Optimization and Control of a Multitype Branching Process

Thesis presented for the purpose of obtaining a Diploma

Option: **Quantitative Life Sciences**

Presented by:

Djou Meli Yannela

Diploma student,

ICTP

Supervised by:

Antonio Celani

Head Senior Research Scientist,

International Centre for Theoretical Physics, ICTP

Academic year **2024-2025**

Dedication

♣ **To my parents**,

♣ **To my brothers and sisters**.

I dedicate this document to you because you have been there for me, and I know that you will always be there for me.

Acknowledgements

First of all, I thank **Almighty God** for the gift of life, health, and the many graces I have received over the years.

I gratefully acknowledge the support provided by the International Centre for Theoretical Physics (ICTP) and the teaching staff, which made this research possible. Their commitment to advancing scientific knowledge and fostering innovative research has been fundamental in completing this thesis.

I am grateful and honored to have reached this milestone under your guidance. I hereby repeat my sincere thanks to each one of you:

♠ I would also like to thank the supervisor of this Thesis, **Antonio Celani**, Head Senior Research Scientist, International Centre for Theoretical Physics, for the time he devoted to me, his help, his enlightened supervision throughout this thesis, and his patience.

♠ I would like to thank, **Matteo Marsili**, coordinator of the Quantitative Life Sciences Department, for his relentless and combined efforts to provide students with unique capacity-building skills.

♠ I cannot finish without thanking **my classmates** for their support, encouragement, and helping me to increase my communication skills during this thesis.

♠ To all those who have contributed in any way to the writing of this dissertation, I offer my sincere thanks.

Contents

List of Figures

List of Tables

List of Abbreviations

DP: Dynamic Programming

MDP: Markov Decision Process

RL: Reinforcement Learning

GWBP: Galton-Watson Branching Process

iid: Independant and Identically Distributed

Abstract

Cancer is a global health challenge in the world. It is difficult to treat due to the fact that cancer cells don't reproduce as normal cells. So, finding a good therapy for killing cancer cells remains an important challenge. In this study, we model cancer cell population dynamics using the Galton-Watson branching process (GWBP) and formulate the problem as a Markov decision process (MDP). The MDP state is defined by the distribution of cell populations across generations, and the available actions correspond to administering or withholding the drug. We apply dynamic programming (DP) with a finite-horizon Bellman equation to determine the optimal drug administration policy that balances tumor reduction and drug toxicity. Our results indicate that the optimal strategy is strongly influenced by drug dosage. At low dosages, continuous administration is optimal due to negligible side effects. However, as dosage and therefore toxicity increase, the optimal policy becomes selective, recommending administration only when necessary to achieve therapeutic goals while minimizing harm.

Keywords: *Cancer Cell, Policy Iteration, Branching process, Markov Decision Process, Drug Toxicity.*

Chapter I

Introduction

The theory of the branching processes is an area of mathematics that describes situations in which an entity (such as a particle, cell, or population) exists at a given time and then may be replaced by one, two, or more entities of a similar or different type. If the resulting offspring are of the same type, the process is known as a single-type branching process; if the offspring are of different types, it is known as a multi-type branching process. Branching processes are studied in many areas of theoretical research and have practical applications.

The first known application of a branching process to biology was made by the mathematician Francis Galton and the statistician Henry William Watson in the 1870s [2]. Since then, branching processes have been applied in population genetics, stem cell differentiation, cancer modeling, and virus evolution [3, 4, 5], just to name a few.

Following the previous work on branching processes by mathematicians, biologists, and experimentalists, we propose to study the optimization and control of a multi-type branching process by combining the mathematical formulation of a Galton-Watson process with reinforcement learning tools, such as the Markov Decision Process (MDP). We formulate the problem of controlling the branching process as a Markov decision process (MDP). We aim to find an optimal protocol for administering a drug to control the evolution of a cell population. Drug administration serves to regulate cell growth.

The main objective of this work is to determine the moment (i.e., depending on the size of the population and on the last generation) the drug has to be administered to avoid the side effects of the accumulation of the drug across generations and to obtain a reduced number of cells in the final generation. To achieve this objective, we combine the mathematical formulation from the Galton-Watson branching process and tools from decision making to find the best policy, that is, when and where to administer the drug.

For a better understanding, this thesis is divided into two main parts, distributed as follows: in the first part, we present the background material on both the branching process and reinforcement learning. We focus more on the Galton-Watson formulation, the Markov Decision Process, and dynamic programming algorithms. The second part, the last, is devoted to the results. We present the case where the previous methods are applied to both single-type and two-type branching processes. These applications can be found in cancer treatment (to find the best strategy for administering a drug to a patient with fewer side effects) and stem cell research.

Chapter II

Background on branching processes and Markov decision process

Before delving into the question of finding the optimal strategy for administering a drug, we must first understand the methods used in this process. First, we briefly introduce the Galton-Watson process, outlining its application and purpose in this context. Second, we discuss what dynamic programming is in reinforcement learning, and the tools used to solve problems based on this approach. Finally, we combine both methods to obtain the tools necessary to achieve our objective.

