time $t$ is not enough to determine the distribution of the process at instant $t + \tau$, the time of the following division. We can retrieve this Markovian property for the generation of the cells by considering the vector of the composition of the living cells, i.e. $\vec{g}(t) := (g_0(t), g_1(t), \ldots)$ where $g_i(t)$, $i \in \mathbb{N}$, denotes the number of cells in generation $i$ alive at time $t$. Indeed, assuming that the composition vector is at state $\vec{x} := (x_1, x_2, \ldots)$ at time $t$, i.e. $\vec{g}(t) = \vec{x}$, we have $Z(t) = \sum_{i \in \mathbb{N}} x_i$, and the exponential lifetime assumption tells us that the next change of state will happen after an exponential time with parameter $\lambda \sum_{i \in \mathbb{N}} x_i$, the minimum of $\sum_{i \in \mathbb{N}} x_i$ cell lifetimes exponentially distributed with parameter $\lambda$. Furthermore, once a division occurs, the composition vector can only step from state $\vec{x}$ to one of the states $\vec{x} + 2\vec{e}_{i+1} - \vec{e}_i$ for some $i \in \mathbb{N}$, where we denote with $\vec{e}_i$ the vector with a 1 in position $i$ and 0 otherwise. Joining these results, we obtain that $\vec{g}(t)$ is a

continuous-time Markov chain with rate transition matrix given by

$$q_{\vec{x},\vec{y}} = \begin{cases} \lambda x_i & \text{if } \vec{x} \neq \vec{y} \text{ and } \vec{y} = \vec{x} + 2\vec{e}_{i+1} - \vec{e}_i \text{ for some } i \in \mathbb{N}, \\ 0 & \text{if } \vec{x} \neq \vec{y} \text{ but } \vec{y} \neq \vec{x} + 2\vec{e}_{i+1} - \vec{e}_i \text{ for all } i \in \mathbb{N}, \\ -\sum_{\vec{y},\vec{y}\neq\vec{x}} q_{\vec{x},\vec{y}} & \text{if } \vec{x} = \vec{y}. \end{cases}$$

Even though we are not exactly going to study the vector $\vec{g}(t)$ in Section 4.2, our approach strongly uses its Markovian property. Thanks to that, the study of the quantity $G(t)/Z(t)$ becomes easier when we condition on the number of divisions occurred in the timeframe $[0, t]$. In particular, in the next section we prove that, for a pure birth process, the quantity $o(t)$ in (4.1) is a random quantity that does not depend on $t$. We show that in the main result of the chapter, that we state in the following proposition and of which we can have a visualisation in Fig. 4.1, obtained by Monte Carlo simulation.

**Proposition 4.1.1.** *For a pure birth process, we have that*

$$\lim_{t\to\infty} Var\left(\frac{G(t)}{Z(t)}\right) = 7.$$

This result is potentially surprising because, according to [101], if we randomly select a living cell in the population at time $t$, its average generation is asymptotically normally distributed with a mean and a variance that grow linearly in time, which might lead one to expect the same of $\mathrm{Var}\,(G(t)/Z(t))$. Furthermore, it is known that the two processes $\{Z(t)\}$ and $\{G(t)\}$ have different growth rates, $e^{\lambda t}$ and $te^{\lambda t}$, respectively, [54, 116] from which one might expect that the variability of the average generation scales as $t^2$ and so be diverging to infinity. Both of these hypotheses are incorrect as $Z(t)$ and $G(t)$ are strongly correlated at the level of sample paths, according to our findings in Chapter 2. Also note that the result in Proposition 4.1.1 does not depend on $\lambda$, the rate at which a cell divides into two new cells; $\lambda$ only influences the speed of convergence in the result.

In the next section, in order to evaluate $\mathrm{Var}\,(G(t)/Z(t))$, we condition the average generation $G(t)/Z(t)$ on the number of living cells at time $t$, $Z(t)$. By the Law of Total Variance (e.g. [14, Theorem 9.5.4])

$$\mathrm{Var}\left(\frac{G(t)}{Z(t)}\right) = \mathbb{E}\left(\mathrm{Var}\left(\frac{G(t)}{Z(t)}\middle| Z(t)\right)\right) + \mathrm{Var}\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}\middle| Z(t)\right)\right) \tag{4.2}$$

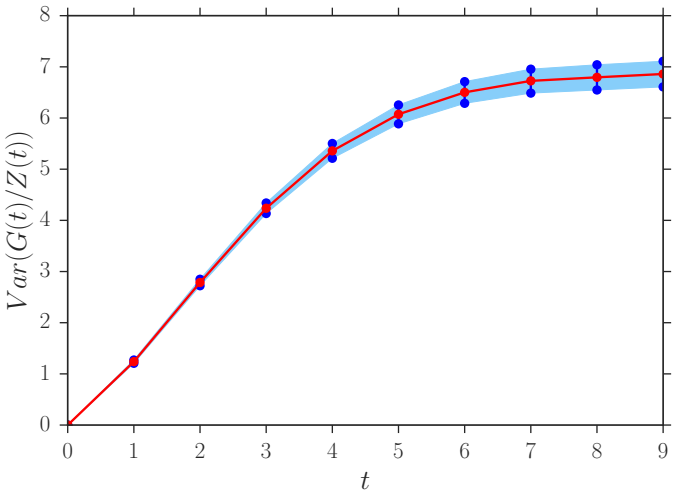



Figure 4.1: **Convergence of variance of the average generation of a cell population in a pure birth process.** From $10^4$ Monte Carlo simulations of a pure birth process with $\lambda = 1$, the red line in the plot denotes the variance of the average generation of the population at time $t$, i.e. $\mathrm{Var}\left(G(t)/Z(t)\right)$, whereas the blue shaded region indicates a 95% confidence interval based on bootstrap percentiles [28, Chapter 13].

and, in order to study the variance of the average generation of the population at time $t$, we study the quantities $\mathbb{E}\left(G(t)/Z(t)|Z(t)\right)$ and $\mathrm{Var}\left(G(t)/Z(t)|Z(t)\right)$ in Theorems 4.2.2 and 4.2.3, respectively. Proposition 4.1.1 then follows.

## 4.2 Results

Before proceeding with the analysis of the two terms on the RHS of (4.2), we prove a lemma that will simplify the proofs of Theorems 4.2.2 and 4.2.3. For that, we introduce a new process, $\{S(t)\}$, denoting the sum of the squares of the generations of the living cells at time $t$, which appears when the second moment of $G(t)/Z(t)$ is studied. In the following we also consider the discrete-time process associated with $\{G(t)\}$ and $\{S(t)\}$, $\{G_k\}$ and $\{S_k\}$ accounting for the sum and the sum of the squares of the generations of the living cells, respectively, when the number of cells alive is $k$.

**Lemma 4.2.1.** *We have that*

- $$\mathbb{E}\left(\frac{G(t)}{Z(t)}\middle| Z(t) = k\right) = \frac{\mathbb{E}\left(G_k\right)}{k} = 2\sum_{i=2}^{k}\frac{1}{i}, \tag{4.3}$$

- $$\frac{\mathbb{E}\left(S_k\right)}{k} = 4\sum_{i=2}^{k-1}\frac{\mathbb{E}\left(G_i\right)}{i(i+1)} + \frac{\mathbb{E}\left(G_k\right)}{k}, \tag{4.4}$$
- $$\mathbb{E}\left(\frac{G(t)^2}{Z(t)^2}\middle| Z(t)=k\right) = \frac{\mathbb{E}\left(G_k^2\right)}{k^2}$$
$$= \frac{k+1}{k}\left(\sum_{i=1}^{k-1}\frac{\mathbb{E}\left(S_i\right)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{1}{(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{\mathbb{E}\left(G_i\right)}{i(i+2)}\right). \tag{4.5}$$

**Proof:** Throughout this proof, we condition on $Z(t) = k$ and denote by $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$ the generation of the $k$ cells present at time $t$, which are not independent. From the definitions, we have $G_k := \sum_{i=1}^k \Gamma_i$ and $S_k := \sum_{i=1}^k \Gamma_i^2$. The idea of the proof is to recover the formulas given above by finding recurrence equations for $\mathbb{E}\left(G_k\right), \mathbb{E}\left(S_k\right)$, and $\mathbb{E}\left(G_k^2\right)$.

For $j \in \{1, 2, \ldots, k\}$, denote by $I_j$ a random variable that takes value 1 if the $j$-th cell is the first one, among the $k$ existing, to divide in two new cells, and 0 otherwise. We have that for fixed $j \in \{1, 2, \ldots, k\}$, $I_j$ is independent of $\{\Gamma_1, \ldots, \Gamma_k\}$ and, due to the memoryless property of the exponential distribution, $\mathbb{P}\left(I_j = 1\right) = 1/k$ for all $j \in \{1, 2, \ldots, k\}$, with $k$ the number of cells in the population. Furthermore, the $I_j$ are not independent of each other because only one of them can assume value 1, i.e. $\sum_{j=1}^k I_j = 1$, implying that $I_j^2 = I_j$ and $I_j I_\ell = 0$ if $j \neq \ell$. With that in mind, we establish the following relations

$$G_{k+1} = G_k + \sum_{j=1}^k I_j\Gamma_j + 2, \qquad S_{k+1} = S_k + \sum_{j=1}^k I_j\left(2(\Gamma_j+1)^2 - \Gamma_j^2\right), \tag{4.6}$$

$$G_{k+1}^2 = G_k^2 + \left(\sum_{j=1}^k I_j\Gamma_j\right)^2 + 4 + 4G_k + 4\sum_{j=1}^k I_j\Gamma_j + 2G_k\sum_{j=1}^k I_j\Gamma_j$$
$$= G_k^2 + \sum_{j=1}^k I_j\Gamma_j^2 + 4 + 4G_k + 4\sum_{j=1}^k I_j\Gamma_j + 2G_k\sum_{j=1}^k I_j\Gamma_j. \tag{4.7}$$

From the first equation in (4.6) we obtain

$$\mathbb{E}\left(G_{k+1}\right) = \mathbb{E}\left(G_k\right) + \sum_{j=1}^k \mathbb{E}\left(I_j\Gamma_j\right) + 2 = \mathbb{E}\left(G_k\right) + \sum_{j=1}^k \mathbb{E}\left(I_j\right)\mathbb{E}\left(\Gamma_j\right) + 2$$
$$= \mathbb{E}\left(G_k\right) + \frac{1}{k}\mathbb{E}\left(\sum_{j=1}^k \Gamma_j\right) + 2 = \mathbb{E}\left(G_k\right) + \frac{1}{k}\mathbb{E}\left(G_k\right) + 2 = \frac{k+1}{k}\mathbb{E}\left(G_k\right) + 2,$$

where we have used that $I_j$ and $\Gamma_j$ are independent. This gives the following recurrence relation

$$\frac{\mathbb{E}\left(G_{k+1}\right)}{k+1} = \frac{\mathbb{E}\left(G_k\right)}{k} + \frac{2}{k+1},$$

that, solved with initial condition $\mathbb{E}\left(G_1\right) = 0$, results in (4.3).

Similarly, using the second equation in (4.6), we have that

$$\begin{aligned}
\mathbb{E}\left(S_{k+1}\right) &= \mathbb{E}\left(S_k\right) + \sum_{j=1}^{k} \mathbb{E}\left(I_j\right) \mathbb{E}\left(2(\Gamma_j+1)^2 - \Gamma_j^2\right) \\
&= \mathbb{E}\left(S_k\right) + \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}\left(\Gamma_j^2 + 4\Gamma_j + 2\right) \\
&= \mathbb{E}\left(S_k\right) + \frac{1}{k}\mathbb{E}\left(S_k\right) + \frac{4}{k}\mathbb{E}\left(G_k\right) + 2 = \frac{k+1}{k}\mathbb{E}\left(S_k\right) + \frac{4}{k}\mathbb{E}\left(G_k\right) + 2,
\end{aligned}$$

from which we get the recurrence equation

$$\frac{\mathbb{E}\left(S_{k+1}\right)}{k+1} = \frac{\mathbb{E}\left(S_k\right)}{k} + \frac{4}{k(k+1)}\mathbb{E}\left(G_k\right) + \frac{2}{k+1}.$$

Solving the above equation for $\mathbb{E}\left(S_1\right) = \mathbb{E}\left(G_1\right) = 0$, we obtain the relation in (4.4).

Using (4.7) and the two results just found, i.e. the formulas in (4.3) and (4.4), we can now find an expression for $\mathbb{E}\left(G_k^2\right)$.

$$\begin{aligned}
\mathbb{E}\left(G_{k+1}^2\right) =& \mathbb{E}\left(G_k^2\right) + \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}\left(\Gamma_j^2\right) + 4 + 4\mathbb{E}\left(G_k\right) + \frac{4}{k}\sum_{j=1}^{k}\mathbb{E}\left(\Gamma_j\right) \\
&+ 2\mathbb{E}\left(G_k\Big(\sum_{j=1}^{k} I_j\Gamma_j\Big)\right) \\
=& \mathbb{E}\left(G_k^2\right) + \frac{1}{k}\mathbb{E}\left(S_k\right) + 4 + \left(4 + \frac{4}{k}\right)\mathbb{E}\left(G_k\right) + \frac{2}{k}\mathbb{E}\left(G_k\sum_{j=1}^{k}\Gamma_j\right) \\
=& \mathbb{E}\left(G_k^2\right) + \frac{\mathbb{E}\left(S_k\right)}{k} + 4 + \frac{4(k+1)}{k}\mathbb{E}\left(G_k\right) + \frac{2}{k}\mathbb{E}\left(G_k^2\right) \\
=& \frac{k+2}{k}\mathbb{E}\left(G_k^2\right) + \frac{\mathbb{E}\left(S_k\right)}{k} + 4 + \frac{4(k+1)}{k}\mathbb{E}\left(G_k\right).
\end{aligned}$$

The equation above can be rewritten as the recurrence equation

$$\frac{\mathbb{E}\left(G_{k+1}^2\right)}{(k+1)(k+2)} = \frac{\mathbb{E}\left(G_k^2\right)}{k(k+1)} + \frac{\mathbb{E}\left(S_k\right)}{k(k+1)(k+2)} + \frac{4}{(k+1)(k+2)} + \frac{4\mathbb{E}\left(G_k\right)}{k(k+2)},$$

that, when solved with initial condition $\mathbb{E}\left(G_1\right) = \mathbb{E}\left(G_1^2\right) = \mathbb{E}\left(S_1\right) = 0$, gives (4.5). □

We now use Lemma 4.2.1 to study the limit behaviour of the first term in the RHS of (4.2).

**Theorem 4.2.2.** *For a pure birth process, we have that*

$$\lim_{t\to\infty} \mathbb{E}\left( Var\left( \frac{G(t)}{Z(t)} \middle| Z(t) \right) \right) = 7 - \frac{2}{3}\pi^2.$$

**Proof:** Given that $\lim_{t\to\infty} Z(t) = \infty$ a.s. [42, Chapter 5], for every fixed $k \in \mathbb{N}$ we have that $\lim_{t\to\infty} \mathbb{P}\left(Z(t) = k\right) = 0$. This implies that

$$\begin{aligned}\lim_{t\to\infty} \mathbb{E}\left( \mathrm{Var}\left( \frac{G(t)}{Z(t)} \middle| Z(t) \right) \right) &= \lim_{t\to\infty} \sum_{k=1}^{\infty} \mathrm{Var}\left( \frac{G_k}{k} \middle| Z(t) = k \right) \mathbb{P}\left(Z(t) = k\right) \\ &= \lim_{k\to\infty} \mathrm{Var}\left( \frac{G_k}{k} \right).\end{aligned}$$

Using Lemma 4.2.1, we can now compute this variance:

$$\begin{aligned}&\mathrm{Var}\left( \frac{G_k}{k} \right) = \frac{\mathbb{E}\left(G_k^2\right)}{k^2} - \frac{\mathbb{E}\left(G_k\right)^2}{k^2} \\ &= \frac{k+1}{k}\left[ \sum_{i=1}^{k-1} \frac{\mathbb{E}\left(S_i\right)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1} \frac{1}{(i+1)(i+2)} + 4\sum_{i=1}^{k-1} \frac{\mathbb{E}\left(G_i\right)}{i(i+2)} \right] - \frac{\mathbb{E}\left(G_k\right)^2}{k^2} \\ &= \frac{k+1}{k}\left[ \sum_{i=1}^{k-1} \frac{\mathbb{E}\left(S_i\right)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1} \frac{1}{(i+1)(i+2)} + 4\sum_{i=1}^{k} \frac{\mathbb{E}\left(G_i\right)}{i^2} \right. \\ &\quad \left. + 4\sum_{i=1}^{k-1} \left( \frac{\mathbb{E}\left(G_i\right)}{i(i+2)} - \frac{\mathbb{E}\left(G_i\right)}{i^2} \right) - 4\frac{\mathbb{E}\left(G_k\right)}{k^2} \right] - \frac{\mathbb{E}\left(G_k\right)^2}{k^2} \\ &= \frac{k+1}{k}\left[ \sum_{i=1}^{k-1} \frac{\mathbb{E}\left(S_i\right)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1} \frac{1}{(i+1)(i+2)} + 4\sum_{i=2}^{k} \frac{1}{i^2} + \frac{\mathbb{E}\left(G_k\right)^2}{k^2} \right. \\ &\quad \left. - 8\sum_{i=1}^{k-1} \frac{\mathbb{E}\left(G_i\right)}{i^2(i+2)} - 4\frac{\mathbb{E}\left(G_k\right)}{k^2} \right] - \frac{\mathbb{E}\left(G_k\right)^2}{k^2} \\ &= \frac{k+1}{k}\left[ \sum_{i=1}^{k-1} \frac{\mathbb{E}\left(S_i\right)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1} \frac{1}{(i+1)(i+2)} + 4\sum_{i=2}^{k} \frac{1}{i^2} \right. \\ &\quad \left. - 8\sum_{i=1}^{k-1} \frac{\mathbb{E}\left(G_i\right)}{i^2(i+2)} - 4\frac{\mathbb{E}\left(G_k\right)}{k^2} \right] + \frac{\mathbb{E}\left(G_k\right)^2}{k^3},\end{aligned}$$

where in the third equality we have added and subtracted the quantity

$$\begin{aligned}4\sum_{i=1}^{k} \frac{\mathbb{E}\left(G_i\right)}{i^2} &= 8\sum_{i=2}^{k} \frac{1}{i} \sum_{j=2}^{i} \frac{1}{j} = 8\sum_{i=2}^{k} \frac{1}{i^2} + 8\sum_{i=2}^{k} \sum_{j=2}^{i-1} \frac{1}{ij} \\ &= 8\sum_{i=2}^{k} \frac{1}{i^2} + 4\left( \left( \sum_{i=2}^{k} \frac{1}{i} \right)^2 - \sum_{i=2}^{k} \frac{1}{i^2} \right) = 4\sum_{i=2}^{k} \frac{1}{i^2} + \frac{\mathbb{E}\left(G_k\right)^2}{k^2}.\end{aligned}$$

Taking the limit as $k \to \infty$, we have that

$$\begin{aligned}\lim_{k\to\infty} \operatorname{Var}\left(\frac{G_k}{k}\right) &= \sum_{i=1}^{\infty} \frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 4\sum_{i=1}^{\infty} \frac{1}{(i+1)(i+2)} \\ &\quad + 4\sum_{i=2}^{\infty} \frac{1}{i^2} - 8\sum_{i=1}^{\infty} \frac{\mathbb{E}(G_i)}{i^2(i+2)} \\ &= \sum_{i=1}^{\infty} \frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 2 + 4\left(\frac{\pi^2}{6} - 1\right) - 8\sum_{i=1}^{\infty} \frac{\mathbb{E}(G_i)}{i^2(i+2)}. \quad (4.8)\end{aligned}$$

Using Lemma 4.2.1, the first term in the RHS of (4.8) becomes

$$\begin{aligned}&\sum_{i=1}^{\infty} \frac{1}{(i+1)(i+2)} \left(4\sum_{k=2}^{i-1} \frac{1}{k+1} \frac{\mathbb{E}(G_k)}{k} + \frac{\mathbb{E}(G_i)}{i}\right) \\ &\quad = 8\sum_{i=1}^{\infty} \frac{1}{(i+1)(i+2)} \sum_{k=2}^{i-1} \frac{1}{k+1} \sum_{j=2}^{k} \frac{1}{j} + 2\sum_{i=1}^{\infty} \frac{1}{(i+1)(i+2)} \sum_{k=2}^{i} \frac{1}{k}. \quad (4.9)\end{aligned}$$

The first term in the RHS of (4.9) is given by

$$\sum_{j=2}^{\infty} \frac{8}{j} \sum_{k=j}^{\infty} \frac{1}{k+1} \sum_{i=k+1}^{\infty} \frac{1}{(i+1)(i+2)} = \sum_{j=2}^{\infty} \frac{8}{j} \sum_{k=j}^{\infty} \frac{1}{k+1} \frac{1}{k+2} = 8\sum_{j=2}^{\infty} \frac{1}{j} \frac{1}{j+1} = 4,$$

whereas the second one is given by

$$2\sum_{k=2}^{\infty} \frac{1}{k} \sum_{i=k}^{\infty} \frac{1}{(i+1)(i+2)} = 2\sum_{k=2}^{\infty} \frac{1}{k} \frac{1}{k+1} = 1.$$

So, the first sum in the RHS of (4.8) is equal to $4 + 1 = 5$. For the last sum in the RHS of (4.8), we have

$$\begin{aligned}-8\sum_{i=1}^{\infty} \frac{1}{i(i+2)} \frac{\mathbb{E}(G_i)}{i} &= -16\sum_{i=1}^{\infty} \frac{1}{i(i+2)} \sum_{j=2}^{i} \frac{1}{j} = -16\sum_{j=2}^{\infty} \frac{1}{j} \sum_{i=j}^{\infty} \frac{1}{i(i+2)} \\ &= -16\sum_{j=2}^{\infty} \frac{1}{j} \frac{1+2j}{2j(j+1)} = -\frac{4}{3}(\pi^2 - 3).\end{aligned}$$

Joining all these results, we obtain

$$\lim_{k\to\infty} \operatorname{Var}\left(\frac{G_k}{k}\right) = 5 + 2 + 4\left(\frac{\pi^2}{6} - 1\right) - \frac{4}{3}(\pi^2 - 3) = 7 - \frac{2}{3}\pi^2.$$

□

Lemma 4.2.1 allows us to also understand the behaviour of the conditional variance of the expected average generation of the population given its size.