2.1 Branching process

Known as a stochastic process, the branching process produces offspring according to a fixed distribution and has several characteristics, such as:

- The particles behave identically and independently of one another.

  The branching process starts with an ancestor that grows indefinitely over time. Let's select one particle from the tree at a given moment in time. Its process is statistically identical to the entire process of the ancestor. However, this process is a subprocess of the whole process, meaning the branching process can be decomposed into subprocesses that are identically distributed with each other and with the whole process.

- The offspring are born at the moment of the mother's death: this moment is called the lifetime τ (the period between birth and death of a particle).

- Suppose that the branching process starts at time $t = 0$ with the birth of a single ancestor. The number of individuals present at time t is a random non-negative integer defined as $Z(t, \omega)$, where ω is the index of a particular realization of the process at the generation. It can be written as:

$$Z(t, \omega) = \begin{cases} \sum_{i=1}^{X(\omega)} Z^{(i)}(t, \tau, \omega) & t \geq \tau \\ 1 & t < \tau, \end{cases} \tag{2.1}$$

where $Z(t, \tau, \omega)$ denotes the number of individuals at time the superscript (i) depicts the i-th independent and identically distributed copy, and $X(w)$ is the number of progeny of the ancestor.

To classify the branching processes, we can use three main properties:

- **The lifetime of a single particle**

The lifetime can be sampled from any distribution and has a significant impact on the analysis of the process. In a Galton-Watson formulation, the lifetime is the same for all particles, equaling one time unit ($\tau = 1$), representing the interval from one generation to the next. This allows the process to be considered at discrete integer times.

- **The state space**:

We can also classify a branching process by the types of offspring that one ancestor can produce. Let's call S the type space of offspring. If

$$S = \begin{cases} \{1\}, & \text{single type} \\ \{1, 2, ..., k\}, & \text{multi-type} \\ \{1, 2, ...\}, & \text{denumerable type} \\ R_+, R, [0, 1]. & \text{continuous type} \end{cases} \tag{2.2}$$

- **The criticality**

This classification is very important because it indicates how the population is expected to behave over time. It is based on the mean progeny count $m = \mathbb{E}[X]$ of a single particle, where X denotes the number of offspring produced by the particle. Let us consider Z_t, the size of the population at time t, and $\mathbb{E}[Z_{t+1} \mid Z_t]$, the expected population size at the next generation. The branching process is:

- subcritical if the expected total population shrinks $\mathbb{E}[Z_{t+1}] < \mathbb{E}[Z_t]$,
- critical if the expected total population is stable $\mathbb{E}[Z_{t+1}] = \mathbb{E}[Z_t]$,
- supercritical if the expected total population grows $\mathbb{E}[Z_{t+1}] > \mathbb{E}[Z_t]$.

This is independent of the time (it only looks at one generation), the types of branching processes, and the model details. Based on these classifications, this work will focus on discrete Galton-Watson processes with a single- and two-type branching.

The Galton-Watson branching process is known as the oldest, simplest, most fundamental, and well-known mathematical formulation of a branching process. It is a classic stochastic and discrete model of the reproduction and extinction of a cell or population from generation to generation, with lifetime $\tau = 1$ as mentioned above. The power of this model lies in its ability to predict whether a population is likely to grow, survive, or eventually die out. The model can describe the average growth rate of the population [4].

2.1.1 Single-type Branching process

A single ancestor particle lives, and at the moment of its death, it produces a random number of offspring. We consider the case where an ancestor can produce up to two (2) offspring (biological process) of the same type (Fig. 2.1). The material in this section follows the style of Athreya and Ney [6].

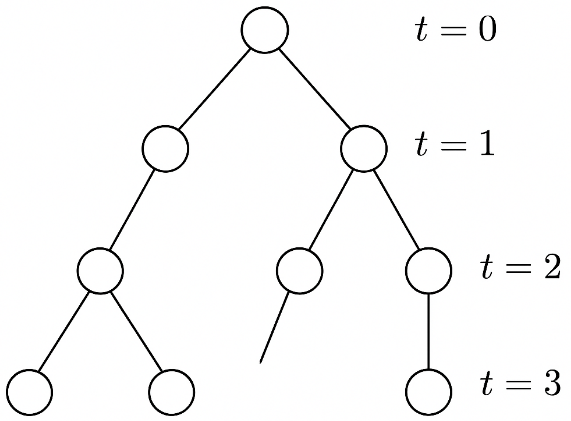


Figure 2.1: Schematic illustration of a single-type Galton-Watson branching process.