**Theorem 4.2.3.** *For a pure birth process, we have that*

$$\lim_{t\to\infty} Var\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}|Z(t)\right)\right) = \frac{2}{3}\pi^2.$$

**Proof:** From Lemma 4.2.1 we know that

$$\begin{aligned}\mathrm{Var}\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}|Z(t)\right)\right) &= \mathrm{Var}\left(2\sum_{i=2}^{Z(t)}\frac{1}{i}\right) = 4\mathrm{Var}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right)\\ &= 4\left(\mathbb{E}\left(\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right)^2\right) - \left(\mathbb{E}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right)\right)^2\right), \qquad (4.10)\end{aligned}$$

where, in the second equality, we have used the fact that the variance of a process doesn't change when a constant is added. Given $Z(t)$ is a pure birth process, its distribution is given by (e.g. [93, pg. 430])

$$\mathbb{P}\left(Z(t)=k\right) = e^{-\lambda t}(1-e^{-\lambda t})^{k-1}, \quad k=1,2,\ldots$$

where $1/\lambda$ is the expected time before a cell divides into two new leaves, which allows us to evaluate the second term in (4.10) exactly:

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right) &= \sum_{k=1}^{\infty}\mathbb{P}(Z(t)=k)\sum_{i=1}^{k}\frac{1}{i} = \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{k=1}^{\infty}(1-e^{-\lambda t})^k\sum_{i=1}^{k}\frac{1}{i}\\ &= \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{i=1}^{\infty}\frac{1}{i}\sum_{k=i}^{\infty}(1-e^{-\lambda t})^k = \frac{1}{(1-e^{-\lambda t})}\sum_{i=1}^{\infty}\frac{1}{i}(1-e^{-\lambda t})^i.\end{aligned}$$

Let $f(t) := \sum_{i=1}^{\infty}(1-e^{-\lambda t})^i/i$. Then

$$f'(t) = \lambda e^{-\lambda t}\sum_{i=1}^{\infty}\frac{1}{i}i(1-e^{-\lambda t})^{i-1} = \lambda e^{-\lambda t}\sum_{i=1}^{\infty}(1-e^{-\lambda t})^{i-1} = \lambda,$$

and, given $f(0)=0$, we have that $f(t)=\lambda t$. This implies that

$$\mathbb{E}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right) = \frac{\lambda t}{(1-e^{-\lambda t})} = \lambda t + o(1),$$

and the second term in the brackets on the RHS of (4.10) is therefore equal to $(\lambda^2 t^2)/(1-e^{-\lambda t})^2$.

Consider the first term on the RHS of (4.10).

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right)^2\right) &= \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{i=1}^{\infty}\left(\sum_{k=1}^{i}\frac{1}{k}\right)^2(1-e^{-\lambda t})^i\\ &= \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\left(\sum_{i=1}^{\infty}\sum_{k=1}^{i}\frac{1}{k^2}(1-e^{-\lambda t})^i + 2\sum_{i=1}^{\infty}\sum_{k=1}^{i}\sum_{j=1}^{k-1}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^i\right). \qquad (4.11)\end{aligned}$$

The first term in the brackets on the RHS of (4.11) is given by

$$\sum_{i=1}^{\infty}\sum_{k=1}^{i}\frac{1}{k^2}(1-e^{-\lambda t})^i = \sum_{k=1}^{\infty}\sum_{i=k}^{\infty}\frac{1}{k^2}(1-e^{-\lambda t})^i = e^{\lambda t}\sum_{k=1}^{\infty}\frac{1}{k^2}(1-e^{-\lambda t})^k.$$

For the second term, we have that

$$2\sum_{i=1}^{\infty}\sum_{k=1}^{i}\sum_{j=1}^{k-1}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^i = 2\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}\sum_{k=i}^{\infty}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^i = 2e^{\lambda t}\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^k.$$

Denoting with $g(t) := 2\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}(1-e^{-\lambda t})^k/(kj)$ and noticing that $g(0) = 0$ and

$$\begin{aligned} g'(t) &= \frac{2\lambda e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}\frac{1}{j}(1-e^{-\lambda t})^k = \frac{2\lambda e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{j=1}^{\infty}\sum_{k=j+1}^{\infty}\frac{1}{j}(1-e^{-\lambda t})^k \\ &= \frac{2\lambda}{1-e^{-\lambda t}}\sum_{j=1}^{\infty}\frac{1}{j}(1-e^{-\lambda t})^{j+1} = 2\lambda f(t) = 2\lambda^2 t, \end{aligned}$$

we obtain that $g_2(t) = \lambda^2 t^2$, and the second term on the RHS of (4.11) is thus $(\lambda^2 t^2)/(1-e^{-\lambda t})$.

So, joining all the results, we have that

$$\begin{aligned} &\lim_{t\to\infty}\mathrm{Var}\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}|Z(t)\right)\right) \\ &\quad = \lim_{t\to\infty} 4\left(\sum_{k=1}^{\infty}\frac{(1-e^{-\lambda t})^{k-1}}{k^2} + \frac{\lambda^2 t^2}{1-e^{-\lambda t}} - \frac{\lambda^2 t^2}{(1-e^{-\lambda t})^2}\right) = 4\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{2}{3}\pi^2 \end{aligned}$$

□

Proposition 4.1.1 follows from equation (4.2) using the results in Theorems 4.2.2 and 4.2.3.

## 4.3 Simulations under the Bellman-Harris framework

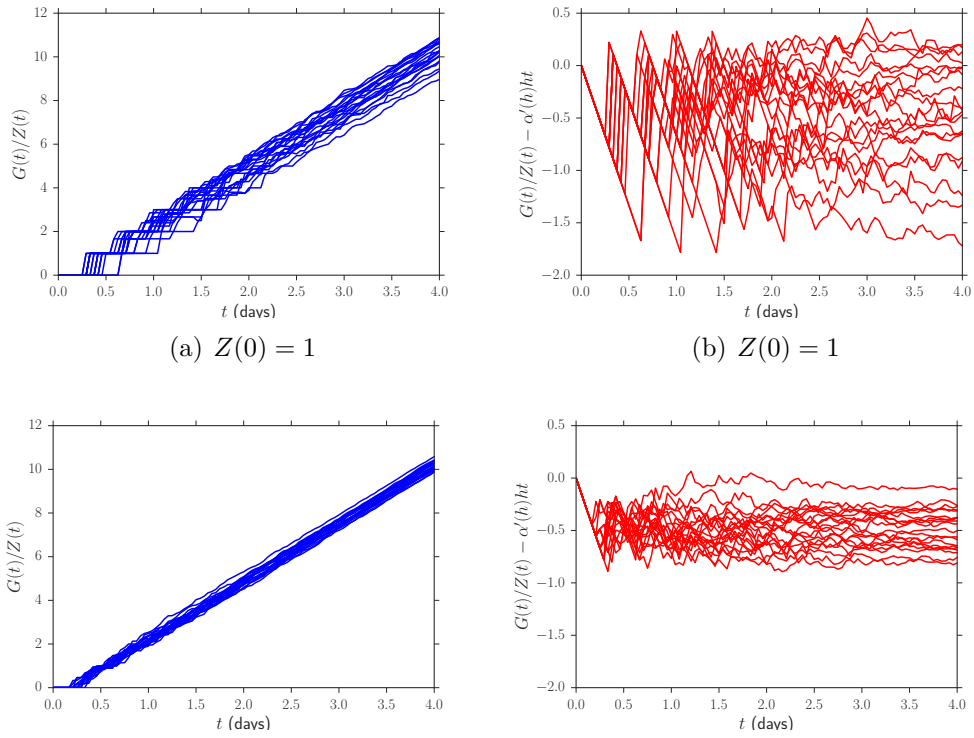
In Section 4.2 we have proved that, for a pure birth process, the variance of the average generation converges to a constant. Even if the same arguments used cannot be extended to a Bellman-Harris branching process, in this section we provide some evidence that it is still true that

$$\frac{G(t)}{Z(t)} \approx \alpha'(h)ht + A, \tag{4.12}$$

where $A$ is a random variable with finite variance. We do that making use of Bellman-Harris branching process simulations, whose parameterisation has been chosen according to [44], as discussed in Section 1.5. In particular, we have considered cell lifetimes following a lognormal distribution with mean 9.3 hours and standard deviation 2.54, and an offspring number distribution, $N$, such that $\mathbb{P}(N = 2) = 4/5 = 1 - \mathbb{P}(N = 0)$.

Fig. 4.2(a) shows us the trend of the average generation, $G(t)/Z(t)$, for 20 cell population simulations that start with one initial cell and grow according to a Bellman-Harris processes. According to (4.1), for large times, the blue paths grow linearly in time, with common slope $\alpha'(h)h$. So, ultimately, all the paths are parallel to each other with a variability in the $y$-intercept that reflects the time necessary for each population to enter in the "linear regime". The variability in such time is a consequence of the strong impact that a division can have in the average generation when only few cells are present in the population and it can be reduced by increasing the number of initial cells. We can see that in Fig. 4.2(c), where the same plot for a pool of 100 initial cells is considered.

In order to obtain more information on the term $O(t)$ in (4.1), in Fig. 4.2(b) we plot the behaviour over time of the quantity $G(t)/Z(t) - \alpha'(h)ht$ (red paths) for different populations that are made up of just one cell at time 0. From Fig. 4.2(b), it seems that no dependency on time is present in $G(t)/Z(t) - \alpha'(h)ht$, suggesting that even for a Bellman-Harris branching process $\mathrm{Var}(G(t)/Z(t))$ converges to a constant and (4.12) holds in these more general circumstances than a pure birth process. Establishing those more general conditions is left for future work beyond this thesis, but it is suggestive that the variability in the average generation from tree to tree might be smaller than one would naively assume.




(a) $Z(0) = 1$ (b) $Z(0) = 1$

(c) $Z(0) = 100$ (d) $Z(0) = 100$

Figure 4.2: **Average generation growth trend.** Each plot presents 20 Monte Carlo simulations of a Bellman-Harris branching process starting at $t = 0$ with a single cell, where paths are conditioned to have living cells at the final time-point of the simulation. The parametrisation used is the same employed in Fig. 2.1, i.e. lifetimes are lognormal with mean 9.3 hours and standard deviation 2.54, whereas the offspring number distribution, $N$, is such that $\mathbb{P}(N = 2) = 4/5$ and $\mathbb{P}(N = 0) = 1/5$. (a) With $Z(t)$ and $G(t)$ being the size and the total generation of the population at time $t$, this figure plots the evolution of $G(t)/Z(t)$. (b) With $\alpha > 0$ being the Malthusian parameter defined in equation (2.4), for the same paths this plot shows $G(t)/Z(t) - \alpha'(h)ht$, which is $O(t)$, according to Corollary 2.3.12. (c)-(d) Corresponding plots to (a)-(b), with 100 initial cells, i.e. $Z(0) = 100$.

CHAPTER 5

# Parameter choice for the average generation estimator

## 5.1 Introduction

The study of the average generation of a population, which we have conducted in previous chapters, was motivated by the DNA coded randomised algorithm that Weber et al. proposed in [116] to infer the average generation of a cell population using the proportion of label-positive cells at a particular time. In particular, if $Z^+(t)$ and $Z(t)$ denote the size of the label-positive and entire population, respectively, at time $t$, as long as $Z^+(t) > 0$, for a small probability of label loss $p$, the proposed estimator was

$$\frac{G(t)}{Z(t)} \approx -\frac{1}{p} \log\left(\frac{Z^+(t)}{Z(t)}\right). \tag{5.1}$$

In previous chapters we focused on the Left-Hand Side (LHS) of that equation, which was the quantity to be estimated. Here, we consider the Right-Hand Side (RHS) of (5.1), contributing to the analysis of the estimator itself. In particular, we wish to understand what impact the choice of the parameter $p$ has on the quality of the estimate and establish how small the parameter has to be in order to obtain the "best" results (in a sense that we clarify later).

As discussed in Chapter 2, Weber et al. provide two different arguments for justifying the estimator in [116]: a weaker result that doesn't assume any

structure in the cell population; and a stronger one under the super-critical Bellman-Harris branching process hypotheses. From both of them we obtain the same suggestion, namely we need a small $p$ for the estimator to approximate the average generation. The idea behind the second of these arguments, for example, is that so long as the label-positive population persists, we have that for large $t$

$$-\frac{1}{t}\log\left(\frac{Z^+(t)}{Z(t)}\right) \approx \alpha(h(1-p)) - \alpha(h),$$

where $\alpha(\cdot)$ is the Malthus parameter (see (2.4) for definition) and $h$ the expected number of offspring of a cell. However, as the Malthus parameter is real analytic [116, Proposition 1], $\alpha(h(1-p))$ coincides with its Taylor expansion around $p = 0$, giving

$$\alpha(h(1-p)) = \alpha(h) - h\alpha'(h)pt + O(p^2).$$

Thus, for large $t$, small $p$, and $Z^+(t) > 0$ we have that

$$-\frac{1}{pt}\log\left(\frac{Z^+(t)}{Z(t)}\right) \approx h\alpha'(h),$$

which is the same constant that appears for the time-rescaled average generation, $G(t)/(tZ(t))$, found in Corollary 2.3.12.

For any finite time, however, $p$ cannot be made arbitrarily small without detriment to the accuracy of the estimate. For fixed time, if $p$ is too small, in all likelihood $Z^+(t) = Z(t)$ and the estimate is 0. In some way, the Weber et al. suggestion is based on the assumption that, even for a very small probability of label loss $p$, the ratio $Z^+(t)/Z(t)$ is different from 1. In their results, this condition is verified because they either let $t$ go to infinity before dropping $p$ to 0, so that for a large population it is unlikely to have all the cells label-positive, or consider the quantity $\mathbb{E}\left(Z^+(t)\right)/Z(t)$ instead of $Z^+(t)/Z(t)$. This last scenario requires that several independent labels can be inserted in each cell, so that for a given family tree, several independent instances of $Z^+(t)$ are available (one per label) and an estimate of the value $\mathbb{E}\left(Z^+(t)\right)$ can be obtained. Thus the core question remains: for a finite system, how should $p$ be chosen?

We first note that the estimator is defined only when label-positive cells are present in the population, which places a distinct requirement as the choice

of the parameter $p$ influences the probability of not getting an estimate at all. Once the population has lost the label, the proportion of the label positive cells cannot be translated to information on the number of divisions that have occurred, given it will remain null despite new divisions taking place. However, even for large values of $p$, the probability of not getting an estimate, i.e. the probability of extinction of the label-positive population, is significantly reduced if the initial number of label positive cells is more than one, which would typically be the case in the adoptive transfer experiments that were the initial motivation for the estimator. Typical initial populations for an adoptive transfer lymphocyte experiment, $n_0$, are of the order of $10^2-10^6$ [16, 74]. Thus, if we ask, for example, for a probability of not getting an estimate below $10^{-2}$, so that no more than one in one hundred experiments provides no estimate, we have that

$$\mathbb{P}\left(Z^+(t)=0|Z^+(0)=n_0\right)=\mathbb{P}\left(Z^+(t)=0|Z^+(0)=1\right)^{n_0}<10^{-2},$$

so long as family trees develop independently, from which we obtain that

$$\mathbb{P}\left(Z^+(t)=0|Z^+(0)=1\right)<10^{-\frac{2}{n_0}}.$$

Thus, in our analysis we will disregard this concern and, on identifying an optimal value for $p$, merely check to ensure that the probability of no estimate is small.

Even if we have seen that the choice of $p$ hardly causes the impossibility of getting an estimate in biological applications, the quality of the estimate can depend upon this choice. Hence, establishing how small $p$ should be in order to obtain a good estimate is an interesting and non-obvious question we try to address in this chapter. Furthermore, by considering different models of population growth other than the Bellman-Harris branching process, we try to understand if these considerations are valid in all generality or if there are situations where the estimator would work even outside the conditions where both the LHS and RHS of(5.1) have been established to have meaning. Both scenarios will indeed be found, but we will see that the models that better capture and describe the complexity of the relationships among cells require a small $p$, and in each case our analysis will provide a suggested order of magnitude.

One oddity of the analysis conducted in [116], that is unimportant in the context studied there where $p \to 0$, is that in effect two different estimators are proposed, which differ from each other by just a prefactor: $-1/\log(1-p)$ instead of $-1/p$. They are essentially equivalent when $p$ is close to 0, as $\log(1-p) = -p + O(p^2)$ by Taylor's theorem, but when the optimal choice of $p$ is not they can provide qualitatively different estimates. In this chapter we shall see that, at least as far as mathematical analysis is concerned, one or other of them may be preferable, depending on the particular population model under study.

This chapter of the thesis, unlike the others, makes use of approximations rather than rigorous proofs. This necessity arises, for example, when we face the problem of computing moments of the estimator, which is a non-linear function of a ratio between two correlated random variables, or when we assume "large" $t$ or "large" number of cells in order to average the behaviour of the population, mitigating in this way the effect of the outliers. We try to be clear on that in the text by stating as Propositions only results that are formally proved, and confining approximate considerations to be outside of that environment.

As mentioned, we use some assumptions in order to smooth the behaviour of the population and be able to describe the size of the population through its expected behaviour. In particular, based on the selected model, we consider one of the following two types of working framework for conducting our analysis:

**F.1** We increase the size of the population keeping unchanged the expected average generation process. We can do that, for example, by increasing the initial size of the population but considering the population always at the same time $t$.

**F.2** We increase the size of the population by letting the population develop for a longer time $t$, changing inevitably the average generation distribution with $t$.

The results coming from the two different approaches are both useful, because they describe interesting experimental settings, but not directly commensu-

rable.

The function we use to quantify the performance of the estimator for a given value of probability loss, $p$, is the Mean Square Error (MSE). This is defined as

$$MSE(\hat{\theta}) = \mathbb{E}\left((\hat{\theta} - \theta)^2\right) = \left(\mathbb{E}\left(\hat{\theta}\right) - \theta\right)^2 + \mathrm{Var}\left(\hat{\theta}\right), \tag{5.2}$$

where we have denoted with $\theta$ the non-random quantity that we want to estimate and with $\hat{\theta}$ the estimator. From (5.2) we can see that the MSE is a function that associates to the estimator a nonnegative number that is the sum of two penalisation terms: the square of the bias, which is the expected difference between the estimate and the parameter object of study, and the variance of the estimator. The bigger is the value of the MSE, the worse the estimator is considered to be. Given in our case the estimator depends on the probability of label loss, the study of the minima (if any) of the MSE as function of $p$ gives us a criteria to select the "best" values for $p$. As said before, the values resulting from this analysis will have no problem in keeping the probability of non getting an estimate, i.e. probability of extinction of the label-positive population, below a certain threshold, due to the large size of the initial label-positive population that are usually considered in biological experiments.

Given the estimator we consider is a non-linear function of the ratio between two dependent random variables, computing the mean and the variance requires the use of some approximations. In particular, using the fact that the function $\log(x/y) = \log(x) - \log(y)$, and that for values of $x$ and $y$ in a neighbourhood of $\mu_x > 0$ and $\mu_y > 0$ respectively, the following Taylor approximation holds

$$\begin{aligned}
\log(x) =& \log(\mu_x) + \frac{(x-\mu_x)}{\mu_x} - \frac{(x-\mu_x)^2}{2\mu_x^2} + O((x-\mu_x)^3), \\
\log^2(x) =& \log^2(\mu_x) + 2\log(\mu_x)\frac{(x-\mu_x)}{\mu_x} + (1-\log(\mu_x))\frac{(x-\mu_x)^2}{\mu_x^2} \\
&+ O((x-\mu_x)^3),
\end{aligned}$$

$$\log(x)\log(y) = \log(\mu_x)\log(\mu_y) + \log(\mu_y)\frac{(x-\mu_x)}{\mu_x} + \log(\mu_x)\frac{(y-\mu_y)}{\mu_y} - \frac{1}{2}\left(\log(\mu_y)\frac{(x-\mu_x)^2}{\mu_x^2} - 2\frac{(x-\mu_x)(y-\mu_y)}{\mu_x\mu_y} + \log(\mu_x)\frac{(y-\mu_y)^2}{\mu_y^2}\right) + O((x-\mu_x)^3) + O((y-\mu_y)^3).$$

These approximations are still valid when computed on positive random variables $X$ and $Y$, around the points $\mu_x = \mathbb{E}(X)$ and $\mu_y = \mathbb{E}(Y)$, respectively. Using the fact that $\mathbb{E}(\log(X/Y)) = \mathbb{E}(\log(X)) - \mathbb{E}(\log(Y))$ and that $\mathrm{Var}(\log(X/Y)) = \mathbb{E}\left(\log^2(X) - 2\log(X)\log(Y) + \log^2(Y)\right) - (\mathbb{E}(\log(X)) - \mathbb{E}(\log(X)))^2$, we obtain the following approximations

$$\mathbb{E}\left(\log\left(\frac{X}{Y}\right)\right) \approx \log\left(\frac{\mathbb{E}(X)}{\mathbb{E}(Y)}\right), \tag{5.3}$$

$$\mathrm{Var}\left(\log\left(\frac{X}{Y}\right)\right) \approx \frac{\mathrm{Var}(X)}{\mathbb{E}(X)^2} - 2\frac{\mathrm{Cov}(X,Y)}{\mathbb{E}(X)\,\mathbb{E}(Y)} + \frac{\mathrm{Var}(Y)}{\mathbb{E}(Y)^2} \tag{5.4}$$

The above approximations, that have an error $O(\mathrm{Var}(X) + \mathrm{Cov}(X,Y) + \mathrm{Var}(Y))$, assume that the distribution of $X$ and $Y$ are concentrated around their mean values.

### 5.1.1 Models and regimes considered

We start by considering a very simple model which is based on the assumption that the label loss process has occurred independently for all the cells of the collection we want to determine the average generation of. Even if the independence assumption is unrealistic for the majority of the biological applications, it gives light to an easier mathematical model that is an interesting starting point to understand the role of $p$ in the quality of the estimates. A straightforward consequence of this assumption is that the number of label-positive cells in the population is Binomial distributed. We study this model in Section 5.2. Our analysis led us to formulate two types of suggestions for the choice of $p$: if the population under study is homogeneous, i.e. the cells are in the same unknown generation $g$ and we can have a rough idea of it, a value of $p \approx 1 - e^{-1.6/g}$ is the one we should aim for; otherwise, we should consider a probability of label loss that scales with the size of the population we are studying, namely $p \approx (1/n)^{1/3}$.

We then insert some dependence among the label loss processes of the cells of the population using a model that preserves the homogeneity in the gen-

erations of all living cells: the Galton-Watson model. Here, similar to the i.i.d homogeneous case treated in Section 5.2.1, we find that the estimator can determine in an accurate way the average generation of the population even for values of $p$ that are not small. Indeed, we find that the Mean Square Error $MSE(p)$ of the estimator reaches its minimum inside the interval $(0, 1)$ and this minimum does not depende on the size of the initial population. In the particular case in which two offspring are produced after each division, we numerically find that this value is $p \approx 0.38$. Analysing the probability of extinction of the label-positive population, we are able to assert that the initial population has to made up of at least 9 cells in order to have a probability of not getting an estimate below $10^{-2}$. We do that in Section 5.3.

We conclude by studying the same problem under the Bellman-Harris branching process framework. Due to the complexity of this model, we can only bring the mathematical analysis up to a point, and draw conclusions using evidence coming from simulations. Even in this case, it seems that the heterogeneity in the generation of the cells forces the "optimal" $p$ to drop to 0 when the initial population, $n_0$, increases. This idea is also supported by the fact that, in this case, we need recourse to approximations that needs large $t$ to be true (see **F.2** in the previous section). Despite that, there exists a special case in which the MSE of the estimator is minimised by a value of $p$ that does not decrease with $n_0$: the Birth-Death branching process. In this case, characterised by Exponential lifetime distributions, we are able to move the analysis further and see, for example, that if two offspring are generated after each division and there are at least 9 cells in the initial population, the value $p = 0.3775$ minimises the MSE and keeps at the same time the probability of not getting an estimate below $10^{-2}$. This behaviour is probably a consequence of the Markovian character that the exponential lifetime distribution confers to the process.

## 5.2 Cells with independent label loss processes

Consider a population in which all the ancestors were equipped with a neutral label, heritably lost, with a small probability $p$, just before division, and assume that they were in generation 0 at the time of first consideration. Assume

that all the living cells have the label loss process independent of each other and that the population we are studying is made up of $n$ cells. Denote with $Z_n^+$ the number of them that still have the label. The generation of each cell is an independent random quantity that is drawn from a distribution on the natural numbers whose associated random variable is $\Gamma$. In order to be able to refer to these assumptions, that will stay in force until the end of the section, we summarise them in the following

**Assumption 5.2.1.** The living cells have the label loss process independent of each other and their generations are drawn, in an independent way, from a common distribution on the natural numbers associated with the random variable $\Gamma$.

Given the only suggestion we have on the value of $p$ is to consider it small, and given the size of the population, $n$, affects the possibility to see cells with no label for small probability of label loss, it is reasonable to think that the choice of the optimal $p$ might depend upon $n$. In order to investigate that, we consider the probability of label loss as being dependent on the sample size, $p_n$. Our aim is to provide a suggestion for a good selection of the parameter $p_n$.

According to Assumption 5.2.1, a cell is label-positive if its ancestors didn't lose the label in the $\Gamma$ divisions that led to it. This means that a cell is equipped with the neutral label with probability $(1-p_n)^\Gamma$, independently of the other members of the sample. For $i \in \{1, 2, \dots, n\}$, let $X_{n,i}$ be the random variable that assumes value 1 if the $i$-th cell of the sample is equipped with the label and 0 otherwise. From what we have said, we have that $X_{n,i} \overset{D}{\sim} \text{Ber}((1-p_n)^{\Gamma_i})$, with $\Gamma_i \overset{D}{\sim} \Gamma$ generation of the cell $i$. Given $\mathbb{P}(X_{n,i}=1) = \mathbb{E}(\mathbb{P}(X_{n,i}=1|\Gamma_i)) = \mathbb{E}\left((1-p_n)^{\Gamma_i}\right)$, we have also that $X_{n,i} \overset{D}{\sim} \text{Ber}\Big(\mathbb{E}\left((1-p_n)^{\Gamma_i}\right)\Big)$, and so that the number of positive cells in the sample of size $n$, $Z_n^+ := \sum_{i=1}^n X_{n,i}$, is a random walk on $\mathbb{N}_{\geq 0}$ that follows the binomial distribution $\text{Bin}\Big(n, \mathbb{E}\left((1-p_n)^\Gamma\right)\Big)$.

It is useful to note that $\mathbb{E}\left((1-p_n)^\Gamma\right) = \mathbb{E}\left(e^{\log(1-p_n)\Gamma}\right) = M_\Gamma(\log(1-p_n))$, with $M_\Gamma(\cdot)$ Moment Generating Function (MGF) of the random variable $\Gamma$, as within the interval where the MGF is finite, it is real analytic (e.g. [23]). Given $\Gamma$ is a positive random variable, we have that $M_\Gamma(s) < \infty$ for $s < 0$. When it is also finite in a neighbourhood of 0, we can substitute it with its

Taylor approximation, that is

$$M_\Gamma(s) = \mathbb{E}\left(e^{s\Gamma}\right) = \mathbb{E}\left(\Gamma\right) + \mathbb{E}\left(\Gamma^2\right)s + O(s^2),$$

where we recall that $\mathbb{E}\left(\Gamma\right)$ is the quantity we want to estimate. Asking for $M_\Gamma(s) = \mathbb{E}\left(e^{s\Gamma}\right) < \infty$ in a neighbourhood of 0 is equivalent to the Probability Generating Function (PGF) $\mathbb{E}\left(s^\Gamma\right) < \infty$ in a neighbourhood of 1. This motivates the following assumption that will be sometimes used within the section.

**Assumption 5.2.2.** The random variable $\Gamma$ is such that its PGF is finite in a neighbourhood of 1, i.e. $\mathbb{E}\left(s^\Gamma\right) < \infty$ for $s \in (1-\delta, 1+\delta)$ and $\delta > 0$. Equivalently, $\Gamma$ has a MGF finite in a neighbourhood of 0, i.e. $\mathbb{E}\left(e^{s\Gamma}\right) < \infty$ for $s \in (-\delta, \delta)$ and $\delta > 0$.

The distributions that are excluded from Assumption 5.2.2 are essentially the ones that have a tail that doesn't decrease fast enough to compensate the behaviour of the exponential function in the integrand of $\mathbb{E}\left(e^{s\Gamma}\right)$ for any $s > 0$. These distributions are known in the literature as heavy-tailed distributions (e.g. [31, pg. 2]) and an example is the lognormal distribution.

When we allow for a probability of label loss that is tailored to the size of the population we are studying we have to deal with sequences $\{p_n\}$. A first observation on the behaviour of the estimator for different types of convergent sequences $\{p_n\}$ comes from the following proposition.

**Proposition 5.2.1.** *For each $n > 0$ and $p_n \in (0,1)$, construct independently a cell population of size $n$ and probability of label loss $p_n$ as described in Assumption 5.2.1. The probability of not getting an estimate is given by*

$$\mathbb{P}\left(Z_n^+ = 0\right) = \left(1 - \mathbb{E}\left((1-p_n)^\Gamma\right)\right)^n. \tag{5.5}$$

*If $p_n \in (0,1)$ is s.t. $p_n \to \bar{p} \in (0,1)$, we have*

$$\lim_{n\to\infty} \frac{1}{\log(1-p_n)} \log\left(\frac{Z_n^+}{n}\right) \overset{a.s.}{=} \frac{\log\left(\mathbb{E}\left((1-\bar{p})^\Gamma\right)\right)}{\log(1-\bar{p})}. \tag{5.6}$$

*If $p_n \to 0$, under Assumption 5.2.2, the limit in (5.6) is equal to $\mathbb{E}\left(\Gamma\right)$.*

**Proof:** The result that appears in (5.5) is a direct consequence of the fact that $Z_n^+ \overset{D}{\sim} \text{Bin}\Big(n, \mathbb{E}\left((1-p_n)^\Gamma\right)\Big)$.

In order for the limit in the LHS of (5.6) to make sense, we need that $Z_n^+ > 0$ for all but finitely many $n$. Using (5.5), we have that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(Z_n^+ = 0\right) = \sum_{n=1}^{\infty} \left(1 - \mathbb{E}\left((1-\bar{p})^\Gamma\right) + o(1)\right)^n < \infty,$$

where we have used the fact that $\lim_{n\to\infty} \mathbb{E}\left((1-p_n)^\Gamma\right) = \mathbb{E}\left((1-\bar{p})^\Gamma\right) < 1$. So, for the Borel-Cantelli lemma, we can conclude that $Z_n^+ > 0$ for all but finitely many $n$.

Applying the Strong Law of Large Number (SLLN) for triangular arrays to $\{X_{n,i}\}$ (e.g. [109, 92]), we have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_{n,i} - \mathbb{E}\left((1-p_n)^\Gamma\right) \overset{a.s.}{=} 0. \tag{5.7}$$

Using (5.7) and the Continuous Mapping Theorem (e.g. [103, pg. 24]), we have that

$$\begin{aligned} \lim_{n\to\infty} \log\left(\frac{\sum_{i=1}^n X_{n,i}}{n}\right) &\overset{a.s.}{=} \lim_{n\to\infty} \log\left(\mathbb{E}\left((1-p_n)^\Gamma\right) + o(1)\right) \\ &\overset{a.s.}{=} \lim_{n\to\infty} \left(\log\left(\mathbb{E}\left((1-p_n)^\Gamma\right)\right) + \log\left(\frac{o(1)}{\mathbb{E}\left((1-p_n)^\Gamma\right)}\right)\right) \\ &\overset{a.s.}{=} \log\left(\mathbb{E}\left((1-\bar{p})^\Gamma\right)\right), \end{aligned} \tag{5.8}$$

which gives us

$$\lim_{n\to\infty} \frac{1}{\log(1-p_n)} \log\left(\frac{Z_n^+}{n}\right) \overset{a.s.}{=} \frac{\log\left(\mathbb{E}\left((1-\bar{p})^\Gamma\right)\right)}{\log(1-\bar{p})}. \tag{5.9}$$

In a similar way, when $p_n \to 0$ and given we have assumed that $\mathbb{E}\left(s^\Gamma\right) < \infty$ around $s = 1$, using (5.8) and the fact that $\lim_{n\to\infty} \log\left(\mathbb{E}\left((1-p_n)^\Gamma\right)\right) / \log(1-p_n) = \mathbb{E}(\Gamma)$, we have that the limit in (5.9) is almost surely equal to $\mathbb{E}(\Gamma)$. □

The proposition just proved says that if we want to recover the value $\mathbb{E}(\Gamma)$ from the proportion of label-positive cells in the population, we should drop to 0 the value of $p_n$ when the size of the population $n$ increases. Other choices

of the sequence $\{p_n\}$ such as $p_n \in (0,1)$ constant for every $n$, or $p_n$ converging to a different value other than 0, will lead the estimator to converge almost surely towards a different quantity.

One exception is the homogeneous setting, where all the cells in the population share the same generation, i.e. $\Gamma = g > 0$. In that case, even a single choice of the probability of label loss $\bar{p}$ for every size of the population leads to the correct estimate of the average generation of the population, $g$. Which $\bar{p}$ is the best option is not clear from Proposition 5.2.1, and this is why, before studying the general case in Section 5.2.2, we analyse this unusual behaviour in the next section.

### 5.2.1 Homogeneous case

Within this section we further restrict the assumptions of the model considering the following set of assumptions

**Assumption 5.2.3.** The living cells have the label loss process independent of each other and share the same generation, $g$. This is equivalent to assuming that Assumption 5.2.1 is in force with $\Gamma = g$.

Despite the fact that there is no real system for which Assumptions 5.2.3 is true, we investigate this case because it is the first model scenario where the deduction differs from the Bellman-Harris branching model, as it is not necessary for $p_n$ to go to 0. Under this assumption, we have that the label-positive population is given by $Z_n^+ \overset{D}{\sim} \text{Bin}(n, (1-p_n)^g)$.

As said in the introduction of the chapter, our criterion to establish the best value of $p_n$ for a given size of the population, $n$, is the to minimise the Mean Square Error (MSE). This means that we want to minimise the quantity in (5.2) for $p_n \in (0,1)$, with $\hat{\theta} = \log(Z_n^+/n)/\log(1-p_n)$ and $\theta = g$. Despite the fact that in this simple case the label positive population, $Z_n^+$, is a binomial random variable, there appear to be no easy expressions for the mean and variance of the estimator. So, we find an approximation of them

using (5.3) and (5.4), obtaining that

$$\mathbb{E}\left(\frac{1}{\log(1-p_n)}\log\left(\frac{Z_n^+}{n}\right)\right) \approx g, \tag{5.10}$$

$$\mathrm{Var}\left(\frac{1}{\log(1-p_n)}\log\left(\frac{Z_n^+}{n}\right)\right) \approx \frac{1-(1-p_n)^g}{n(1-p_n)^g\log^2(1-p_n)}. \tag{5.11}$$

These approximations are good as long as $Z_n^+$ is concentreted around its mean, i.e. as long as $\mathrm{Var}\,(Z_n^+) = np_n(1-p_n)$ is small. This will be the case for the value of $p$ we will find.

Given the expected estimator value is approximated by the average generation of the population, $g$, the study of the MSE of the estimator reduces to study the variance in (5.11). Computing the derivative of the expression in (5.11), we find that the value of $p_n$ that minimises the variance of the estimator solves the equation

$$\frac{g\log(1-p_n)+2[1-(1-p_n)^g]}{n(1-p_n)\log(1-p_n)^3} = 0.$$

We can see that the location of the point of minimum of the MSE does not depend on the size of the population, $n$. The solution of the above equation, for $p = p_n$, can be expressed using the principal branch of the Lambert-W function $W(x)$, a quantity that appears frequently in the solutions of a variety of mathematical problems and is defined as the solution of the functional equation $W(x)e^{W(x)} = x$ (e.g. [21, 85]). In this way, for every given $n$, the value of $p_n$ that minimises the MSE, is given by

$$p^*(g) := 1 - e^{\frac{-W(-2/e^2)-2}{g}} \approx 1 - e^{-\frac{1.6}{g}}, \tag{5.12}$$

where the value of $W(-2/e^2)$ has been computed numerically.

Thus, as long as $\mathrm{Var}\,(Z_n^+)$ is small and we can use the approximations for the mean and the variance found in (5.10) and (5.11), by defining $p_n$ for every $n$ as in (5.12), i.e. $p_n = p^*(g)$, we minimise the Mean Square Error of the estimator. Given the quantity $g$ is what we want to estimate, it is not possible to determine the value of (5.12) in advance, but if we have some information on the cycle of the cell and we can have a rough idea of $g$, we can use it to determine the probability of label loss for the experiment.

Assuming $p_n = p^*(g)$, we can check how large the population has to be in order to keep the probability of not getting an estimate below $10^{-2}$. Using (5.5), we

have that

$$\mathbb{P}\left(Z_n^+ = 0\right) = \left(1 - \left(1 - \left(1 - e^{\frac{-1.6}{g}}\right)^g\right)\right)^n = \left(1 - e^{-1.6}\right)^n \leq 10^{-2},$$

from which we obtain

$$n \geq \frac{\log(10^{-2})}{\log(1 - e^{-1.6})} \approx 20.$$

So, if the population is made up of at least 20 cells, by defining $p_n = 1 - e^{-1.6/g}$ we minimise the MSE of the estimator and keep at the same time the probability of not getting an estimate below $10^{-2}$.