2.1.1.1 Fundamental equation

Any particle in the process can be assigned to a subprocess that is traceable (except the ancestor) to a particular generation offspring of the ancestor. Thus, the number of cells at generation $t+1$, Z_{t+1} is equal to the sum of all the cells in generation n from each subprocess that started with the first generation.

$$Z_{t+1} = \begin{cases} Z_{1,t+1}^{(1)} + ... + Z_{1,t+1}^{(Z_1)}, & Z_1 > 0 \\ 0, & Z_1 = 0 \end{cases}. \tag{2.3}$$

$Z_{1,t+1}^{(j)}$ denotes the number of particles at the generation $t+1$, started by a single ancestor born at the generation 1. (j) represents the jth iid copy. More simply, from one generation to the next, we have

$$Z_{t+1} = \sum_{j=1}^{Z_t} X_j, \tag{2.4}$$

where X_j is a stochastic variable that denotes the number of offspring of the cell j from generation t. X_j is distributed according to the following reproduction distribution:

$$X_j = \begin{cases} 0 & \text{wp} & p_0 & \text{death} \\ 1 & \text{wp} & p_1 & \text{quiescence} \\ 2 & \text{wp} & p_2 & \text{duplication} \end{cases}. \tag{2.5}$$

With the condition $\sum_k p_k = 1$. Each individual gives rise to k progeny with the same probability p_k.

Probability generating function

It is a mathematical tool that helps to model the reproduction of particles over generations. The probability generating function (PGF) of the offspring distribution, where X is a nonnegative random variable, is defined as

$$g_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} p_k s^k. \tag{2.6}$$

$s \in [0,1]$ represents a symbolic argument and $p_k = P(X = k)$ the probability that the random variable X takes the value k.

This function has many properties such as

$\star$ g_x is a non-negative function of s.

$\star$ If X is proper, $g_x(1) = 1$; otherwise $g_X(1) = P[X < \infty]$.

$\star$ The moments of the population size can be computed by differentiating PGF $\mu'_i = \mathbb{E}[X^i] = \sum_i k \cdot P(X = k)$ $\quad \mu'_1 = \mathbb{E}[X]$ and $\mu'_2 = \mathbb{E}[X^2]$.

$\star$ If X and Y are two *iid* random variables, then $g_{X+Y} = g_X(s)g_Y(s)$.

$\star$ The PGF of the population size in the $(t+1)$-generation, denoted $g_{t+1}(s)$ is obtained by the PGF of the population size in the generation t, denoted $g_t(s)$. So, $g_{t+1}(s) = g_t\big(g(s)\big)$.

The criticality classification for a single type is defined as :

$$\begin{cases} \mu > 1, & \text{supercritical} \\ \mu = 1, & \text{critical} \\ \mu < 1, & \text{subcritical} \end{cases} \tag{2.7}$$

where $\mu = \mathbb{E}[X] = g'_X(1)$ is the average offspring of a single particle.

2.1.2 Two-type Branching process

This part of the work involves branching processes with two types of particles as shown in Figure 2.2. In this case, the distribution of the offspring depends on the type of parents. We have the following hypothesis:

$\star$ There are two types of cells in the population, type-0 and type-1 cells.

$\star$ Each cell at the moment of division gives birth to a maximum of two daughters. The type of the daughter cell is not necessarily the same as the mother cell.

$\star$ During lifetime, a type-i cell can undergo many transformations, such as: reversible transformation to type-j cell, dying, dividing, or staying quiescent. The same phenomenon can be observed with a type-j cell.

2.1.2.1 Fundamental equation

As with a single-type branching process, a two-type is also traceable and divides into subprocesses. However, the process is no longer independent and identically distributed (*iid*) for all populations as previously, but rather, it is *iid* with type conditions. This means that populations of the same type reproduce identically. In other words, individuals of the same type-i, $i \in [0, 1]$, produce offspring according

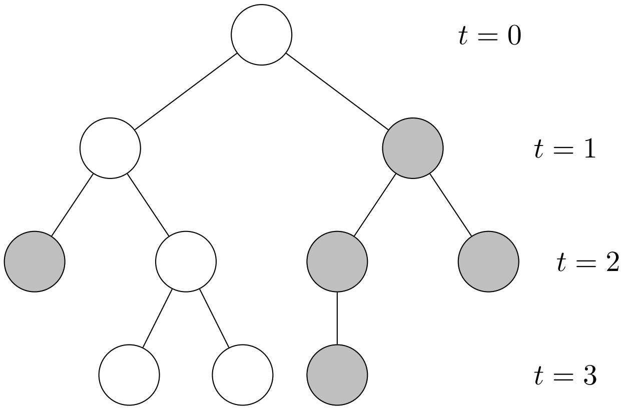


Figure 2.2: Illustration of the two-type Galton-Watson branching process.

to the same probability distribution.

Let $Z_t^{(i)}$ represent the number of type-i cells in the generation t. The total population size in generation t, is given by

$$Z_t = Z_t^{(0)} + Z_t^{(i)}.$$

The mathematical formulation of the process is stated as

$$Z_{t+1}^{(0)} = \sum_{i=1}^{Z_t^{(0)}} X_{ti}^{0\to 0} + \sum_{i=1}^{Z_t^{(1)}} X_{ti}^{1\to 0}, \tag{2.8a}$$

$$Z_{t+1}^{(1)} = \sum_{i=1}^{Z_t^{(0)}} X_{ti}^{0\to 1} + \sum_{i=1}^{Z_t^{(1)}} X_{ti}^{1\to 1}, \tag{2.8b}$$

where $Z_{t+1}^{(1)}, Z_{t+1}^{(0)}$ represent the number of cells of type-1, and type-0 cells at the generation $t+1$, respectively. $X_{ti}^{k\to j}$ depicts the offspring of type-i cell at the generation t that changes from $type - k$ in the generation t to $type - j$ at the generation $t+1$.