We can also notice that when $g$ grows, the value of $p^*(g)$ drops to 0 (see Fig. 5.1(a)). Given the principal branch of the Lambert-W function is analytic (e.g. [15]), if we make a change of variable $y = 1/g$ in (5.12) and make a Taylor expansion of $p^*(y)$ around $y = 0$, we can determine the speed at which this decay happens

$$p^*(g) = \frac{W(-2/e^2) + 2}{g} + O(1/g^2) \approx \frac{1.6}{g} + O(1/g^2).$$

Given the approximation that we have used for the mean makes the estimator appear approximately unbiased, one can wonder if an additional term in the Taylor expansion of the mean of the estimator in (5.10) would provide different results. If we do so, we obtain that (5.10) becomes

$$\mathbb{E}\left(\frac{1}{\log(1 - p_n)} \log\left(\frac{Z_n^+}{n}\right)\right) \approx g - \frac{1 - (1 - p_n)^g}{2n \log(1 - p_n)(1 - p_n)^g},$$

and that the approximation for the Mean Square Error is given by

$$MSE_n(p_n) \approx \left(\frac{1 - (1 - p_n)^g}{2n \log(1 - p_n)(1 - p_n)^g}\right)^2 + \frac{1 - (1 - p_n)^g}{n(1 - p_n)^g \log^2(1 - p_n)}. \quad (5.13)$$

Given for any $n \in \mathbb{N}$ $\lim_{p\to 0} MSE_n(p) = \infty = \lim_{p\to 1} MSE_n(p)$, the continuity of the function tells us that the expression in (5.13) admits at least a global minimum in $(0, 1)$. While we were unable to find a closed form solution to this equation, it is readily computable numerically. In Fig. 5.1(b) we can see the behaviour of the $MSE_n(p_n)$ for different values of $n$, when the average generation $g$ is assumed to be 7. There, with a black dashed line, we have highlighted the value of $p^*(7) \approx 0.2036$. Independently of the size of the population, $n$,

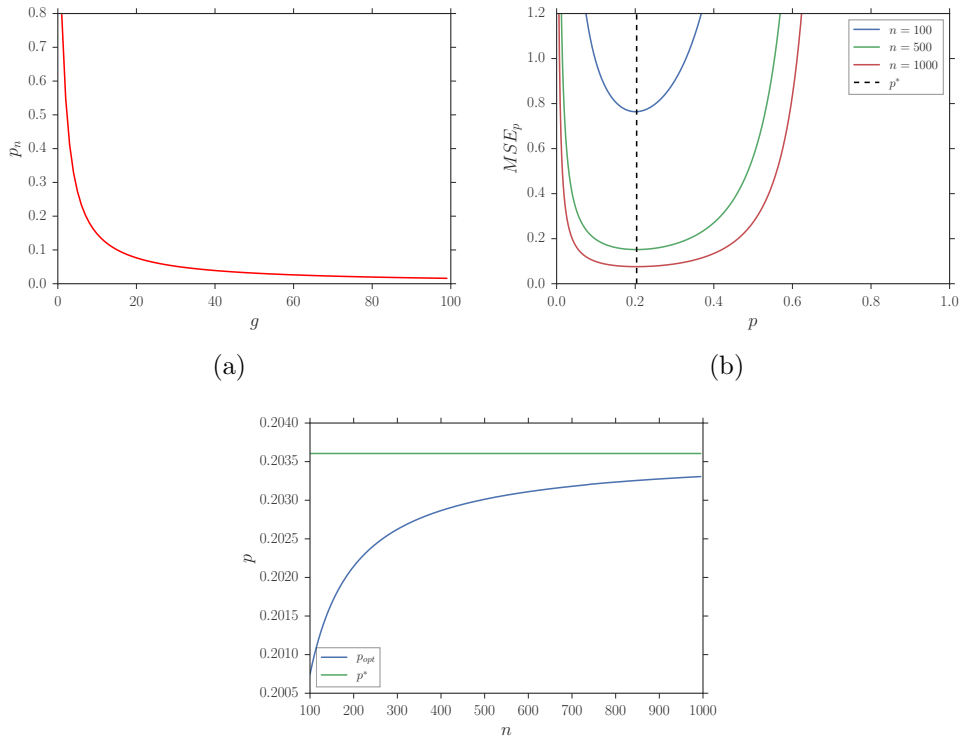



Figure 5.1: **Behaviour of $MSE_n(p_n)$ and $MMSE_n$ for different values of $n$.** (a) The plot shows the trend of $p^*(g) = 1 - \exp((-W(-2/e^2) - 2)/g)$ in function of the average generation $g$, where $W(\cdot)$ is the linear branch of Lambert-W function (e.g. [21, 85]). The function plotted, introduced in (5.12), describes the behaviour of the probability of label loss $p_n$ that, for a given $n$, minimises the approximation of the variance of the estimator in (5.11). This quantity is also called Minimum Mean Square Error (MMSE). (b) For the average generation $g = 7$ and values of $n$ in $\{100, 500, 1000\}$, the approximations of the MSE in (5.13) is plotted in function of $p$. Coloured solid curves correspond to different values of $n$, whereas the black dashed line describes the function $p = p^*(7)$, where $p^*(7)$ is the point that minimises the approximation of variance of the estimator in (5.11). (c) The panel shows a comparison between the numerically computed minimum of the MSE in (5.13) (blue line) and the function $y = p^*(7)$ (green line), when the size of the population $n$ grows.

we can see that the minimum previously found in (5.12) represents a good approximation of the minimum of the quantity in (5.13). This is made more clear in Fig. 5.1(c) where the the value of $p^*(7)$ is compared with the real minimum of the MSE in (5.13), computed with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) minimisation algorithm, for different values of $n$.

In the next section we see how the situation changes when we admit some heterogeneity in the generation of the cells.

### 5.2.2 Heterogeneous case

Recall the assumptions made in Assumptions 5.2.1 and 5.2.2. We have seen that a consequence of these assumptions is that $Z_n^+ \overset{D}{\sim}\text{Bin}\Big(n, \mathbb{E}\left((1-p_n)^\Gamma\right)\Big)$, where $n$ is the size of the population.

As in the homogeneous case, we can use the Taylor approximations in (5.3) and (5.4) to obtain an approximation of the mean and variance of the estimator

$$\mathbb{E}\left(\frac{1}{\log(1-p_n)}\log\left(\frac{Z_n^+}{n}\right)\right) \approx \frac{\log\left(\mathbb{E}\left((1-p_n)^\Gamma\right)\right)}{\log(1-p_n)}, \tag{5.14}$$

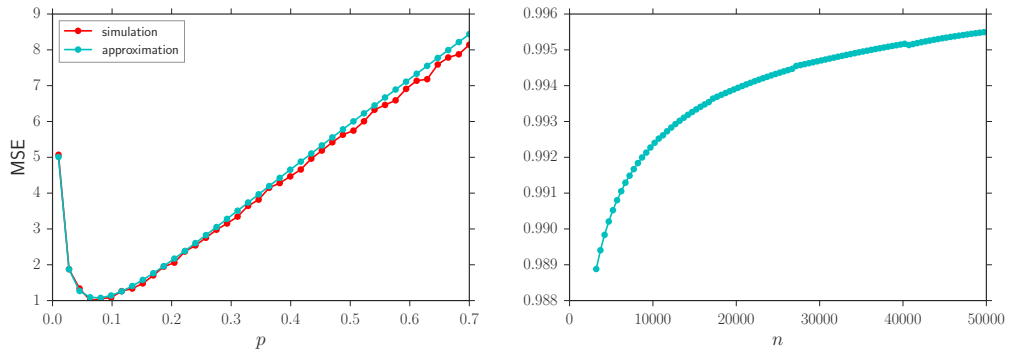
$$\text{Var}\left(\frac{1}{\log(1-p_n)}\log\left(\frac{Z_n^+}{n}\right)\right) \approx \frac{1}{n}\frac{1-\mathbb{E}\left((1-p_n)^\Gamma\right)}{\mathbb{E}\left((1-p_n)^\Gamma\right)\log(1-p_n)^2}. \tag{5.15}$$

Given we are interested in the value of $p_n$ that minimises the MSE of the estimator, we should find the minimum of

$$MSE_n(p_n) \approx \left(\frac{\log\left(\mathbb{E}\left((1-p_n)^\Gamma\right)\right)}{\log(1-p_n)} - \mathbb{E}\left(\Gamma\right)\right)^2 + \frac{1}{n}\frac{1-\mathbb{E}\left((1-p_n)^\Gamma\right)}{\mathbb{E}\left((1-p_n)^\Gamma\right)\log(1-p_n)^2}. \tag{5.16}$$

A comparison between the approximation found above and the MSE computed using Monte Carlo simulations is made in Fig. 5.2(a), for a generation Poisson distributed, $\Gamma \overset{D}{\sim}\text{Poi}(5)$, and a population size of $n = 100$. Despite only few terms being considered in the Taylor expansion of the mean and variance in (5.14) and (5.15), we can see that the expression in (5.16) provides a good estimate of the MSE.

Due to the continuity of the function in (5.16) on the interval $(0, 1)$ and the fact that $\lim_{p\to 0} MSE_n(p) = \infty = \lim_{p\to 1} MSE_n(p)$, the existence of a global




(a) (b)

Figure 5.2: $MSE_n(p_n)$ **and** $MMSE_n$ **in the i.i.d. heterogeneous case.** (a) For a generation Poisson distributed, $\Gamma \overset{D}{\sim} \text{Poi}(5)$, and $n = 100$, the panel shows a comparison between the behaviour of the approximation of the MSE described in (5.16) (cyan line) and the MSE obtained using 1000 Monte Carlo simulation for each probability of label loss $p$ considered (red line). (b) Assuming a generation Geometric distributed, $\Gamma \overset{D}{\sim} \text{Geo}(1/5)$, the plot shows the trend of the Minimum Mean Square Error, obtained through a numerical minimisation of (5.16), multiplied by a factor $n^{1/3}$. It seems that the curve is converging to a constant when $n$ goes to $\infty$ and this support the thesis that the $MMSE_n$ decreases as $n^{-1/3}$ when $n$ increases.

minimum in that interval is assured. An explicit expression for this point was not forthcoming, and so, in order to find its location, we have to resort to some further approximations.

We have seen in Proposition 5.2.1 that using a value of $p_n$ that drops to 0 when the size of the population $n$ increases is what we want. Thus, we try to establish the position of the point of minimum of the MSE of the estimator recurring to the Laurent approximations of the quantities in the RHS of (5.14) and (5.15) around $p = 0$. These quantities are analytic thanks to the assumption $\mathbb{E}\left(s^{\Gamma}\right) < \infty$ in a neighbourhood of 1. Using that

$$\frac{\log\left(\mathbb{E}\left((1-p_n)^{\Gamma}\right)\right)}{\log(1-p_n)} = \mathbb{E}(\Gamma) - \frac{1}{2}\left(\text{Var}(\Gamma) - \mathbb{E}(\Gamma)\right)p_n + O({p_n}^2),$$

$$\frac{1}{n}\frac{1-\mathbb{E}\left((1-p_n)^{\Gamma}\right)}{\mathbb{E}\left((1-p_n)^{\Gamma}\right)\log(1-p_n)^2} = \frac{\mathbb{E}(\Gamma)}{np_n} + \frac{1}{n}O(1),$$

we obtain the following approximation for $n$ large and $p_n$ around 0

$$\begin{aligned} MSE_n(p_n) &\approx \frac{\mathbb{E}(\Gamma)}{np_n} + \frac{a_0}{n} + \frac{a_1}{n}p_n + \left(\frac{a_2}{n} + \frac{1}{4}\left(\mathrm{Var}(\Gamma) - \mathbb{E}(\Gamma)\right)^2\right){p_n}^2 + \frac{1}{n}O({p_n}^3) \\ &\approx \frac{\mathbb{E}(\Gamma)}{np_n} + \frac{1}{4}\left(\mathrm{Var}(\Gamma) - \mathbb{E}(\Gamma)\right)^2 {p_n}^2, \end{aligned} \tag{5.17}$$

for some constants $a_0, a_1, a_2 \in \mathbb{R}$. The last step of the above approximation is consequence of the fact that $p_n \in (0,1)$ and that $n$ is assumed to be large.

Computing the derivative in $p_n$ of the quantity in (5.17), we obtain that for a given $n$, the value of $p$ that minimises the approximation of the MSE of the estimator in equation (5.17), is given by

$$p_n^* = \left(\frac{4\mathbb{E}(\Gamma)}{\left(\mathrm{Var}(\Gamma) - \mathbb{E}(\Gamma)\right)^2}\frac{1}{n}\right)^{\frac{1}{3}} \sim \left(\frac{1}{n}\right)^{\frac{1}{3}}. \tag{5.18}$$

In Fig. 5.2(b), through the use of Monte Carlo simulations, we have computed the Minimum Mean Square Error (MMSE), i.e. the value of $p_n$ that minimises the MSE, in function of $n$, and we have plotted its behaviour when multiplied by a factor of $n^{1/3}$. To do that, we have assumed that the generations are drawn from a Geometric distribution with mean 5, i.e. $\Gamma \overset{D}{\sim} \mathrm{Geo}(1/5)$. Given the cyan line seems to be converging to a constant, the plot corroborates the thesis that, given a population size $n$, the value of $p_n$ the minimises the MSE of the estimator decreases as $n^{-1/3}$ in this model.

Defining $p_n = (1/n)^{1/3}$ and using (5.5), we can compute the minimum size of the population for which the value of $p_n$ gives a probability of not getting an estimate below $10^{-2}$. In particular, we have

$$\mathbb{P}\left(Z_n^+ = 0\right) = \left(1 - \mathbb{E}\left((1-p_n)^\Gamma\right)\right)^n \approx \left(\mathbb{E}(\Gamma)\, p_n\right)^n \approx \left(\frac{1}{n}\right)^{-\frac{n}{3}} \leq 10^{-2},$$

from which we obtain

$$n \geq e^{W(6\log(2)+\log(5))} \approx 7,$$

with $W(x)$ the Lambert-W function. So, when the population has more than 7 cells, the value of $p_n = (1/n)^{1/3}$ minimises the MSE and keep at the same time the probability of not getting an estimate below $10^{-2}$. For this reason, given a cell sample with size $n$, a value of $p_n$ such as (5.18) is our suggestion to maximise the quality of the estimate of the average generation.

## 5.3 Galton-Watson branching process

After having considered in Section 5.2 a population where the cells had independent label loss processes, in this section we introduce some cell dependencies by considering the most simple form of branching process: the Galton-Watson process (see Section 1.3 and [42, 9] for more details).

Consider a cell population that behaves according to a Galton-Watson branching process with number offspring distribution $N$. This means that each cell of the population, independently of the others, generates a random number of offspring distributed as an independent copy of $N$. Given all the cells in the population always divide simultaneously, a Galton-Watson branching process is generally studied across generations, without focusing on the time at which these divisions occur. In the following, we summarise the assumptions that will stay in force within this section.

**Assumption 5.3.1.** The initial population is made up of $n_0$ cells, each of which is equipped with a heritable neutral label that either is lost before division with probability $p$, or is transmitted to the offspring with probability $1-p$. We denote with $Z_{n_0}(k)$ and $Z^+_{n_0}(k)$ respectively the sizes of the whole and the label-positive population in generation $k$ generated from $n_0$ ancestors. When $n_0 = 1$, we simplify the notation in $Z(k)$ and $Z^+(k)$. We assume that $p \in (0, 1 - 1/h)$, where $h := \mathbb{E}(N)$, so that these last two processes are supercritical, i.e. the expected number of label-positive cells obtained after a label-positive cell division, $h(1-p)$, is greater than 1.

There are phenomenological differences between branching processes lattice distributed, i.e. where the lifetime distributions are such that every possible value is of the form $a + bn$ for $a, b \neq 0$ fixed and $n$ the varies in the integers, and the ones that are not. The Galton-Watson model is an example of the first type, given the lifetimes of the cells are equal and constant. This is why it would not be surprising if the results for the estimator differed for a Galton-Watson than for a Bellman-Harris branching process. Furthermore, notice that there is no cell cycle model where lifetimes are constant, and so the results we find in this section are just for mathematical interest, or perhaps for other applications.

Due to these phenomenological differences, all the results proved by Weber et al. in [116] in the context of a Bellman-Harris branching process are not directly applicable here. Indeed, one of the assumptions used there is that the cells lifetimes are non-lattice distributed. However, it is not hard to recover a result equivalent to [116, Theorem 1], reported in this Thesis in the second equation of (2.37) and (2.38). This time, however, the homogeneity in the generation of the living cells allows us to drop the requirement of having a small $p$, as long as the estimator uses the prefactor $1/\log(1-p)$ instead of $-1/p$. We show this result in the following proposition, along to a similar result where the regularity in the population is obtained by increasing the initial size of the population instead of increasing the timeframe after which the population is studied (a longer timeframe means a higher number of divisions and so a higher generation $k$). Both results are a direct application of the Strong Law of Large Number (SLLN), the Continuous Mapping Theorem, and well known results on Galton-Watson branching processes.

**Proposition 5.3.1.** *Under Assumption 5.3.1 and as long as lim inf$_k$ $Z^+_{n_0}(k) > 0$, we have that*

$$\lim_{k\to\infty} \frac{1}{k\log(1-p)} \log\left(\frac{Z^+_{n_0}(k)}{Z_{n_0}(k)}\right) \overset{a.s.}{=} 1.$$

*Furthermore, for any $p \in (0,1)$ and $k \geq 0$, we have that*

$$\lim_{n_0\to\infty} \frac{1}{\log(1-p)} \log\left(\frac{Z^+_{n_0}(k)}{Z_{n_0}(k)}\right) \overset{a.s.}{=} k,$$

*where the almost sure limit refers to the probability space on which i.i.d. copies of $Z(k)$, namely $Z_{(i)}(k)$ for $i = 1, 2, \ldots$, are constructed s. t. $Z_{n_0}(k) = \sum_{i=1}^{n_0} Z_{(i)}(k)$for each $n_0 \in \mathbb{N}$.*

**Proof:** It is a well known result that in a Galton-Watson branching process $\mathbb{E}\left(Z^+_{n_0}(k)\right) = n_0(h(1-p))^k$ and $\mathbb{E}\left(Z_{n_0}(k)\right) = n_0 h^k$ (e.g. [42, 10]). Furthermore, under the super-critical assumption, the normalised processes $\mathcal{Z}_{n_0}(k) := Z_{n_0}(k)/(n_0 h^k)$ and $\mathcal{Z}^+_{n_0}(k) := Z^+_{n_0}(k)/(n_0(h(1-p))^k)$ converge almost surely to the nonnegative random variable $\mathcal{Z}$ and $\mathcal{Z}^+$, respectively. As long as the positive-label population does not become extinct, using the Continuous Map-

ping Theorem it follows that

$$\lim_{k\to\infty} \frac{1}{k\log(1-p)} \log\left(\frac{Z^+_{n_0}(k)}{Z_{n_0}(k)}\right) = \lim_{k\to\infty} \frac{1}{k\log(1-p)} \log\left((1-p)^k \frac{\mathcal{Z}^+_{n_0}(k)}{\mathcal{Z}_{n_0}(k)}\right)$$
$$= 1 + \lim_{k\to\infty} \frac{1}{k\log(1-p)} \log\left(\frac{\mathcal{Z}^+}{\mathcal{Z}}\right) = 1.$$

This concludes the proof of the first statement of the proposition.

To prove the second part, notice that $Z_{n_0}(k) = \sum_{i=1}^{n_0} Z_{(i)}(k)$ and $Z^+_{n_0}(k) = \sum_{i=1}^{n_0} Z^+_{(i)}(k)$ where $Z_{(i)}(k)$ and $Z^+_{(i)}(k)$ are i.i.d. copies of $Z(k)$ and $Z^+(k)$, respectively. Denote with $\mathcal{Z}^+(k)$ and $\mathcal{Z}(k)$ the normalised processes $\mathcal{Z}^+_{n_0}(k)$ and $\mathcal{Z}_{n_0}(k)$ when $n_0 = 1$, and assume that $\mathcal{Z}^+_{(i)}(k)$ and $\mathcal{Z}_{(i)}(k)$, for $0 \leq i \leq n_0$ are i.i.d. copies of $\mathcal{Z}^+(k)$ and $\mathcal{Z}(k)$. Given for construction $\mathbb{E}\left(\mathcal{Z}^+(k)\right)$ and $\mathbb{E}\left(\mathcal{Z}(k)\right)$ are equal to 1, using the SLLN and the Continuous Mapping Theorem, we have that

$$\lim_{n_0\to\infty} \frac{1}{\log(1-p)} \log\left(\frac{Z^+_{n_0}(k)}{Z_{n_0}(k)}\right)$$
$$= \lim_{n_0\to\infty} \frac{1}{\log(1-p)} \log\left((1-p)^k \frac{\sum_{i=1}^{n_0} Z^+_{(i)}(k)}{n_0(h(1-p))^k} \frac{n_0 h^k}{\sum_{i=1}^{n_0} Z_{(i)}(k)}\right)$$
$$= k + \lim_{n_0\to\infty} \frac{1}{\log(1-p)} \log\left(\frac{\frac{1}{n_0}\sum_{i=1}^{n_0} \mathcal{Z}^+_{(i)}(k)}{\frac{1}{n_0}\sum_{i=1}^{n_0} \mathcal{Z}_{(i)}(k)}\right)$$
$$= k + \frac{1}{\log(1-p)} \log\left(\frac{\mathbb{E}\left(\mathcal{Z}^+(k)\right)}{\mathbb{E}\left(\mathcal{Z}(k)\right)}\right) = k.$$

□

The above theorem tells us that every time the label-positive population survives, regardless of the value of $p$ we have chosen, the estimator is asymptotic to the average generation $k$ of the population, when the number of divisions increases. Furthermore, according to the second statement of Proposition 5.3.1, even if the cell population is studied for a short timeframe that allows the birth of only few generations, the estimator converges to the value $k$ when $n_0$ increases. This last setting, where $k$ is fixed but $n_0$ varies, is the setting we focus on in the following. Given the probability of extinction decreases when $n_0$ increases, it seems reasonable to think the value of $p$ that allows the estimator to give the best performance depends on the size of the initial population, $n_0$. So, in the following, we denote the probability of label loss with $p_{n_0}$.

To perform a qualitative analysis of the estimator, even in this section, we use the notion of Mean Square Error (MSE). To compute the MSE of the estimator, we need to compute its mean and variance. As in Section 5.2, we approximates these values using the Taylor expansion in (5.3) and (5.4). But, before doing that, we need to introduce a lemma that allows to describe the behaviour of the Covariance between $Z^+(k)$ and $Z(k)$.

**Proposition 5.3.2.** *If $\mathbb{E}(N^2) < \infty$ and under Assumption 5.3.1, we have that*

$$\mathbb{E}\left(Z^+(k)Z(k)\right) = h^{k-1}(1-p)^k\left(\frac{v(1-h^k)}{1-h} + h\right),$$

*where $h = \mathbb{E}(N)$ and $v = \mathbb{E}(N(N-1))$.*

**Proof:** By conditioning on the first generation, we obtain

$$Z(k) = \sum_{\ell=1}^{Z(1)} Z_{(\ell)}(k-1), \quad \text{and} \quad Z^+(k) = \sum_{\ell=1}^{Z^+(1)} Z^+_{(\ell)}(k-1),$$

where the $(Z_{(\ell)}(k-1), Z^+_{(\ell)}(k-1))$ are independent copies of $(Z(k-1), Z^+(k-1))$. If we denote with $f_k(s,r) := \mathbb{E}\left(s^{Z^+(k)}r^{Z(k)}\right)$ and with $g_k(r) := \mathbb{E}\left(r^{Z(k)}\right)$, we obtain the following recurrence relation

$$\begin{aligned} f_k(s,r) &= \sum_{i,j} \mathbb{P}\left(Z^+(1) = i, Z(1) = j\right) \mathbb{E}\left(\prod_{\ell=1}^{i} s^{Z^+_{(\ell)}(k-1)} r^{Z_{(\ell)}(k-1)}\right) \mathbb{E}\left(\prod_{\ell=i+1}^{j} r^{Z_{(\ell)}(k-1)}\right) \\ &= \sum_{i,j} \mathbb{P}\left(Z^+(1) = i, Z(1) = j\right) f_{k-1}(s,r)^i g_{k-1}(r)^{j-i}. \end{aligned}$$

Computing the derivative of the above equation firstly for $s$, secondly for $r$, and evaluating then the expression found in $(s,r) = (1,1)$, after some arrangements we find

$$\begin{aligned} \mathbb{E}\left(Z^+(k)Z(k)\right) = \mathbb{E}\left(Z^+(k-1)\right)\mathbb{E}\left(Z(k-1)\right)\left(\mathbb{E}\left(Z^+(1)Z(1)\right) - \mathbb{E}\left(Z^+(1)\right)\right) \\ + \mathbb{E}\left(Z^+(1)\right)\mathbb{E}\left(Z^+(k-1)Z(k-1)\right). \end{aligned}$$

Using that $\mathbb{E}(Z^+(k)) = (h(1-p))^k$ and $\mathbb{E}(Z(k)) = h^k$ (e.g. [42, 10]), we can solve the above recurrence equation and conclude the proof. □

Using that $\mathbb{E}(Z(k)) = h^k$, $\mathrm{Var}(Z(k)) = h^k(h^k-1)(v+h-h^2)/(h(h-1))$ (e.g. [42, pg. 6]), the respective formulas for $Z^+(k)$, and Proposition 5.3.2

we can proceed to approximate the mean and variance of the estimator. From (5.3) and (5.4), it follows that

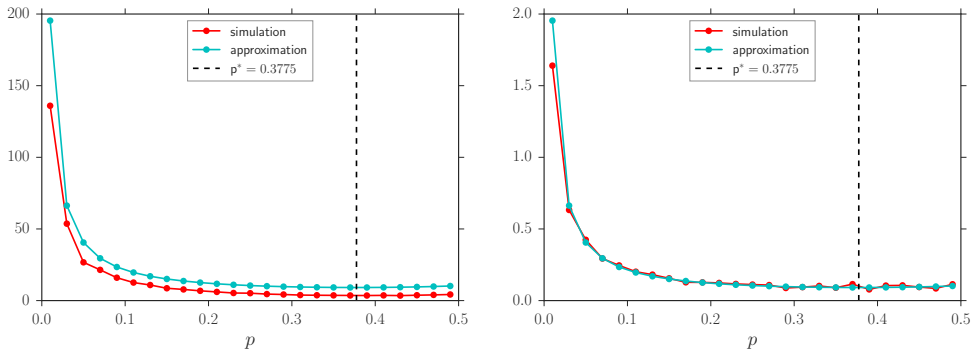
$$\mathbb{E}\left(\frac{1}{\log(1-p)}\log\left(\frac{Z^+_{n_0}(k)}{Z_{n_0}(k)}\right)\right) \approx k$$

$$\mathrm{Var}\left(\frac{1}{\log(1-p)}\log\left(\frac{Z^+_{n_0}(k)}{Z_{n_0}(k)}\right)\right) \approx \frac{1}{n_0\log(1-p)^2}\left[-\frac{(v+h-h^2)(h^k-1)}{h^{k+1}(h-1)}\right.$$
$$\left.+\frac{(1-p)(v+h-(1-p)h^2)((h(1-p))^k-1)}{(h(1-p))^{k+1}(h(1-p)-1)}\right]. \tag{5.19}$$

From the approximations above, we obtain that the MSE coincides with the variance of the estimator expressed above. Given that this approximation is a continuous map in $p$ that converges to $\infty$ when $p$ approaches 0 or 1, we have that a global minimum inside the interval is guaranteed. Despite the simplicity of the expression in (5.19), it seems hard to get an explicit formula for the minimum point. However, for a given initial population size $n_0$ and number offspring distribution $N$, it is possible to numerically determine the location of one of such minimum.

We explore the simple case $N = 2$ a.s.. Given $h = \mathbb{E}(N) = 2$ the range of values for $p$ that allows to remain under the supercritical assumptions is $(0, 1/2)$. Fig. 5.3 shows a comparison between the approximation of the variance that we have found in (5.19) (cyan lines) and the MSE of the estimator computed using 1000 Monte Carlo simulations for each configuration (red lines). We can see that even for one starting cell (Fig. 5.3(a)), $n_0 = 1$, the two lines seem to have the same trend. When we increase the size of the initial population, the fit improves (Fig. 5.3(b)). The point $p^* = 0.3775$, computed using numerical methods, represents the point of minimum of the function of $p$ in (5.19) in the particular case $N = 2$. This doesn't depend on the value of the starting population $n_0$.

In this particular scenario considered in Fig. 5.3, we can check how large the initial population has to be in order for us to have a probability of not getting an estimate below $10^{-2}$ when the probability of label loss is $p^* = 0.3775$. Under Assumption 5.3.1, the probability of extinction of the label-positive population is defined as the solution in $(0, 1)$ of the equation $s = \mathbb{E}\left(s^{N^+}\right)$, with $N^+$ number of label-positive offspring distribution. Given we are considering




(a) $n_0 = 1$ (b) $n_0 = 100$

Figure 5.3: $MSE_n(p_n)$ **in Galton-Watson branching process.** The panels show a comparison between the approximation of the MSE given by the quantity in the RHS of (5.19) (cyan lines) and the MSE computed using 1000 Monte Carlo simulations for each configuration (red lines). In both panels is assumed $N = 2$ . The black dashed lines show the minimum of the MSE approximation in (5.19), computed through numerically methods. The value of this point is $p^* = 0.3775$. In plot (a) we have considered a starting population of $n_0 = 1$, whereas in (b) $n_0 = 100$.

the case $N = 2$, we have that

$$\mathbb{P}\left(Z_{n_0}^{+}(k) = 0\right) \leq \mathbb{P}\left(\lim_{k\to\infty} W_{n_0}^{+}(k) = 0\right) = \left(\frac{p}{1-p}\right)^{n_0} \leq 10^{-2}.$$

from which we obtain that

$$n_0 \geq \frac{\log(10^{-2})}{\log\left(\frac{p}{1-p}\right)} = \frac{\log(10^{-2})}{\log\left(\frac{0.3775}{0.3775-1}\right)} \approx 9 \tag{5.20}$$

So, it suffices that the growth of the population starts with at least 9 cells in order for $p = 0.3775$ to be the point that minimises the MSE ensuring also a probability of not getting an estimate below $10^{-2}$.

In general, for other distributions of the number of offspring $N$, the same method of finding numerical solutions for the minimum of (5.19) and checking the probability of extinction that it determines can be used. Anyway, from what we have seen in Fig. 5.3, it seems that picking up a large value of $p$, i.e. something in the interval $(r, 1-1/h)$ for some $r \in (0, 1-1/h)$, can be a better choice than selecting a small $p$.

## 5.4 Bellman-Harris branching process

In this final section of the chapter, we continue the qualitative study of the estimator in the RHS of (5.1) in order to understand how to select $p$ to minimise the Mean Square Error, but this time under the more general framework of a Bellman-Harris branching process. Compared to the previous section, where a Galton-Watson model was assumed, here we allow for asynchronism in the lifetimes of the cells, which in turn causes heterogeneity in the generation of the cells of the population. We will see that the introduction of this heterogeneity makes the results of the analysis qualitatively similar to the ones found in Section 5.2.2, where the cells were supposed i.i.d. but with random generations. In the following, the assumptions we make within this section.

**Assumption 5.4.1.** The growth of the population under study starts from $n_0$ cells, each of one equipped with a neutral and heritable label that at each division can be either lost, with probability $p$, or passed to the offspring, with probability $1-p$. The lifetime and the number of offspring of a cell are denoted with $L$ and $N$, respectively, and it is assumed that $L$ is non-lattice distributed and $\mathbb{E}(N^2) < \infty$. $Z_{n_0}(t)$ and $Z^+_{n_0}(t)$ denote the size of the entire and label-positive populations generated from $n_0$ ancestors at time $t$ (when $n_0 = 1$ we simplify the notation to $Z(t)$ and $Z^+(t)$). Denote $h := \mathbb{E}(N)$ and $v := \mathbb{E}(N(N-1))$ and suppose that $h(1-p) > 1$, i.e. the mean number of label-positive offspring generated after a label-positive cell division is greater than 1. This puts the two branching process $Z_{n_0}(t)$ and $Z^+_{n_0}(t)$ in the supercritical regime and gives them a positive probability of non-extinction.

Given we want to study the MSE of the estimator, we start by finding approximations for its mean and variance. Using equations (5.3) and (5.4), we obtain that

$$\mathbb{E}\left(\frac{1}{\log(1-p_{n_0})}\log\left(\frac{Z^+_{n_0}(t)}{Z_{n_0}(t)}\right)\right) \approx \frac{1}{\log(1-p_{n_0})}\log\left(\frac{\mathbb{E}(Z^+(t))}{\mathbb{E}(Z(t))}\right), \tag{5.21}$$

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{\log(1-p_{n_0})}\log\left(\frac{Z^+_{n_0}(t)}{Z_{n_0}(t)}\right)\right) \approx \frac{1}{n_0\log(1-p_{n_0})^2}\Big(&\frac{\mathrm{Var}\left(Z^+(t)\right)}{\mathbb{E}\left(Z^+(t)\right)^2} \\ &- 2\frac{\mathrm{Cov}(Z^+(t),Z(t))}{\mathbb{E}\left(Z(t)\right)\mathbb{E}\left(Z^+(t)\right)} + \frac{\mathrm{Var}\left(Z(t)\right)}{\mathbb{E}\left(Z(t)\right)^2}\Big).\end{aligned} \tag{5.22}$$

The first two moments of the processes $Z^+(t)$ and $Z(t)$ that appear in the RHS of (5.21) and (5.22) are quantities well studied in the literature [42, 9]. There are not explicit expressions for them in general, but approximations for large $t$ are available. What we need to investigate is instead the behaviour of the covariance of $Z(t)$ and $Z^+(t)$ for large $t$, which we do in the following Lemma 5.4.1.

**Lemma 5.4.1.** *Under Assumption 5.4.1, we have*

$$\lim_{t\to\infty} \frac{\mathbb{E}\left(Z(t)Z^+(t)\right)}{e^{\alpha(h)t}e^{\alpha(h(1-p))t}} = \frac{(1-p)cc_p v \int_0^\infty e^{-\alpha(h)u}e^{-\alpha(h(1-p))u}d\mathbb{P}(L\le u)}{1-h(1-p)\int_0^\infty e^{-\alpha(h)u}e^{-\alpha(h(1-p))u}d\mathbb{P}(L\le u)}, \quad (5.23)$$

*where*

$$c = \lim_{t\to\infty} \frac{\mathbb{E}\left(Z(t)\right)}{e^{\alpha(h)t}} \quad \textit{and} \quad c_p = \lim_{t\to\infty} \frac{\mathbb{E}\left(Z^+(t)\right)}{e^{\alpha(h(1-p))t}} \quad (5.24)$$

*are positive constants defined in* (2.7).

**Proof:** As in Lemma 2.3.6, in order to prove (5.23), we can find an integral equation solved by $\mathbb{E}\left(Z(t)Z^+(t)\right)$ and conclude using one of the versions of the Renewal Theorem.

An integral equation solved by the Probability Generating Function (PGF) $F(s,r) = \mathbb{E}\left(s^{Z(t)}r^{Z^+(t)}\right)$ of $(Z(t), Z^+(t))$ is given by

$$\begin{aligned} F(s,r,t) =& sr\mathbb{P}(L>t) + p\int_0^t \rho(\mathbb{E}\left(s^{Z(t-u)}\right))d\mathbb{P}(L\le u) \\ &+ (1-p)\int_0^t \rho(F(s,r,t-u))d\mathbb{P}(L\le u) \end{aligned} \quad (5.25)$$

where $\rho(s) = \mathbb{E}\left(s^N\right)$ is the PGF of the offspring number after a cell division. Computing the derivative of (5.25) first for $s$, secondly for $r$, and evaluating it at $(1,1,t)$, we obtain

$$\begin{aligned} \mathbb{E}\left(Z(t)Z^+(t)\right) =& \mathbb{P}(L>t) + (1-p)v\int_0^t \mathbb{E}\left(Z(t-u)\right)\mathbb{E}\left(Z^+(t-u)\right)d\mathbb{P}(L\le u) \\ &+ h(1-p)\int_0^t \mathbb{E}\left(Z(t-u)Z^+(t-u)\right)d\mathbb{P}(L\le u). \end{aligned} \quad (5.26)$$

Dividing this expression by $e^{\alpha(h)t}e^{\alpha(h(1-p))t}$, and denoting with

$$\begin{aligned} K(t) =& \frac{\mathbb{E}\left(Z(t)Z^+(t)\right)}{e^{\alpha(h)t}e^{\alpha(h(1-p))t}}, \\ d\overline{\mathbb{P}}(L\le u) =& h(1-p)e^{-\alpha(h)u}e^{-\alpha(h(1-p))u}d\mathbb{P}(L\le u), \\ f(t) :=& \mathbb{P}(L>t)e^{-\alpha(h)t}e^{-\alpha(h(1-p))t} \\ &+ \frac{(1-p)v}{h(1-p)}\int_0^t \frac{\mathbb{E}(Z(t-u))}{e^{\alpha(t-u)}}\frac{\mathbb{E}(Z^+(t-u))}{e^{\alpha(h(1-p))(t-u)}}d\overline{\mathbb{P}}(L\le u), \end{aligned} \quad (5.27)$$

we have

$$K(t) = f(t) + \int_0^t K(t-u)d\overline{\mathbb{P}}(L \leq u). \tag{5.28}$$

We can notice that $\overline{\mathbb{P}}$ is a defective measure, in fact

$$h(1-p)\int_0^\infty e^{-\alpha(h)u}e^{-\alpha(h(1-p))u}d\mathbb{P}(L \leq u) < h(1-p)\int_0^\infty e^{-\alpha(h(1-p))u}d\mathbb{P}(L \leq u) = 1.$$

We conclude the proof applying Theorem 2.3.3 to (5.28). □

Now that we have information on the limit behaviour of $\mathbb{E}\left(Z^+(t)Z(t)\right)$, we can rewrite the approximations in (5.21) and (5.22).

Using that $\mathbb{E}\left(Z(t)\right) \approx ce^{\alpha(h)t}$ and $\mathbb{E}\left(Z^+(t)\right) \approx c_p e^{\alpha(h(1-p))t}$ for large $t$, with $c, c_p > 0$ constants defined in (2.7), we have that

$$\begin{aligned}\mathbb{E}\left(-\frac{1}{p_{n_0}}\log\left(\frac{Z^+_{n_0}(t)}{Z_{n_0}(t)}\right)\right) &\approx -\frac{1}{p_{n_0}}\log\left(\frac{\mathbb{E}(Z^+(t))}{\mathbb{E}(Z(t))}\right)\\ &\approx -\frac{\log(c_p/c)}{p_{n_0}} + \frac{[\alpha(h)-\alpha(h(1-p))]t}{p_{n_0}}\end{aligned} \tag{5.29}$$

Let's now observe that

$$\begin{aligned}&\frac{\mathrm{Var}\left(Z^+(t)\right)}{\mathbb{E}\left(Z^+(t)\right)^2} - 2\frac{\mathrm{Cov}(Z^+(t),Z(t))}{\mathbb{E}\left(Z(t)\right)\mathbb{E}\left(Z^+(t)\right)} + \frac{\mathrm{Var}\left(Z(t)\right)}{\mathbb{E}\left(Z(t)\right)^2}\\ &\qquad = \frac{\mathbb{E}(Z^+(t)^2)}{\mathbb{E}(Z^+(t))^2} - 2\frac{\mathbb{E}(Z^+(t)Z_1(t))}{\mathbb{E}(Z^+(t))\mathbb{E}(Z(t))} + \frac{\mathbb{E}(Z(t)^2)}{\mathbb{E}(Z(t))^2}.\end{aligned}$$

Using the information we know about the behaviour of the first two moments of $Z^+(t)$ and $Z(t)$ [42, Chapter VI, Sections 16 and 18] and of their covariance (Lemma 5.4.1), we have that for large $t$

$$\frac{\mathbb{E}(Z^+(t)^2)}{\mathbb{E}(Z^+(t))^2} \approx \frac{v(1-p_{n_0})\int_0^\infty e^{-2\alpha(h(1-p_{n_0}))u}d\mathbb{P}(L \leq u)}{1-h(1-p_{n_0})\int_0^\infty e^{-2\alpha(h(1-p_{n_0}))u}d\mathbb{P}(L \leq u)} \tag{5.30}$$

$$\frac{\mathbb{E}(Z^+(t)Z(t))}{\mathbb{E}(Z^+(t))\mathbb{E}(Z(t))} \approx \frac{v(1-p_{n_0})\int_0^\infty e^{-\alpha((1-p_{n_0})h)u}e^{-\alpha(h)u}d\mathbb{P}(L \leq u)}{1-h(1-p_{n_0})\int_0^\infty e^{-\alpha((1-p_{n_0})h)u}e^{-\alpha(h)u}d\mathbb{P}(L \leq u)} \tag{5.31}$$

$$\frac{\mathbb{E}(Z(t)^2)}{\mathbb{E}(Z(t))^2} \approx \frac{v\int_0^\infty e^{-2\alpha(h)u}d\mathbb{P}(L \leq u)}{1-h\int_0^\infty e^{-2\alpha(h)u}d\mathbb{P}(L \leq u)} =: k. \tag{5.32}$$

where we remember that $h := \mathbb{E}\left(N\right)$ and $v := \mathbb{E}\left(N(N-1)\right)$.

The quantities in (5.29),(5.30),(5.31), and (5.32) are not easy to compute in general. One of the few cases where this is possible and at the same time

interesting is the exponential lifetime case. Even if it is well known that the exponential distribution does not describe properly the behaviour of the cell-division cycle, this assumption is largely employed because of the easier mathematical framework that this condition creates.

Before dealing with the general case, where we need to use further approximations, in the next section we continue the study of the problem under the more restricted conditions of a Birth-Death branching process. This model is essentially a generalisation of the Pure-birth process that we have considered in Chapter 4, where $L$ is exponential distributed and the only two outcomes for the cells after division are either 2 or 0 offspring.