Depending on the type of the mother cell (type-i), the offspring $X_{ti}^{i\to j}$ follows the distribution r_{ij}^{β} where β represents the cell type of the mother and (i, j) the type of the daughter cell, i stand for type-0 and j for type-1. So,

$$X_{ij}^{\beta} \ \sim \ r_{ij}^{\beta}.$$

2.1.2.2 Probability generating function

The PGF has the same two main components, depending on the type of cell. The properties presented for the single-type should be verified in this case as well, but they are not applied to the entire population; instead, they are applied to one type. The formulation of the moment also changes.

The PGF of the total offspring produced by a type-β mother is defined as

$$h_k(s_0, s_1) = \mathbb{E}\,[s_0^i s_1^j\,], \tag{2.9}$$

where s_0, s_1 are variables used to track the number of type-0, type-1 cells, and (i, j) is the number of type-0 type-1 daughters produced by the mother of type-k.

For type-0 mother, this can be written as

$$h_0(s_0, s_1) = r^0_{00} + r^0_{01}s_1 + r^0_{10}s_0 + r^0_{02}s_1^2 + r^0_{20}s_1^2 + r^0_{11}s_0s_1.$$

Knowing the PGF of the total offspring of each cell type, we can compute the moments of each individual.

⋆ **The first moment**:

It gives the expected number of particles of each type at each generation. It is defined as

$$M = \begin{bmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{bmatrix} \in \mathbb{R}^{2\times 2}, \tag{2.10}$$

where:

- m_{ij} represents the expected number of type-j offspring produced by a single particle of type-i,

- i represents the mother type and j the daughter type.

For the mother of type-0, we have: $m_{00} = \sum_{k,l} k\ r^0_{kl}$ and $m_{01} = \sum_{k,l} l\ \ r^0_{kl}$ with r^0_{kl} the probability of type-0 with k offspring of type-0 and l offspring of type-1.

The criticality classification gives us the following:

Instead of having μ, we have λ, which is the largest eigenvalue of the mean matrix. So, if

$$\begin{cases} \lambda\ >\ 1, & \text{supercritical} \\ \lambda, =\ 1, & \text{critical} \\ \lambda\ <\ 1. & \text{subcritical} \end{cases} \tag{2.11}$$

2.2 Markov Decision process

A Markov decision process (MDP) is a fundamental tool for modeling decision-making problems in a stochastic environment. It is mostly used in control problems and fields such as reinforcement learning, robotics. In this document, the MDP is applied to the field of Reinforcement Learning, and the references used are books of Puterman and Martin [7], Sutton et *al* [8], Bertsekas and Dimitri [9].

This concept evolves over discrete time steps $t = 0, 1, 2, ...$ and at each time, it is characterized by a set of 5 elements: the state (s_t), the action (a_t), the transition probability $\big(P(s_{t+1}|s_t, a_t)\big)$, the reward $\big(\mathrm{R}(s_t, a_t,$ $s_{t+1}\ )\big)$ and the discount factor γ.

- ⋆ **State** (s_t):

  This represents all the configurations in which the environment can be in. We call $\mathbb{S}$ the set of states with $s_t \in \mathbb{S}$. We can have a finite or an infinite set of states.

- ⋆ **Action** (a_t):

  This represents all the possible actions that can be taken by the agent. $\mathbb{A}$ is called set of actions, so $a_t \in \mathbb{A}$.

- ⋆ **Transition probability** $\big(P(s_{t+1}|s_t, a_t)\big)$:

  This represents the probability that the environment moves from the state s_t to the state s_{t+1} if the agent takes the action a_t

- ⋆ **Reward** $\big(\mathrm{R}(s_t, a_t,\ s_{t+1}\ )\big)$ This function depicts the immediate reward that the environment receives when it reaches the state s_{t+1} by taking the action a_t at the state s_t.

- ⋆ **Discount factor** $(\gamma) \in [0, 1]$:

  This measures how much future rewards are worth compared to the immediate ones.

The steps of an MDP are as follows: First, the environment/system is in a state s_t at time t. Then an action is chosen from the set of actions $\mathbb{A}$. Using the transition probability, the system moves to the state s_{t+1} and receives an immediate reward $R(s_t, a_t, s_{t+1})$. The next state becomes the current state, and the process restarts (see Figure 2.3).

Note: The chosen action and the new state depend only on the current state.

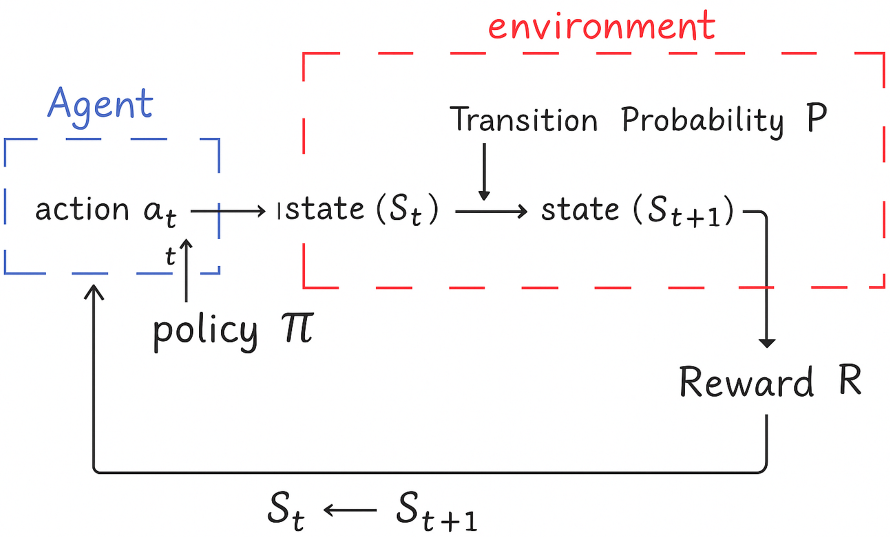


Figure 2.3: Illustration of a Markov Decision Process steps.

The objective of the MDP

The goal of an MDP is to find the optimal policy $\pi(a_t|s_t)$ (probability of taking an action a_t in a given state s_t) that maximizes (or minimizes) the cumulative expected reward (or cost) over time, which is called the objective function. The objective function differs depending on the type of process [7]. We have

- ⋆ Finite-horizon expected reward

$$G^\pi(s) = \mathbb{E}_\pi\big[\sum_{t=0}^{T} r(s_t, a_t)\big]\,, \tag{2.12}$$

- ⋆ Infinite-horizon expected reward

$$G^\pi(s) = \mathbb{E}_\pi\big[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t)\big]\,. \tag{2.13}$$

The optimal policy for the infinite-horizon (the general case) is written as

$$\begin{aligned}\pi^* &= \arg\max_\pi G^\pi(s).\\ &= \arg\max_\pi \mathbb{E}_\pi\big[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\big].\end{aligned} \tag{2.14}$$

When the initial state is set, we obtain what is called the state-value function, denoted $V^\pi(s)$, written as:

$$V^\pi(s) = \mathbb{E}_\pi\big[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\,|\, s_0 = s\big]. \tag{2.15}$$

The expanded form of the state value function $V^\pi(s)$ is given by

$$V^\pi(s_t) = \sum_a \pi(a_t|s_t) \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t) \left[r(s_t, a_t) + \gamma V^\pi(s_{t+1})\right] \tag{2.16}$$

The optimal value function satisfied the Bellman optimality equation, given as

$$V^*(s_t) = \max_{a \in A} \left[r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t) V^*(s_{t+1})\right] \tag{2.17}$$

Eq. (2.17) expresses the value of a given state under the optimal policy. The optimal policy is then given by:

$$\pi^*(s_t) = \arg\max_{a \in A} \left[r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t) V^*(s_{t+1})\right] \tag{2.18}$$

To find the optimal policy, we can use several types of algorithms, including linear programming, reinforcement learning, and dynamic programming, just to name a few. In this work, we use dynamic programming.

2.2.1 Dynamic programming (DP)

Mathematically speaking, dynamic programming is a method used to solve complex problems by breaking them into simpler subproblems that can be solved once and the solutions combined according to a recurrence of optimal substructure. From a reinforcement learning perspective, DP is a classical and powerful method to solve complex decision-making problems, in which the agent learns to make optimal choices. DP is used when the full model of the environment is known, including transition probabilities and rewards [10].

The algorithms used in DP to solve MDP are value iteration and policy iteration.

⋆ The value iteration focuses on directly finding the optimal value function. It iteratively updates the value function $V(s_t)$ using the Bellman equation Eq. (2.17) in each state and then derives the optimal policy from the final value function using Eq. (2.18).

⋆ The policy iteration alternates between two steps: policy evaluation using Eq. (2.16) and policy improvement, using Eq. (2.18) with the value function obtained in the previous step.