### 5.4.1 Birth-Death branching process

Within this section we consider an idealised model that is commonly used in biology and other fields for mathematical convenience, but does not well model what in the experiments has been observed. The assumptions we make are summarised in the following.

**Assumption 5.4.2.** The cell lifetimes are i.i.d. and distributed according to an exponential random variable with mean $\lambda$, i.e. $L \overset{D}{\sim} \text{Exp}(\lambda)$. Concerning the numbers of offspring $N$, we assume that just two realisations are possible: 0 and 2. If $h := \mathbb{E}(N)$, we assume $h(1-p) > 1$, i.e. the mean number of label-positive offspring generated after a label-positive cell division is greater than 1. This condition guarantees the supercritical regime in both processes $Z_{n_0}(t)$ and $Z^+_{n_0}(t)$.

A consequence on the assumption made on $N$ is that $v = h$, with $h := \mathbb{E}(N)$ and $v := \mathbb{E}(N(N-1))$. From the definition of Malthus parameter in (2.4), we have that $\alpha(h) = \lambda(h-1)$. Furthermore, the constants $c$ and $c_p$ defined through the equations (5.24) and (2.7), are both equal to 1. This implies that the first term in (5.29) is null, and in particular

$$\mathbb{E}\left(-\frac{1}{p_{n_0}}\log\left(\frac{Z^+_{n_0}(t)}{Z_{n_0}(t)}\right)\right) \approx \frac{[\lambda(h-1) - \lambda(h(1-p_{n_0})-1)]t}{p_{n_0}} = \lambda h t \qquad (5.33)$$

Furthermore, the integrals that appear in (5.30), (5.31), and (5.32) can be

easily computed when $L \overset{D}{\sim} \text{Exp}(\lambda)$, giving us

$$\frac{\mathbb{E}(Z^+(t)^2)}{\mathbb{E}(Z^+(t))^2} \approx \frac{h(1-p)}{h(1-p)-1}, \frac{\mathbb{E}(Z^+(t)Z(t))}{\mathbb{E}(Z^+(t))\mathbb{E}(Z(t))} \approx \frac{h(1-p)}{h-1}, \frac{\mathbb{E}(Z(t)^2)}{\mathbb{E}(Z(t))^2} \approx \frac{h}{h-1} \tag{5.34}$$

This leads to

$$\text{Var}\left(-\frac{1}{p_{n_0}}\log\left(\frac{Z^+_{n_0}(t)}{Z_{n_0}(t)}\right)\right) \approx \frac{1}{n_0 p_{n_0}} \frac{h[(2h-1)-2hp_{n_0}]}{(h-1)[(h-1)-hp_{n_0}]}. \tag{5.35}$$

Using (5.33) and (5.35) we can approximate the MSE of the estimator with

$$\begin{aligned} MSE_{n_0}(p_{n_0}) = &\left(\mathbb{E}\left(-\frac{1}{p_{n_0}}\log\left(\frac{Z^+_{n_0}(t)}{Z_{n_0}(t)}\right)\right) - \alpha'(h)ht\right)^2 \\ &+ \text{Var}\left(-\frac{1}{p_{n_0}}\log\left(\frac{Z^+_{n_0}(t)}{Z_{n_0}(t)}\right)\right) \approx \frac{1}{n_0 p_{n_0}} \frac{h[(2h-1)-2hp_{n_0}]}{(h-1)[(h-1)-hp_{n_0}]} \end{aligned} \tag{5.36}$$

Computing the derivative in $p_{n_0}$ of the expression above, we obtain

$$\begin{aligned} &\frac{\partial}{\partial p_{n_0}}\left(\frac{1}{n_0 p_{n_0}} \frac{h[(2h-1)-2hp_{n_0}]}{(h-1)[(h-1)-hp_{n_0}]}\right) \\ &\qquad = -\frac{h(2h^2 {p_{n_0}}^2 + 2h(h+1)p_{n_0} + 2h^2 - 3h + 1)}{n_0(h-1){p_{n_0}}^2(h-1-hp_{n_0})^2}, \end{aligned}$$

which is negative for $p_{n_0} \in (0, 1-1/h)$. This means that when the lifetimes of the cells are exponentially distributed, condition on $Z^+_{n_0}(t)$ being positive, a value of $p_{n_0}$ around $1-1/h$ should be the best choice to minimise the MSE of the estimator. Given for any $n_0$ the probability of extinction of the label-positive population increases with $p_{n_0}$, also this time our suggestion is to balance the two effects choosing the highest probability of label loss that still allow $\mathbb{P}\left(Z^+_{n_0}(t)=0\right)$ to be under a certain limit, namely $10^{-2}$. Given the probability of extinction for a supercritical Bellman-Harris branching process, $Z^+(t)$, is given by the solution $s \in (0,1)$ of the equation $s = \mathbb{E}\left(s^{N^+}\right)$, with $N^+$ offspring number distribution of label-positive cells obtained from a label-positive mother, in a Birth-Death branching process we have

$$\mathbb{P}\left(Z^+_{n_0}(t)=0\right) \leq \mathbb{P}\left(\lim_{t\to\infty} Z^+(t)=0\right)^{n_0} = \left(\frac{2-h(1-p_{n_0})}{h(1-p_{n_0})}\right)^{n_0}.$$

So, if we want that the probability of not getting an estimate is below $10^{-2}$, we have to consider the following constraint

$$p_{n_0} \leq 1 - \frac{2}{h(10^{-2/n_0}+1)}.$$

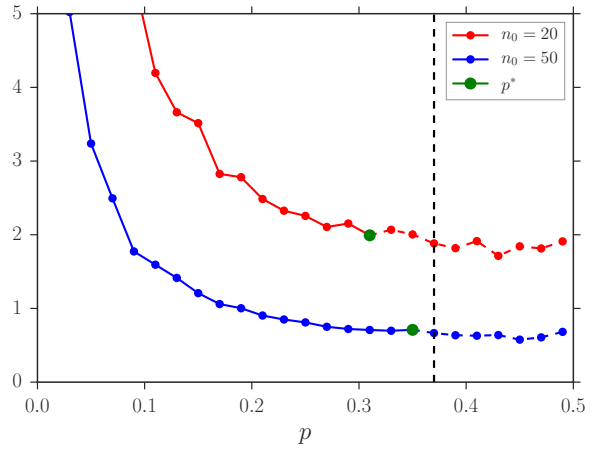



Figure 5.4: $MSE_n(p_n)$ **in a Birth-Death process.** The plot describes the behaviour of the empirical MSE of the estimator, obtained using 1000 Monte Carlo simulations, in function of the probability of label loss $p$. We have assumed an exponential lifetime distribution with mean 9.3 (see [44] for the choice of the mean) and an expected number of offspring $h = 8/5$, i.e. at each division can be generated 0 offspring with a probability of $1/5$. The two curves correspond to different sizes of the initial population, namely $n_0 = 20$ (red line) and $n_0 = 50$ (blue line). Solid lines correspond to range of $p$ that allows a probability of not getting an estimate below $10^{-2}$, whereas dashed lines the opposite. The black dashed line represents the value $p_{max} = 1 - 1/h = 3/8$, maximum value of $p$ that guarantees the supercriticality of the processes $Z^+(t)$ and $Z(t)$ (see Assumption 5.4.2).

Note that the set of solutions of the inequality above can also be empty for small $n_0$. For example, if $n_0 = 1$, for any choice of $p_{n_0}$, the probability of extinction cannot go below $(2 - h)/h$.

In Fig. 5.4.1 we can see that decreasing monotone trend of the MSE of the estimator, computed using Monte Carlo simulations for $h = 8/5$. Two different sizes of initial population have been considered, $n_0 = 20$ (red line) and $n_0 = 50$ (blue line), and the points $p_{n_0} = 1 - 2/(h(10^{-2/n_0} + 1))$ have been highlighted (green markers). We can see that the choice $p_{n_0} = 1 - 2/(h(10^{-2/n_0} + 1))$, when this quantity is positive, seems to be a good option.

### 5.4.2 General lifetimes

After having explored the Birth-Death branching process subcase, we deal with the general one. In order to do that we assume that the minimum of the MSE of the estimator, when $n_0$ grows, drops to 0. This allows us to

approximate the position of the point of minimum by finding the one of the second-order Taylor expansion of the MSE.

Using the definition of $c_p$ in (2.7), we can compute the Taylor expansion of the first term in the RHS of (5.29), obtaining

$$\log\left(\frac{c_p}{c}\right) = b_1 p_{n_0} + b_2 {p_{n_0}}^2 + O({p_{n_0}}^3)$$

for some constants $b_1,\ b_2 \in \mathbb{R}$. Notice that they can also be 0, as in the Exponential lifetime case.

Given $\alpha(x)$ is analytic around $x = h > 1$ [116, Proposition 1], there exists a $\bar{n} > 0$ s.t. for $n_0 > \bar{n}$ we have

$$\alpha(h(1-p_{n_0})) = \alpha(h) - \alpha'(h)hp_{n_0} + \frac{\alpha''(h)}{2}(hp_{n_0})^2 + O((hp_{n_0})^3).$$

So, we have that

$$-\frac{1}{p_{n_0}}\log\left(\frac{\mathbb{E}(Z^+(t))}{\mathbb{E}(Z(t))}\right) \approx b_1 + \alpha'(h)ht + \left(b_2 + \frac{\alpha''(h)}{2}h^2 t\right)p_{n_0} + O({p_{n_0}}^2). \tag{5.37}$$

Given the fact that numerators and denominators in the RHS of the previous equations are analytic functions of $p_{n_0}$ around 0, and divisions among analytic functions are still analytic (as long as the denominator is different from 0) [98, pg. 197-198], the two fractions in the right-hand side of (5.30) and (5.31), are analytic functions in $p_{n_0}$ too.

So, for $p_{n_0}$ in a neighbourhood of 0, we can approximate them with their first-order Taylor approximation, obtaining

$$\begin{aligned}
\frac{\mathbb{E}(Z_1^+(t)^2)}{\mathbb{E}(Z_1^+(t))^2} \approx & k + p_{n_0}\frac{v\int_0^\infty (2\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\le u)}{[1-h\int_0^\infty (2\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\le u)]^2},\\
\frac{\mathbb{E}(Z_1^+(t)Z_1(t))}{\mathbb{E}(Z_1^+(t))\mathbb{E}(Z_1(t))} \approx & k + p_{n_0}\frac{v\int_0^\infty (\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\le u)}{[1-h\int_0^\infty (\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\le u)]^2}.
\end{aligned}$$

Using these results, equation (5.38) can be rewritten as

$$\mathrm{Var}\left(-\frac{1}{p_{n_0}}\log\left(\frac{Z_{n_0}^+(t)}{Z_{n_0}(t)}\right)\right) \approx \frac{k_2}{n_0 p_{n_0}} + \frac{1}{n_0}O(1), \tag{5.38}$$

where

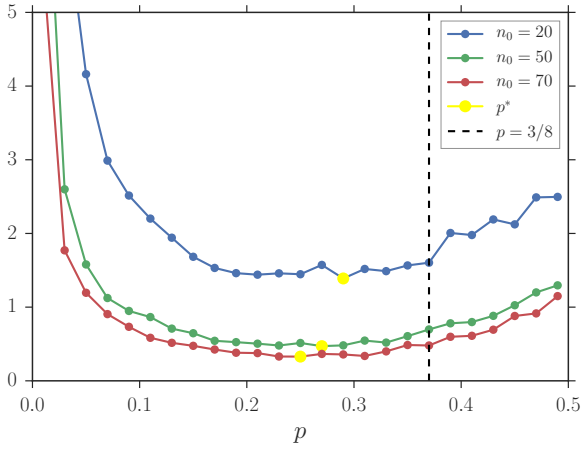
$$k_2 := \frac{v\int_0^\infty (2\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\leq u)}{[1-h\int_0^\infty (2\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\leq u)]^2} - \frac{2v\int_0^\infty (\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\leq u)}{[1-h\int_0^\infty (\alpha'(h)hu-1)e^{-2\alpha(h)u}d\mathbb{P}(L\leq u)]^2}.$$

So, we have that

$$MSE_{n_0}(p_{n_0}) \approx \frac{k_2}{n_0 p_{n_0}} + b_1^2 + 2b_1\left(b_2+\frac{\alpha''(h)}{2}h^2t\right)p_{n_0} + \left(b_2+\frac{\alpha''(h)}{2}h^2t\right)^2 {p_{n_0}}^2 \tag{5.39}$$

We are not able to analytically obtain information on the minimum of the above expression, given the expressions for the constants $k_2$, $b_1$, $b_2$, and the function $\alpha''(h)$ are not easy to handle. What we can do is to numerically study a particular case and draw conclusions from that.

In Fig. 5.4.2, we can see the behaviour of the MSE in a Bellman-Harris branching process with offspring distribution $N$ s.t. $\mathbb{P}(N=0) = 1/5 = 1-\mathbb{P}(N=2)$, and lifetimes lognormal distributed with mean 9.3 and std 2.54 (see [44] for the parameterisation). Three sizes for the initial population are considered, $n_0 = 20$ (blue line), $n_0 = 50$ (green line), and $n_0 = 70$ (red line), and for each parameterisation 1000 Monte Carlo simulations have been used to describe the MSE. We can see how the minimum points of these curves (yellow markers) seem to depend on $n_0$ and decrease when $n_0$ increases. The rate of this decay is not clear, but it is likely that it depends on the first non null term that appears in the polynomial approximation of the MSE in (5.39). In fact, in the special case of a Birth-Death process, we have for example that $\alpha(h) = \lambda(1-h)$ and so $\alpha''(h) = 0$. If we compare (5.39) with (5.36), we can see that, for large $n_0$, all the terms $O(1)$ that appear in (5.39) can be disregarded, explaining the decreasing behaviour of the MSE. However, in the case that the terms $O(1)$ that appear in (5.39) cannot be ignored, we think that the value of $p$ that minimises the MSE drops to 0 when $n_0$ increases.




(a) $n_0 = 1$

Figure 5.5: $MSE_n(p_n)$ **in a Bellman-Harris branching process.** The panel describes the empirical behaviour of the MSE of a Bellman-Harris branching process with lognormal distributed lifetimes (with mean 9.3 and std 2.54) and number of offspring distribution $N$ s.t. $\mathbb{P}(N = 0) = 1/5 = 1 - \mathbb{P}(N = 2)$, for different sizes of the initial population, $n_0$. For each parameterisation 1000 Monte Carlo simulations have been used, simulating the growth of cell populations for a time frame of $t = 36$ hours. Blue, green, and red lines correspond to the initial populations $n_0 = 20$, $n_0 = 50$, and $n_0 = 70$, respectively. The yellow markers denote the points with the lowest value for the MSE. The black dashed line denotes the value $p_{max} = 3/8$, maximum value of $p$ that can be considered in order for the processes $Z^+(t)$ and $Z(t)$ to remain in the supercritical regime (see Assumption 5.4.2).

CHAPTER 6

# Average generation in a Renewal Process

## 6.1 Introduction

In the previous chapters of this thesis, we have provided additional mathematical support for the methodology proposed by Weber et al. in [116] to estimate the average generation of a growing cell population (Chapter 2), but we have also focused on aspects left behind from the original analysis, such as the generalisation to a multi-type setting (Chapter 3) and a qualitative analysis of the estimator as a function of the parameters of the model (Chapter 5). Using a simplification of the model, we have also seen that the average generation is a quantity well behaved over time, given its variance is not exploding but converging to a constant (Chapter 4).

In all these situations, even if the models considered sometimes differed from each other, we have always assumed an expanding cell population, that we have obtained considering for each cell an average number offspring per division greater than 1. There are biological systems that to a first approximation can be described by a model that contemplates only a growing phase, such as cancer growth dynamics [51, 40, 27], but many others have more complex population dynamics that are the result of the interchange of expansion, homeostatic, and contraction phases.

In this chapter we want to see what happens when environment boundaries,

lack of stimulatory signals or nutrients force the population to stop its growth, staying in an homeostatic stage, i.e. keeping the size approximately constant. We are interested in knowing whether the formulas in (2.37) and (2.38) are appropriate in this case, and if so how robust they are.

The main reason a cell population maintains its size constant is to preserve the correct functioning of the body [34] in absence of critical situations that alter the normal equilibrium. This is the case for example after an adaptive immune response, when a constant population of memory B and T cells enables an organism to keep ready its defences against further exposure to the same threat (see Fig. 1.4). Similarly, even if this is still subject to debate, somatic stem cells seem to keep their number constant in the human body using asymmetric cell divisions, i.e. producing one cell of the same type (self-renewal) and another one with a smaller potency, with exceptions when the stem-cells pool is first established during development and when they are regenerated after injuries [60, 79]. Failure in homoeostatic regulation can cause unregulated division of cells and is symptomatic of major health problems, such as cancer or autoimmune diseases [22].

In Chapter 2, we have used a super-critical Bellman-Harris branching process to mathematically describe the growth of an expanding cell population. Allowing the expected offspring after each division to be equal to one, we might think that a critical Bellman-Harris branching processes can do the same job as an homeostatic population. This is not the appropriate model to describe such dynamics, however, due to the fact that a nondegenerate Bellman-Harris branching process, even a critical one, can only have two possible fates: infinite growth or extinction [42]. As keeping the size of the population approximately constant is one of the key features we want to represent, we will use a model that, despite its simplicity, allows us to conclude strong results concerning the average generation. In particular, we will see the population as a sum of renewal processes, one for each member of the starting population.

The main result of the chapter, Theorem 6.2.1, constitutes the equivalent of Corollary 2.3.12 for the homeostatic case. This time the strong path by path relationship between the average generation of the population and the proportion of label-positive cells is a consequence of Renewal Theory [93] and sub-critical Bellman-Harris branching processes [42], necessary to describe the

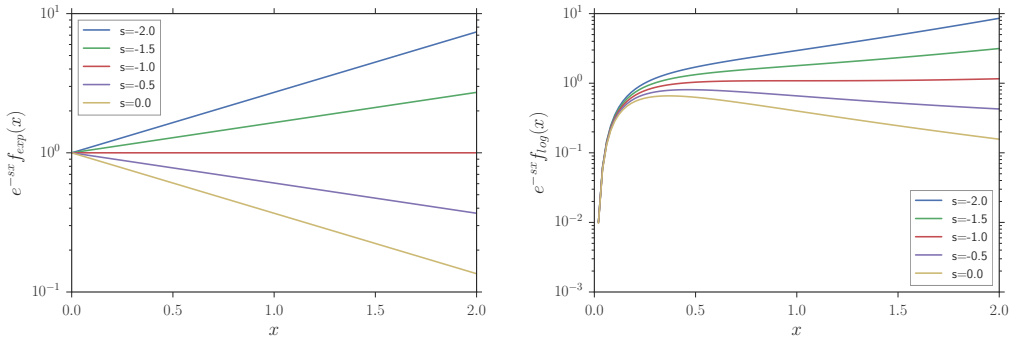
delabelling of the population.

## 6.2 The renewal model

We now describe the model we use throughout this chapter, adopting notation that is coherent with that introduced in the previous chapters, in particular Chapter 2.

Suppose we have an initial population of $n_0$ cells, where every cell, independently of each other, after a random lifetime $L$ is substituted by a newborn cell, so that the size of the population is kept constant. Let $L$ be strictly positive and non-lattice distributed. A proportion $\gamma$ of the starting population is equipped with a neutral label, i.e. one that doesn't change the population dynamics, which is passed on to the newborns after each substitution with probability $1-p$. We allow for $\gamma \neq 1$ because we can think of the initial collection of cells as the final population obtained after an expansion phase, in which some of the cells have already lost their labels. Even if a higher proportion of initial label-positive cells improves the quality of the average generation estimates, the value of $\gamma$ does not affect the asymptotic result that we find in Theorem 6.2.1, as long as $\gamma \neq 0$. According to the assumptions, we have that the label survives in the lineage of each of the starting cells equipped with it for a time $L^+ = \sum_{i=0}^{Y} L^{(i)}$, where $\{L^{(i)}\}_{i\geq 0}$ are i.i.d. copies of $L$ and $Y \overset{D}{\sim} \text{Geo}(p)$, i.e. $Y$ has the same distribution of a geometric random variable with parameter $p$.