Chapter III

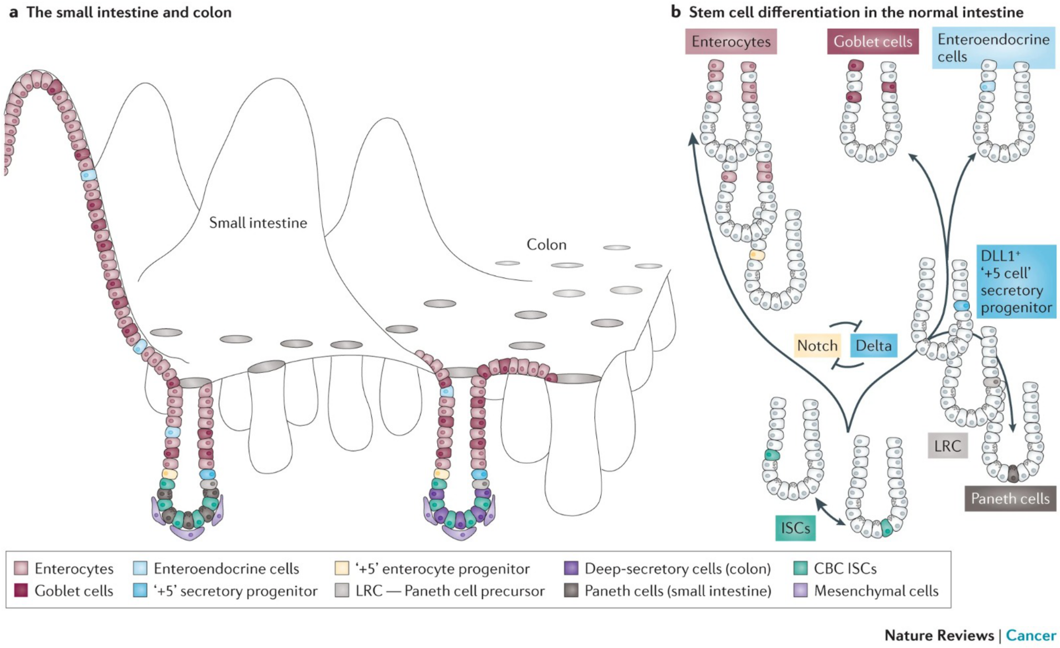
Results and Discussion

In the previous chapter, we briefly explained the branching process, the Markov decision process, and the tools to solve related problems. In this chapter, we will first present the problem of drug administration. Secondly, based on that problem, we will map the branching process into a Markov decision process to find the optimal policy for drug administration.

3.1 Problem description

Cancer is one of the world's major diseases. In general, it is not a single disease, but rather a group of diseases. This is why there is so much interest in cancer treatment and drug optimization strategies, particularly in the field of biology. Cancer is characterized by the uncontrolled growth and abnormal division of cells. Cancer cells have many properties, including uncontrolled proliferation (they divide faster and do not die as easily as normal cells), invasion, and metastasis (they spread into tissues and travel to other organs). If left untreated, cancer can cause serious illness or death. Chemotherapy is one of the most common cancer treatment methods used nowadays [11], but when the therapy is not used effectively (i.e. when it does not distinguish between cell types, or it is constantly applied, ...), the patient has what are so-called side effects. Therefore, developing an optimal chemotherapy strategy that minimizes both the side effects of drug administration and tumor size remains a challenge for oncologists [12].

A stem cell is a type of cell that does not perform a specific function. It can self-renew, producing identical copies of itself, and differentiate, transforming into a specialized type of cell. For example, in the intestine, there is a type of cell known as an intestinal stem cell. These cells divide to produce transit-amplifying progenitors, which then differentiate into several mature cell types, such as enterocytes, goblet cells, endocrine cells, and Paneth cells, as shown in Fig. 3.1.



a The small intestine and colon

b Stem cell differentiation in the normal intestine

Small intestine
Colon
Enterocytes
Goblet cells
Enteroendocrine cells
DLL1⁺ '+5 cell' secretory progenitor
Notch
Delta
LRC
Paneth cells
ISCs
Enterocytes
Goblet cells
Enteroendocrine cells
'+5' secretory progenitor
'+5' enterocyte progenitor
LRC — Paneth cell precursor
Deep-secretory cells (colon)
Paneth cells (small intestine)
CBC ISCs
Mesenchymal cells

Nature Reviews | Cancer

Figure 3.1: Intestinal epithelium: stem cells and their differentiations [1].

Cancer stem cells are a subpopulation of cancer cells within a tumor that share key properties with normal stem cells, such as self-renewal and differentiation. These cells drive tumor growth and progression by continuously regenerating cancer cells. They are the root of the tumor, maintaining and expanding it, whereas differentiated cancer cells are less likely to promote long-term growth. In cancer therapy, they are difficult to target because they survive and grow using the same signals as normal cells [13].

In this thesis, as in the works of Nazila et al. in Bazrafshan [12], we find an optimal strategy for cancer therapy. This strategy is designed to minimize the side effects of drug administration and reduce the size of the cancer cells. This applies to cancer stem cells and their differentiation. Furthermore, the plasticity of cancer cells is also taken into consideration.

3.2 Mapping of a branching process into a Markov Decision process

To formulate the problem of optimal strategy for drug administration in a cancer treatment application, we model the cancer cell evolution as a branching process. These cells may die, divide, remain quiescent (not changing from one generation to the next), or mutate (depending on the type of the considered branching process). This natural process of cancer evolution is called an uncontrolled branching process.

Once the treatment is introduced into the system, the process is no longer called an uncontrolled branching process, but a controlled branching process. It is called controlled because the treatment affects the natural evolution of cancer. In our case, the drug affects the probability distribution of the offspring.

The MDP can be used to formalize this problem of decision-making, where the state represents the size of the tumor at time or generation t. In this work, the actions are to administer or not administer the drug. The transition probability and the reward will depend on the assumptions and the type of branching process chosen. The objective is to find the best strategy (optimal policy) that minimizes the side effects of drug accumulation and cancer progression. The model considered is the Galton-Watson process for the branching process.

In the following, a finite-horizon MDP model with finite state and action space is developed for a period of cancer chemotherapy treatment. To do so, it is necessary to define the five basic components of MDP suitable for cancer chemotherapy treatment.

3.2.1 Single-type branching process

In this section, the MDP characteristics used are those explained above in Sect. 2.1.1

3.2.1.1 State

Let us suppose that the tree is made of just one type of cancer cell. We denote by Z_t and $\mathbb{S}_t$ the population size and the space of states in generation t.

$\star$ In generation $t = 0$, we start with one single cell $\mathbb{S}_0 = 1$.

$\star$ In generation $t = 1$, the possible result of the number of cells is $\{0, 1, 2\}$. So we have $\mathbb{S}_1 = \{0, 1, 2\}$.

$\star$ In generation $t = 2$, we have $\mathbb{S}_2 = \{0, 1, 2, 3, 4\}$

$\star$...

$\star$ In generation t, we have $\mathbb{S}_t = \{0, 1, ..., 2^t\}$. The state is then $Z_t \in \mathbb{S}_t$

3.2.1.2 Action

We consider the case in which the actions are to administer or not to administer the drug. Thus, the set of actions is given by $\mathbb{A} = \{0, 1\}$, where 1 stands for administering the drug and 0 does not.

Each action $a \in \mathbb{A}$ modifies the distribution of the offspring as follows.

$$\begin{aligned}(p_0, p_1, p_2) \rightarrow & (p_0 + a\epsilon_1,\ p_1 + a\epsilon_1,\ p_2 + a\epsilon_2) = (q_0,\ q_1,\ q_2),\\ q_i = & p_i + a\epsilon_i, \qquad \sum_i \epsilon_i = 0 \text{ and } \sum_i q_i = 1. \end{aligned} \tag{3.1}$$

With ϵ_i, the effect of the drug on the probability of offspring p_i.

3.2.1.3 Transition probability

The transition probability is based on the probability generating function of the offspring of a single cell, given by Eq. (2.6).

The transition probability from state Z_t to state Z_{t+1} is then given by

$$P(Z_{t+1} = j | Z_t = i, a_t) = [s^j]\big(g_{a_t}(s)\big)^i.$$

Where $[s^j]\big(g_{a_t}(s)\big)^i$ represents the coefficient of s^i in the expansion of $\big(g_{a_t}(s)\big)^i$ and $g_{a_t}(s)$ the distribution of the entire offspring under the drug effects with the offspring probability q_i.

3.2.1.4 Reward

Written as $r(Z_t, a_t)$, the immediate reward function in the case of this work will be the cost function because the organism is more damaged when we are in an uncontrolled situation. So, we have $r(Z_t, a_t) = -c(Z_t, a_t)$.

For cancer treatment, we can observe two different cost functions.

⋆ The terminal cost:

It depends only on the final state $f(Z_T)$, which is an increasing function of the state, and it is given as $f(Z_T) = \log(1 + Z_t)$,

⋆ The step action cost:

This represents the cumulative effect of a drug from generation to generation. It is represented by $\alpha\, A_t$. A_t is the binary action (treat or not) and α the dosage modulation.

The choice of the log function for the terminal cost is motivated by the desire to balance the effects of drug toxicity and tumor growth. From one generation to the next, the step cost increases linearly while tumor size increases exponentially. Therefore, choosing the wrong terminal cost can cause the optimizer to focus only on one cost instead of both.

Combining both, the total cost function is

$$c(Z_T) = \log(1 + Z_t) + \alpha \sum_{t=0} A_t \tag{3.2}$$

The objective of solving the MDP is to find a policy for choosing actions that minimizes the expected costs. Let $C_t^*(s)$ be the optimal value of the expected total cost when the state in generation t is Z_t. Then, the minimal expected total cost and the corresponding optimal policy for all generation t and all states $Z \in \mathbb{S}$ can be found by solving the following Bellman equation (Eq. (2.17)) iteratively.

$$C_t^*(Z_t) = \begin{cases} C_T(Z_T, a_T) & t = T, \\ \min_{a_t \in \mathbb{A}} \left\{ c_t(Z_t, a_t) + \sum_{Z_{t+1}} P(Z_{t+1}|Z_t, a_t)\, C_{t+1}^*(Z_{t+1}) \right\} & t < T. \end{cases} \tag{3.3}$$

Where:

− $C_T(Z_T, a_T)$ is the final cost obtained at the end of the process,

− c_t is the intermediate cost from generation t to generation $t + 1$,

− T denotes the final generation.

At each generation, we choose the following action (Eq. (2.18)):

$$a_t^* \in \arg\min_{a_t \in \mathbb{A}} \left\{ c_t(Z_t, a_t) + \sum_{Z_{t+1}} P(Z_{t+1}|Z_t, a_t)\, C_{t+1}^*(Z_{t+1}) \right\},\ t = 1, 2, ..., N - 1. \tag{3.4}$$

a_t^* is the action that minimizes the expected total cost in generation t for state Z.