Our desire is to understand if the proportion of label-positive cells in the population at a particular time $t$ can be used to estimate the average generation of the entire population at the same instant, as we have seen in Chapters 2 and 3 for a single and two-type supercritical Bellman-Harris branching process population, respectively. To do so, we need to study the behaviour of $Z^+_{n_0}(t)$, the number of label-positive cells at time $t$, where the initial size of the population is $n_0$ and $Z^+_{n_0}(0) = \lfloor \gamma n_0 \rfloor$. The population dynamics of the label-positive population can be described by a sub-critical Bellman-Harris branching process with starting population $\lfloor \gamma n_0 \rfloor$, offspring distribution $N \overset{D}{\sim}$ Ber(1-p), and lifetime $L$. As the expected number of label-positive cells obtained as a consequence of a label-positive cell division is $h^+_p := 1-p < 1$, from the theory of sub-critical branching processes [42] we already know that $\lim_{t\to\infty} Z^+_{n_0}(t) = 0$




(a) $L \overset{D}{\sim} \text{Exp}(1)$ (b) $\log(L) \overset{D}{\sim} \mathcal{N}(0,1)$

Figure 6.1: The plot illustrates the behaviour of the integrand of the Laplace transform $\mathbb{E}(e^{-sL})$, for different values of $s$ and different distributions of $L$. (a) If $L \overset{D}{\sim} \text{Exp}(1)$, $e^{-sx} f_{exp}(x)$ is an exponential function, where $f_{exp}(x)$ is the density function of $L$ (notice the logarithmic scale for the y-axis). The plot suggests that the quantity $\mathbb{E}(e^{-sL})$ is finite only for $s \in (-1, \infty]$ (we can easily check that doing the computations). (b) When $L$ is lognormal distributed with $\log(L) \overset{D}{\sim} \mathcal{N}(0,1)$, we obtain that for all negative $s$, the quantity $e^{-sx} f_{log}(x)$, where $f_{log}(x)$ is the density function of $L$, is either increasing or decreasing slower than an exponential. This has as consequence that $\mathbb{E}(e^{-sL})$ is not finite for all negative $s$.

almost surely (a.s.). In order to describe the decrease rate of $Z^+_{n_0}(t)$, we use the concept of Malthusian parameter already introduced in (2.4), i.e. we define $\alpha^+ = \alpha(h^+_p)$ as the solution of

$$h^+_p \mathbb{E}(e^{-\alpha^+ L}) = 1, \tag{6.1}$$

where $h^+_p = 1 - p$. The first difference we can notice with respect to the expanding case treated in Chapter 2, where $h^+_p > 1$, is that now this solution does not necessarily exist. In particular, its existence depends on the tail of the cell's lifetime distribution $L$. Indeed, $h^+_p < 1$ forces a potential solution $\alpha^+$ of (6.1) to be negative, but $\mathbb{E}(e^{-sL})$, for $s < 0$, may be not even finite when the tail of $L$ is not decreasing fast enough to compensate for the exponential in the integral (see Figure 6.1). We are in this situation for example when $L$ follows a lognormal distribution, whose Laplace transform $\mathbb{E}(e^{-sL})$ is not finite, and so not defined, for every negative $s$.

What we need to assume in order to assure the existence and uniqueness of the solution $\alpha^+$ to (6.1) is that the Laplace transform of the lifetime $\mathbb{E}(e^{-sL})$ is finite for $s$ in a neighbourhood of zero, i.e.

**Assumption 6.2.1.** The lifetime $L$ is such that $\mathbb{E}(e^{-sL}) < \infty$ for $s \in [s_{\min}, \infty)$, with $s_{\min} < 0$.

As we will consider values of $p$ arbitrarily small and $h_p^+ = 1-p$ converges from below to one for $p$ tending to zero, under Assumption 6.2.1, we can always suppose, for a $p$ small enough, that $1 = \mathbb{E}(e^{-0L}) < 1/h_p^+ < \mathbb{E}(e^{-s_{\min}L})$. Given $\mathbb{E}(e^{-sL})$ is real analytic on the interior of the domain on which it is finite, i.e. for $s > s_{\min}$, and it is decreasing in $s$, we can apply the real analytic version of the Implicit Function Theorem (e.g. Theorem 2.5.3 [67]) concluding that the solution $\alpha(h_p^+)$ to equation (6.1) exists. The kind of behaviour that we are requiring from $L$ is also referred in the literature as being not heavy-tailed distributed [31, pg. 2].

Sometimes within this chapter we will need to reference the solution of (6.1) with a quantity $h > h_p^+$ instead of $h_p^+$. When this will be the case, this solution will be denoted $\alpha(h)$, i.e. we will see the Malthusian parameter as a function of $h$. The argument in the previous paragraph allows us to say that if $h > h_p^+$, then $\alpha(h)$ exists. Furthermore, as we have already said in Chapter 2, the function $\alpha(h)$ is also real analytic [116, Proposition 1] and such that

$$\alpha'(h) = \frac{1}{h^2 \int_0^{+\infty} u e^{-\alpha u} d\mathbb{P}(L \leq u)}. \tag{6.2}$$

As in the previous chapters, we call the generation of a cell the number of divisions that led to that cell, where 0 is assumed to be the generation of the initial member of the population. Denoting with $G_{n_0}(t)$ the total generation process given $n_0$ starting cells, i.e. the sum of the generations of all the cells alive at time $t$, the initial set up is given by $G_{n_0}(0) = 0$.

Using a combination of results on renewal and Bellman-Harris processes, we obtain the following theorem, analogous for the homeostatic case to Corollary 2.3.12.

**Theorem 6.2.1.** *Under Assumption 6.2.1, we have*

$$\lim_{t\to\infty} \frac{G_{n_0}(t)}{tn_0} \overset{a.s.}{=} \frac{1}{\mathbb{E}(L)}, \quad \lim_{t\to\infty} \lim_{n_0\to\infty} -\frac{1}{pt} \log\left(\frac{Z_{n_0}^+(t)}{n_0}\right) \overset{a.s.}{=} -\frac{\alpha(h_p^+)}{p}, \tag{6.3}$$

*and*

$$\lim_{p\to 0} -\frac{\alpha(h_p^+)}{p} = \frac{1}{\mathbb{E}(L)}.$$

**Proof:** We start by proving the second equality in (6.3). The proportion of label-positive cells at time $t$ is given by

$$\frac{Z^+_{n_0}(t)}{n_0} \overset{a.s.}{=} \frac{1}{n_0}\sum_{i=1}^{\lfloor \gamma n_0 \rfloor} \mathbb{1}_{\{L_i^+ > t\}} \overset{a.s.}{=} \frac{\lfloor \gamma n_0 \rfloor}{n_0}\frac{1}{\lfloor \gamma n_0 \rfloor}\sum_{i=1}^{\lfloor \gamma n_0 \rfloor} X_i \tag{6.4}$$

where $L_i^+$ and $X_i$ are i.i.d. copies of $L^+$ and $\text{Ber}(\mathbb{P}(L^+ > t))$, respectively. Using the Strong Law of Large Numbers (SLLN) (e.g. [30, pg. 259]) in (6.4), we obtain that for every $t \geq 0$

$$\lim_{n_0\to\infty} \frac{Z^+_{n_0}(t)}{n_0} \overset{a.s.}{=} \gamma \mathbb{P}(L^+ > t).$$

From [10, Theorem 1, pg. 162] or [59] we know that

$$\lim_{t\to\infty} e^{-\alpha(h_p^+)t}\mathbb{P}(L^+ > t) = c, \tag{6.5}$$

where $\alpha(h_p^+)$ is defined in (6.1) and $c \in (0,1)$, from which follows that $\lim_{t\to\infty} \log(\mathbb{P}(L > t))/t = \alpha(h_p^+)$. So, using the Continuous Mapping Theorem (e.g. [103, pg. 24]), we can conclude that

$$\lim_{t\to\infty}\lim_{n_0\to\infty} -\frac{1}{pt}\log\left(\frac{Z^+_{n_0}(t)}{n_0}\right) \overset{a.s.}{=} -\frac{\alpha(h_p^+)}{p}.$$

Remembering that $h_p^+ = 1-p$ and $\alpha(1) = 0$, we have

$$\lim_{p\to 0} -\frac{\alpha(h_p^+)}{p} = \lim_{p\to 0}\frac{\alpha(1)-\alpha(1-p)}{p} = \alpha'(1) \overset{(6.2)}{=} \frac{1}{\mathbb{E}(L)}. \tag{6.6}$$

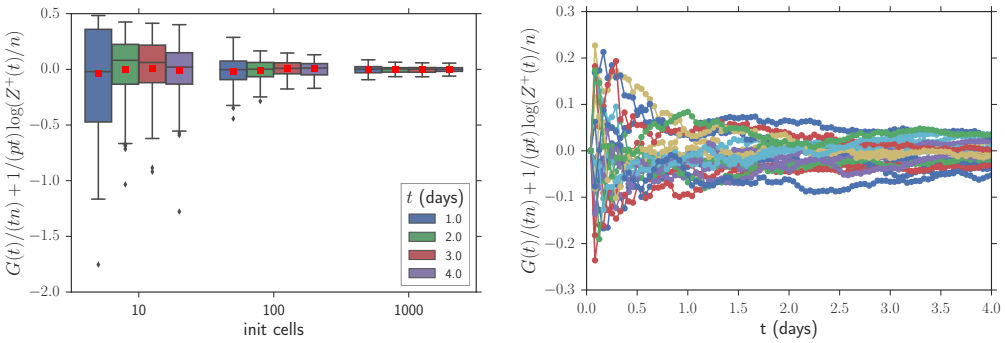
Let's now take a look to the first equality in (6.3). If $R_i(t)$ denotes the renewal process that counts the number of descendants generated by the ancestor $i$, we have for every $n_0 \in \mathbb{N}$

$$\lim_{t\to\infty}\frac{G_{n_0}(t)}{tn_0} \overset{a.s.}{=} \lim_{t\to\infty}\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{R_i(t)}{t} \overset{a.s.}{=} \frac{n_0}{n_0\mathbb{E}(L)} = \frac{1}{\mathbb{E}(L)}, \tag{6.7}$$

where we have used that $\lim_{t\to\infty} R_i(t)/t = 1/\mathbb{E}(L)$ a.s. [93, Theorem 3.3.2, pg.189]. Equation (6.7) is still valid even if we consider a very large $n_0$ before taking the limit $t \to \infty$. In fact, for the SLLN

$$\lim_{t\to\infty}\lim_{n_0\to\infty}\frac{G_{n_0}(t)}{tn_0} = \lim_{t\to\infty}\frac{1}{t}\lim_{n_0\to\infty}\frac{\sum_{i=1}^{n_0} R_i(t)}{n_0} = \lim_{t\to\infty}\frac{\mathbb{E}(R_1(t))}{t} \overset{a.s.}{=} \frac{1}{\mathbb{E}(L)}, \tag{6.8}$$

where we have used that $\lim_{t\to\infty}\mathbb{E}(R_1(t))/t = 1/\mathbb{E}(L)$ [93, Theorem 3.3.3, pg.191].




(a) (b)

Figure 6.2: **Difference between average generation and estimator divided by $t$.** The plots in the figure are obtained using 100 Monte Carlo simulations describing the evolution of a homeostatic cell population. The dynamic of each cell in the initial population is described by a renewal process with lifetime following a lognormal distribution with mean 9.3 hours and std 2.54. Each member of the initial pool of cells is equipped with a neutral label, $\gamma = 1$, that is lost after division with probability $p = 0.01$. (a) For different sizes of the initial population, the difference between $G_{n_0}(t)/(tn_0)$ and $-1/(pt)\log\left(Z^+_{n_0}(t)/n_0\right)$ is studied. Simulated data after each day for 4 days are showed using a set of boxplots. (b) 20 paths describing the same difference are showed for a time range of four days. In the plot, the starting population of each population is made up of 1000 cells.

Combining (6.6) and (6.7), we obtain (6.3). □

A graphical representation of the results in Theorem 6.2.1 can be found in Fig. 6.2, where an homeostatic cell dynamic, with lognormal lifetimes with mean 9.3 hours and standard deviation 2.54 (see [44] for the choice of the parameters), has been recreated with the use of computer simulations. In the first of these plots, Fig. 6.2(a), the difference between the average generation and the estimator, both divided by time, is displayed using a boxplot at the end of each day for 4 days, for different sizes of the starting population. The contraction of the boxplots to a single point, both in time and in the number of initial cells, suggests almost sure convergence of $G_{n_0}(t)/(tn_0)$ and $-1/(pt)\log\left(Z^+_{n_0}(t)/n_0\right)$, for $n_0$ and $t$ larges, to the same quantity. The same can be concluded from Fig. 6.2(b), where the path-by-path behaviour of the estimation's error over time is displayed for $n_0 = 1000$. The fact that each of the 20 paths showed looks to be converging to 0 when time advance supports the idea that the error over time converges to 0 with probability 1.

It is worth noticing that (6.3) is still true even if the lifetimes of the first $k$ generations of cells, $k \in \mathbb{N}$, follow $k$ different distributions from the cells that follow. For example, let us suppose that we are in the situation where an expansion phase is followed by an homeostatic phase due to the occurrence of an event at time $t^*$ that changes the statistics of the population dynamics. Let us denote by $L^{(i)}$ the common lifetime distribution of the cells in generation $i$. Conditioned on $\{t \geq t^*\}$, the residual lifetime $L^{(0)}$ of the cells alive at time $t^*$, that we consider in generation 0, is different from the others, exception made if $L$ follows an exponential distribution. An useful generalisation of Theorem 6.2.1 that suits these kind of applications is the following:

**Corollary 6.2.2.** *Let $L^{(i)}$, $i \in \mathbb{N}$, be independent each other and s.t. the first $k$ are not necessarily equally distributed to the other $L^{(i)}$, $i \geq k+1$, which follow a common distribution $L$. If all the $L^{(i)}$, $0 \leq i \leq k$, satisfy the hypothesis made in Assumption 6.2.1 for $L$, then equation* (6.3) *is still valid.*

**Proof:** All the results proved in Theorem 6.2.1 are still valid in this new case. The only result we are going to check is the one in (6.5) for $k = 1$, as the extension to any finite value of $k$ is an easy adaptation.

Let's denote $L^* = \sum_{i=0}^{Y} L^{(i)}$ the new label lifetime, where $\{L^{(i)}\}_{i \geq 1}$ are i.i.d. copies of $L$, independent from $L^{(0)}$, the lifetime of members of the initial population. From the fact that $L^{(0)}$ satisfies Assumption 6.2.1 and in particular from $\mathbb{E}(e^{-\alpha^+ L^{(0)}}) < \infty$, with $\alpha^+ < 0$ defined in (6.1), using integration by parts, we have

$$\lim_{t\to\infty} e^{-\alpha^+ t}\mathbb{P}(L^{(0)} > t) = \int_0^\infty e^{-\alpha^+ u} d\mathbb{P}(L^{(0)} > u) - \int_0^\infty \alpha^+ e^{-\alpha^+ u}\mathbb{P}(L^{(0)} > u)du$$
$$+ \lim_{t\to 0} e^{-\alpha^+ t}\mathbb{P}(L^{(0)} > t) = c_0 < \infty.$$

Using the memorylessness of the geometric distribution and the fact that, given $Y > 0$, $\widetilde{L}^+ := \sum_{i=1}^{Y} L^{(i)} \overset{D}{\sim} L^+$, with $L^+$ as in Theorem 6.2.1, we have

that

$$\begin{aligned}
&\lim_{t\to\infty} e^{-\alpha^+ t}\mathbb{P}(L^* > t) = \lim_{t\to\infty} e^{-\alpha^+ t}\mathbb{P}(L^* > t|Y = 0)P(Y = 0)+ \\
&\quad + \lim_{t\to\infty} e^{-\alpha^+ t}\mathbb{P}(L^* > t|Y > 0)P(Y > 0) \\
&= \lim_{t\to\infty} e^{-\alpha^+ t}\mathbb{P}(L^{(0)} > t)p + \lim_{t\to\infty} e^{-\alpha^+ t}\mathbb{P}(\widetilde{L}^+ > t - L^{(0)})(1-p) \\
&= c_0 p + \lim_{t\to\infty}(1-p)\int_0^\infty e^{-\alpha^+(t-u)}\mathbb{P}(\widetilde{L}^+ > t-u)e^{-\alpha^+ u}d\mathbb{P}(L^{(0)} \le u) \\
&= c_0 p + (1-p)\int_0^\infty c e^{-\alpha^+ u}d\mathbb{P}(L^{(0)} \le u) \\
&= c_0 p + c(1-p)\mathbb{E}(e^{-\alpha^+ L^{(0)}}) < \infty,
\end{aligned}$$

where we swapped the symbols of limit and integral and used $c$ defined in (6.5). □

The importance of Theorem 6.2.1 is that it enlarges the range of applicability of the random delabelling average generation estimator to a population dynamic that is different from expansion: homeostasis. Given cell populations are normally able to switch between expansion, homeostasis, and contraction in a response to environmental signals they receive, this study provides a further important piece toward the applicability of this method in more complex population dynamics. Indeed, Corollary 6.2.2 shows the robustness of the method for different initial lifetimes, opening to a description in which the homeostasis is just a secondary phase of the overall population dynamic.

# Bibliography


[1] K. Akashi, D. Traver, T. Miyamoto, and I. L. Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature, 404(6774):193, 2000. 49

[2] O. Akinduro, T. S. Weber, H. Ang, M. L. R. Haltalli, N. Ruvio, D. Duarte, N. M. Rashidi, E. D. Hawkins, K. R. Duffy, and C. Lo-Celso. Proliferation dynamics of acute myeloid leukaemia and haematopoietic progenitors competing for bone marrow space. Nat. Commun., 9:519, 2018. 21, 49

[3] B. Alberts, A. Johson, J. Lewis, M. Raff, and K. Roberts. Molecular Biology of the Cell. Garland Science, 4th edition, 2002. 12

[4] R. C. Allsopp, H. Vaziri, C. Patterson, S. Goldstein, E. V. Younglai, A. B. Futcher, C. W. Greider, and C. B. Harley. Telomere length predicts replicative capacity of human fibroblasts. Proc. Natl. Acad. Sci. U.S.A, 89(21):10114–10118, 1992. 13

[5] T. Antal and P. L. Krapivsky. Exact solution of a two-type branching process: models of tumor progression. Journal of Statistical Mechanics: Theory and Experiment, 2011(08):P08018, 2011. 54

[6] S. Asmussen. A probabilistic look at the wiener–hopf equation. SIAM review, 40(2):189–201, 1998. 26, 31, 34

[7] S. Asmussen and H. Hering. Branching processes., 1983. 18

[8] K. B. Athreya and P. Jagers. Classical and modern branching processes, volume 84. Springer Science & Business Media, 2012. 18

[9] K. B. Athreya and N. Kaplan. Convergence of the age distribution in the one-dimensional supercritical age-dependent branching process. Ann. Probability, 4(1):38–50, 1976. 90, 97

[10] K. B. Athreya and P. E. Ney. Branching processes. Courier Corporation, 1972. 14, 18, 24, 91, 93, 110

[11] N. Bacaër. A short history of mathematical population dynamics. Springer Science & Business Media, 2011. 14

[12] R. Bellman and T. E. Harris. On the theory of age-dependent stochastic branching processes. Proceedings of the National Academy of Sciences, 34(12):601–604, 1948. 16

[13] J. E Bestman, J. Lee-Osbourne, and H. T. Cline. In vivo time-lapse imaging of cell proliferation and differentiation in the optic tectum of xenopus laevis tadpoles. Journal of Comparative Neurology, 520(2):401–433, 2012. 10

[14] J. K. Blitzstein and J. Hwang. Introduction to probability. Chapman and Hall/CRC, 2014. 63

[15] J. M. Borwein and S. B. Lindstrom. Meetings with lambert w and other special functions in optimization and analysis. Pure and Applied Functional Analysis, 1(3):361–396, 2016. 85

[16] V. R. Buchholz, M. Flossdorf, I. Hensel, L. Kretschmer, B. Weissbrich, P. Gräf, A. Verschoor, M. Schiemann, T. Höfer, and D. H. Busch. Disparate individual fates compose robust CD8+ T cell immunity. Science, 340(6132):630–635, 2013. 4, 75

[17] C. A. Carlson, A. Kas, R. Kirkwood, L. E. Hays, B. D. Preston, S. J. Salipante, and M. S. Horwitz. Decoding cell lineage from acquired mutations using arbitrary deep sequencing. Nat. Methods, 9(1):78–80, 2012. 13

[18] G. Casella and R. L. Berger. Statistical inference, volume 2. Duxbury Pacific Grove, CA, 2002. 43

[19] S. N Catlin, L. Busque, R. E. Gale, P. Guttorp, and J. L. Abkowitz. The replication rate of human hematopoietic stem cells in vivo. Blood, 117(17):4460–4466, 2011. 1

[20] G. M. Cooper and R. E. Hausman. The cell: a molecular approach. Sinauer Associates, 4th edition, 2007. 55

[21] R. M. Corless, G. H. Gonnet, D. E.G. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambertw function. Advances in Computational mathematics, 5(1):329–359, 1996. 84, 86

[22] B. Crimeen-Irwin, K. Scalzo, S. Gloster, P.L. Mottram, and M. Plebanski. Failure of immune homeostasis-the consequences of under and over reactivity. Curr. Drug. Targets Immune Endocr. Metabol. Disord., 5(4):413, 2005. 106

[23] J. H. Curtiss. A note on the theory of moment generating functions. The Annals of Mathematical Statistics, 13(4):430–433, 1942. 80

[24] R. J. De Boer and A. S. Perelson. Quantifying T lymphocyte turnover. J. Theor. Bio., 327:45–87, 2013. 7

[25] E. K. Deenick, J. Hasbold, and P. D. Hodgkin. Switching to IgG3, IgG2b, and IgA is division linked and independent, revealing a stochastic framework for describing differentiation. J. Immunol., 163(9):4707–4714, 1999. 50

[26] K. R. Duffy, C. J. Wellard, J. F. Markham, J. H. S. Zhou, R. Holmberg, E. D. Hawkins, J. Hasbold, M. R. Dowling, and P. D. Hodgkin. Activation-induced B cell fates are selected by intracellular stochastic competition. Science, 335(6066):338–341, 2012. 7, 50

[27] R. Durrett, J. Foo, K. Leder, J. Mayberry, and F. Michor. Evolutionary dynamics of tumor progression with random fitness values. Theoretical population biology, 78(1):54–66, 2010. 105

[28] B. Efron and R. J. Tibshirani. An introduction to the bootstrap. CRC press, 1994. 64

[29] R. J. Errington, S. C. Chappell, I. A Khan, N. Marquez, M. Wiltshire, V. D Griesdoorn, and P. J. Smith. Time-lapse microscopy approaches to

track cell cycle and lineage progression at the single-cell level. Current protocols in cytometry, pages 12–4, 2013. 10

[30] W. Feller. An introduction to probability theory and its applications. Vol. I. John Wiley & Sons Inc., 1968. 28, 110

[31] S. Foss, D. Korshunov, and S. Zachary. An Introduction to Heavy-Tailed and Subexponential Distributions. Springer Science & Business Media, 2013. 81, 109

[32] A. Foudi, K. Hochedlinger, D. Van Buren, J. W Schindler, R. Jaenisch, V. Carey, and H. Hock. Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. Nat. Biotechnol., 27(1):84–90, 2009. 11

[33] S. A. Frank, Y. Iwasa, and M. A. Nowak. Patterns of cell division and the risk of cancer. Genetics, 163(4):1527–1532, 2003. 5

[34] A. A. Freitas and B. B. Rocha. Lymphocyte lifespans: homeostasis, selection and competition. Immunology today, 14(1):25–29, 1993. 106

[35] C. A. Giurumescu, S. Kang, T. A. Planchon, E. Betzig, J. Bloomekatz, D. Yelon, P. Cosman, and A. D. Chisholm. Quantitative semi-automated analysis of morphogenesis with single-cell resolution in complex embryos. Development, 139(22):4271–4279, 2012. 10

[36] F. L. Gomes, G. Zhang, F. Carbonell, J. A. Correa, W. A. Harris, B. D. Simons, and M. Cayouette. Reconstruction of rat retinal progenitor cell lineages in vitro reveals a surprising degree of stochasticity in cell fate decisions. Development, 138(2):227–235, 2011. 10

[37] A. Görgens, S. Radtke, M. Möllmann, M. Cross, J. Dürig, P. A. Horn, and B. Giebel. Revision of the human hematopoietic tree: granulocyte subtypes derive from distinct hematopoietic lineages. Cell reports, 3(5):1539–1552, 2013. 3

[38] G. Grimmett and D. Welsh. Probability: an introduction. Oxford University Press, 2014. 43

[39] P. Haccou, P. Haccou, P. Jagers, and V. A. Vatutin. Branching processes: variation, growth, and extinction of populations. Number 5. Cambridge University Press, 2005. 16, 18

[40] H. Haeno, Y. Iwasa, and F. Michor. The evolution of two mutations during clonal expansion. Genetics, 177(4):2209–2221, 2007. 105

[41] C. B. Harley, A. B. Futcher, and C. W. Greider. Telomeres shorten during ageing of human fibroblasts. Nat. Genet., 345(6274):458–460, 1990. 13

[42] T. E. Harris. The theory of branching processes. Springer-Verlag, 1963. 14, 20, 24, 26, 27, 28, 29, 36, 37, 39, 41, 42, 43, 45, 49, 53, 62, 67, 90, 91, 93, 97, 98, 106, 107

[43] E. D. Hawkins, M. Hommel, M. L. Turner, F. L. Battye, J. F Markham, and P. D. Hodgkin. Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. Nat. Protoc., 2(9):2057–2067, 2007. 11

[44] E. D. Hawkins, J. F. Markham, L. P. McGuinness, and P. D. Hodgkin. A single-cell pedigree analysis of alternative stochastic lymphocyte fates. Proc. Natl. Acad. Sci. U.S.A, 106(32):13457–13462, 2009. 10, 23, 29, 57, 71, 101, 103, 111

[45] L. Hayflick. The limited in vitro lifetime of human diploid cell strains. Experimental cell research, 37(3):614–636, 1965. 12

[46] M. Hills, K. Lücke, E. A. Chavez, C. J. Eaves, and P. M. Lansdorp. Probing the mitotic history and developmental stage of hematopoietic cells using single telomere length analysis (STELA). Blood, 113(23):5765–5775, 2009. 7, 13

[47] P. D. Hodgkin, J-H Lee, and A. B. Lyons. B cell differentiation and isotype switching is related to division cycle number. J. Exp. Med., 184(1):277–281, 1996. 7, 50

[48] W. K. Hong, R. C. Jr. Bast, W. N. Hait, D. W. Kufe, R. E. Pollock, R. R. Weichselbaum, J. F. Holland, and E. Frei III. Holland-Frei Cancer Medicine. PMPH-USA, 8th edition, 2010. 49

[49] M. B. Horton, G. Prevedello, J. M. Marchingo, J. H. S. Zhou, K. R. Duffy, S. Heinzel, and P. D. Hodgkin. Multiplexed division tracking dyes for proliferation-based clonal lineage tracing. J. Immunol., page ji1800481, 2018. 11

[50] K. E. Huffman, S. D. Levene, V. M. Tesmer, J. W. Shay, and W. E. Wright. Telomere shortening is proportional to the size of the g-rich telomeric 3'-overhang. Journal of Biological Chemistry, 275(26):19719–19722, 2000. 12

[51] Y. Iwasa, M. A. Nowak, and F. Michor. Evolution of resistance during clonal expansion. Genetics, 172(4):2557–2566, 2006. 105

[52] H. Iwasaki, C. Somoza, H. Shigematsu, E. A. Duprez, J. Iwasaki-Arai, S. Mizuno, Y. Arinobu, K. Geary, P. Zhang, T. Dayaram, et al. Distinctive and indispensable roles of pu. 1 in maintenance of hematopoietic stem cells and their differentiation. Blood, 106(5):1590–1600, 2005. 2

[53] P. Jagers. The proportions of individuals of different kinds in two-type populations. a branching process problem arising in biology. Journal of Applied Probability, 6(2):249–260, 1969. 22, 50, 52, 56

[54] P. Jagers. Renewal theory and the almost sure convergence of branching processes. Arkiv för Matematik, 7(6):495–504, 1969. 28, 45, 49, 55, 63

[55] P. Jagers. Branching Processes with Biological Applications. John Wiley and Sons, 1975. 14, 18

[56] P. Jagers. General branching processes as markov fields. Stoch. Process. Their Appl., 32(2):183–212, 1989. 18

[57] P. Jagers. Stabilities and instabilities in population dynamics. J. Appl. Probab., 29(4):770–780, 1992. 18

[58] P. Jagers. Some notes on the history of branching processes, from my perspective. http://www.math.chalmers.se/~jagers/Branching%20History.pdf, 2009. Lecture at the Oberwolfach Symposium on "Random Trees". 14

[59] P. Jagers, F. C. Klebaner, and S. Sagitov. On the path to extinction. Proc. Natl. Acad. Sci. U.S.A., 104(15):6107–6111, 2007. 110

[60] Y. N. Jan and L. Y. Jan. Asymmetric cell division. Nature, 392(6678):775, 1998. 106

[61] M. K. Jenkins, H. H. Chu, J. B. McLachlan, and J. J. Moon. On the composition of the preimmune repertoire of t cells specific for peptide–major histocompatibility complex ligands. Annual review of immunology, 28:275–294, 2009. 4

[62] S. M. Kaech, E. J. Wherry, and R. Ahmed. Vaccines: Effector and memory t-cell differentiation: Implications for vaccine development. Nature Reviews Immunology, 2(4):nri778, 2002. 6, 60

[63] M. Kimmel and D. E. Axelrod. Branching Processes in Biology. Springer, 2002. 14, 20, 24, 27

[64] I. Kinjyo, J. Qin, S. Tan, C. J. Wellard, P. Mrass, W. Ritchie, A. Doi, L. L. Cavanagh, M. Tomura, A. Sakaue-Sawano, et al. Real-time tracking of cell cycle progression during cd8+ effector and memory t-cell differentiation. Nature communications, 6:6301, 2015. 4

[65] A. Klenke. Probability theory: a comprehensive course. Springer Science & Business Media, 2013. 52, 62

[66] M. Kondo, I. L. Weissman, and K. Akashi. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell, 91(5):661–672, 1997. 49

[67] S. G. Krantz and H. R. Parks. The implicit function theorem. Birkhäuser Boston, 2002. 109

[68] H. Landecker. Seeing things: from microcinematography to live cell imaging. Nature methods, 6(10):707, 2009. 10

[69] N. Levinson et al. Limiting theorems for age-dependent branching processes. Illinois Journal of Mathematics, 4(1):100–118, 1960. 42

[70] N. Liu, M. Ren, J. Song, Y. Ríos, K. Wawrowsky, A. Ben-Shlomo, S. Lin, and S. Melmed. In vivo time-lapse imaging delineates the zebrafish pituitary proopiomelanocortin lineage boundary regulated by fgf3 signal. Developmental biology, 319(2):192–200, 2008. 10

[71] A. B. Lyons. Analysing cell division in vivo and in vitro using flow cytometric measurement of CFSE dye dilution. J. Immunol. Methods, 243(1):147–154, 2000. 11

[72] A. B. Lyons, S. J Blake, and K. Doherty. Flow cytometric analysis of cell division by dilution of cfse and related dyes. 04 2013. 11

[73] A. B. Lyons and C. R. Parish. Determination of lymphocyte division by flow cytometry. J. Immunol. Methods, 171(1):131–137, 1994. 11

[74] J. M. Marchingo, A. Kan, R. M. Sutherland, K. R. Duffy, C. J. Wellard, G. T. Belz, A. M. Lew, M. R. Dowling, S. Heinzel, and P. D. Hodgkin. Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion. Science, 346(6213):1123–1127, 2014. 7, 75

[75] J. M. Marchingo, G. Prevedello, A. J. Kan, S. Heinzel, P. D. Hodgkin, and K. R. Duffy. T cell stimuli independently sum to regulate an inherited clonal division fate. Nat. Commun., 7:13540, 2016. 11

[76] J. Mendelsohn, P. M. Howley, M. A. Israel, J. W. Gray, and C. B. Thompson. The Molecular Basis of Cancer. Saunders, 2015. 4, 49

[77] L. M. F. Merlo, J. W. Pepper, B. J. Reid, and C. C. Maley. Cancer as an evolutionary and ecological process. Nat. Rev. Cancer, 6(12):924–935, 2006. 5

[78] C. J. Mode. Multitype branching processes. Theory and applications. Elsevier, New York, 1971. 14

[79] S. J. Morrison and J. Kimble. Asymmetric and symmetric stem-cell divisions in development and cancer. Nature, 441(7097):1068, 2006. 106

[80] A. Mukherjea, M. Rao, and S. Suen. A note on moment generating functions. Statistics & probability letters, 76(11):1185–1189, 2006. 42

[81] K. Murphy and C. Weaver. Janeway's immunobiology. Garland Science, 9th edition, 2016. 12

[82] C. M. O'Connor, J. U. Adams, and J. Fairman. Essentials of cell biology. Cambridge, MA: NPG Education, 1, 2010. 54

[83] A. M. Olovnikov. A theory of marginotomy: the incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. Journal of theoretical biology, 41(1):181–190, 1973. 12

[84] J. A Owen, Je. Punt, S. A. Stranford, et al. Kuby immunology. WH Freeman New York, 2013. 4, 5, 6

[85] P. M. Pardalos and T. M. Rassias. Contributions in Mathematics and Engineering: In Honor of Constantin Carathéodory. Springer, 2016. 84, 86

[86] S. Pauklin and L. Vallier. The cell-cycle state of stem cells determines cell fate propensity. Cell, 155(1):135–147, 2013. 50

[87] L. Perié and K. R. Duffy. Retracing the in vivo haematopoietic tree using single-cell methods. FEBS letters, 590(22):4068–4083, 2016. 2

[88] L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, and T. N. Schumacher. The branching point in erythro-myeloid differentiation. Cell, 163(7):1655–1662, 2015. 2

[89] E. O. Powell. Some features of the generation times of individual bacteria. Biometrika, 42:16–44, 1955. 10

[90] B. J. C. Quah and C. R. Parish. New and improved methods for measuring lymphocyte proliferation in vitro and in vivo using CFSE-like fluorescent dyes. J. Immunol. Methods, 379(1):1–14, 2012. 11

[91] Y. Reizel, N. Chapal-Ilani, R. Adar, S. Itzkovitz, J. Elbaz, Y. E. Maruvka, E. Segev, L. I. Shlush, N. Dekel, and E. Shapiro. Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. PLoS genetics, 7(7):e1002192, 2011. 13

[92] G. Rempala and J. Wesolowski. Symmetric functionals on random matrices and random matchings problems, volume 147. Springer Science & Business Media, 2007. 82

[93] S. I. Resnick. Adventures in Stochastic Processes. Birkhäuser Boston, 1992. 14, 21, 26, 34, 43, 62, 69, 106, 110

[94] T. Reya, S. J. Morrison, M. F. Clarke, and I. L. Weissman. Stem cells, cancer, and cancer stem cells. Nature, 414(6859):105, 2001. 3

[95] J. L. Richards, A. L. Zacharias, T. Walton, J. T. Burdick, and J. I. Murray. A quantitative model of normal caenorhabditis elegans embryogenesis and its disruption after stress. Dev. Biol., 374(1):12–23, 2013. 10

[96] A. W. Roberts, M. S. Davids, J. M. Pagel, B. S. Kahl, S. D. Puvvada, J. F. Gerecitano, T. J. Kipps, M. A. Anderson, J. R. Brown, L. Gressick, et al. Targeting bcl2 with venetoclax in relapsed chronic lymphocytic leukemia. New England Journal of Medicine, 374(4):311–322, 2016. 5

[97] W. Rudin. Principles of mathematical analysis. McGraw-Hill Book Co., New York, third edition, 1976. International Series in Pure and Applied Mathematics. 28, 43, 44

[98] W. Rudin. Real and complex analysis. McGraw-Hill International Edition, 3rd edition, 1987. 102

[99] N. Rufer, T. H. Brümmendorf, S. Kolvraa, C. Bischoff, K. Christensen, L. Wadsworth, M. Schulzer, and P. M. Lansdorp. Telomere fluorescence measurements in granulocytes and T lymphocyte subsets point to a high turnover of hematopoietic stem cells and memory T cells in early childhood. J. Exp. Med., 190(2):157–168, 1999. 13

[100] A. Sakaue-Sawano, H. Kurokawa, T. Morimura, A. Hanyu, H. Hama, H. Osawa, S. Kashiwagi, K. Fukami, T. Miyata, H. Miyoshi, T. Imamura, M. Ogawa, H. Masai, and A. Miyawaki. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. Cell, 132(3):487–498, 2008. 19

[101] M. L. Samuels. Distribution of the branching-process population among generations. Journal of Applied Probability, 8(04):655–667, 1971. 21, 25, 63

[102] T. E. Schlub, V. P. Badovinac, J. T. Sabel, J. T. Harty, and M. P. Davenport. Predicting cd62l expression during the cd8+ t-cell response in vivo. Immunology & Cell Biology, 88(2):157–164, 2010. 6

[103] R. J. Serfling. Approximation Theorems of Mathematical Statistics. Wiley, New York, 1980. 47, 82, 110

[104] D. Shibata, W. Navidi, R. Salovaara, Z.-H. Li, and L. A. Aaltonen. Somatic microsatellite mutations as molecular tumor clocks. Nat. Med., 2(6):676–681, 1996. 13

[105] D. Shibata and S. Tavaré. Counting divisions in a human somatic cell tree. Cell Cycle, 5(6):610–614, 2006. 13

[106] J. A. Smith and L. Martin. Do cells cycle? Proc. Natl. Acad. Sci. U.S.A, 70(4):1263–1267, 1973. 10

[107] J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson. The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol., 100(1):64–119, 1983. 10

[108] S. G. Tangye, D. T. Avery, and P. D. Hodgkin. A division-linked mechanism for the rapid generation of ig-secreting cells from human memory b cells. The Journal of Immunology, 170(1):261–269, 2003. 7

[109] T. Tao. The weak and strong law of large numbers. https://terrytao.wordpress.com/2015/10/23/275a-notes-3-the-weak-and-strong-law-of-large-numbers, 2015. Course in Probability Theory (275A), Notes 3. 82

[110] C. Tomasetti and B. Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science, 347(6217):78–81, 2015. 5

[111] J.-L. Tsao, Y. Yatabe, R. Salovaara, H. J. Järvinen, J.-P. Mecklin, L. A. Aaltonen, S. Tavaré, and D. Shibata. Genetic reconstruction of individual colorectal tumor histories. Proc. Natl. Acad. Sci. U.S.A, 97(3):1236–1241, 2000. 13

[112] M. Turner, E. Hawkins, and P.D. Hodgkin. Quantitative regulation of B cell division destiny by signal strength. J. Immunol., 181(1):374–382, 2008. 7

[113] H. Vaziri, W. Dragowska, R. C. Allsopp, T. E. Thomas, C. B. Harley, and P. M. Lansdorp. Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age. Proc. Natl. Acad. Sci. U.S.A, 91(21):9857–9860, 1994. 13

[114] A. Wasserstrom, D. Frumkin, R. Adar, S. Itzkovitz, T. Stern, S. Kaplan, G. Shefer, I. Shur, L. Zangi, Y. Reizel, A. Harmelin, Y. Dor, N. Dekel, Y. Reisner, D. Benayahu, E. Tzahor, E. Segal, and E. Y. Shapiro. Estimating cell depth from somatic mutations. PLoS Comput. Biol., 4(5), 2008. 13

[115] H. W. Watson and F. Galton. On the probability of the extinction of families. The Journal of the Anthropological Institute of Great Britain and Ireland, 4:138–144, 1875. 14, 27

[116] T. S. Weber, L. Perié, and K. R Duffy. Inferring average generation via division-linked labeling. Journal of mathematical biology, 73(2):491–523, 2016. iii, 9, 18, 19, 20, 21, 22, 24, 25, 27, 28, 31, 35, 37, 39, 47, 48, 49, 51, 52, 54, 60, 61, 63, 73, 74, 76, 91, 102, 105, 109

[117] S. L. Weinrich, R. Pruzan, L. Ma, M. Ouellette, V. M. Tesmer, S. E. Holt, A. G. Bodnar, S. Lichtsteiner, N. W. Kim, J. B. Trager, R. D. Taylor, R. Carlos, W. H. Andrews, W. E. Wright, J. W. Shay, C. B. Harley, and G. B. Morin. Reconstitution of human telomerase with the template RNA component hTR and the catalytic protein subunit hTRT. Nat. Genet., 17(4):498–502, 1997. 13

[118] B. Zhang, M. Dai, Q.-J. Li, and Y. Zhuang. Tracking proliferative history in lymphocyte development with cre-mediated sister chromatid recombination. PLoS Genet., 9(10):e1003887, 10 2013. 7