Note : We are working with a finite-horizon problem, so the discount factor γ is set to 1.
After computing the Bellman equation, we got many results presented in the following

Figure. 3.2 illustrates the action that minimizes the side effects of the drug and reduces the size of the cancer cell in each generation. We observe that in each generation, there is a critical population size above which the drug is administered. Below the threshold, do not administer a drug (the balls are colored yellow), and above, always administer a drug. For example, in generation $t = 4$, the threshold is 3 while in generation $t = 6$ it is 19. Also, in the Galton-Watson process, the risk of extinction is high when the population is small [14], so the treatment is not necessary in such cases. This indicates that the policy

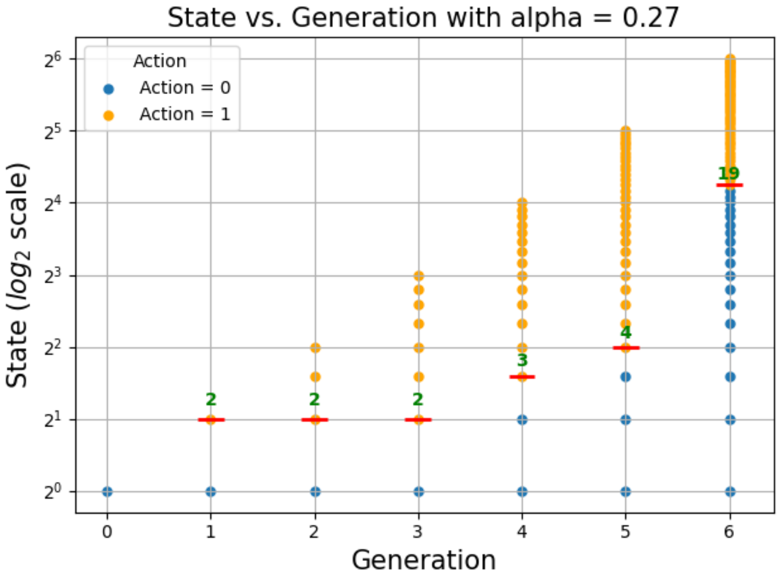


Figure 3.2: Optimal policy for a fixed dose of a drug.

shown prefers to avoid early intervention unless the tumor has already become dangerous, because early treatment may kill cancer cells, but it also has a cost. It is important to mention that the graph shown above was obtained for a fixed dose of drug $\alpha = 0.27$.

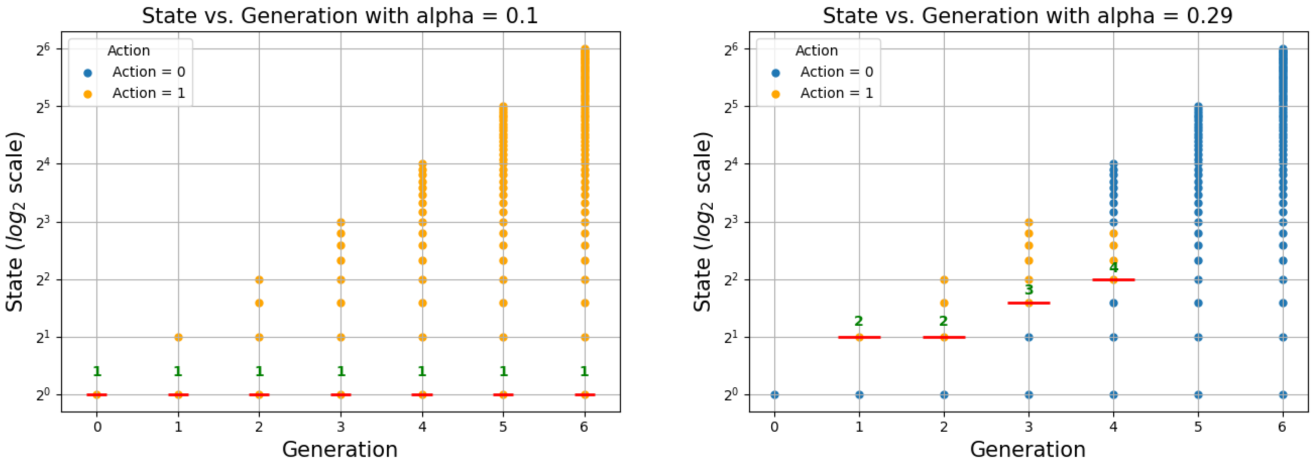


Figure 3.3: Optimal policy for different values of α

On the left-hand side of Fig. 3.3, the policy tells us to always administer the drug. This can be explained by the fact that the dose of the drug is too small to have a significant effect on drug accumulation. On the right-hand side, always administering the drug is too costly because the drug dose is too high. So, it is better to administer the drug when necessary.

Figures 3.2 and 3.3 both allow us to understand that the policy depends not only on the generation, but also on the dose of the drug. Figure 3.4 represents the evolution of the average logarithm of population size in the last generation as a function of α. This shows that, initially (for small values of α), the average population size at the final generation is kept under control because the drug is administered almost all the time. However, once α exceeds a certain (close to 0.20), the toxicity of the drug becomes significant, leading to high cost when treating at a certain population size. As a result, the optimal policy strategy is to allow the population to grow without administering the drug. This implies the case of high-dose treatment (large α) with an uncontrolled process (increasing curve).

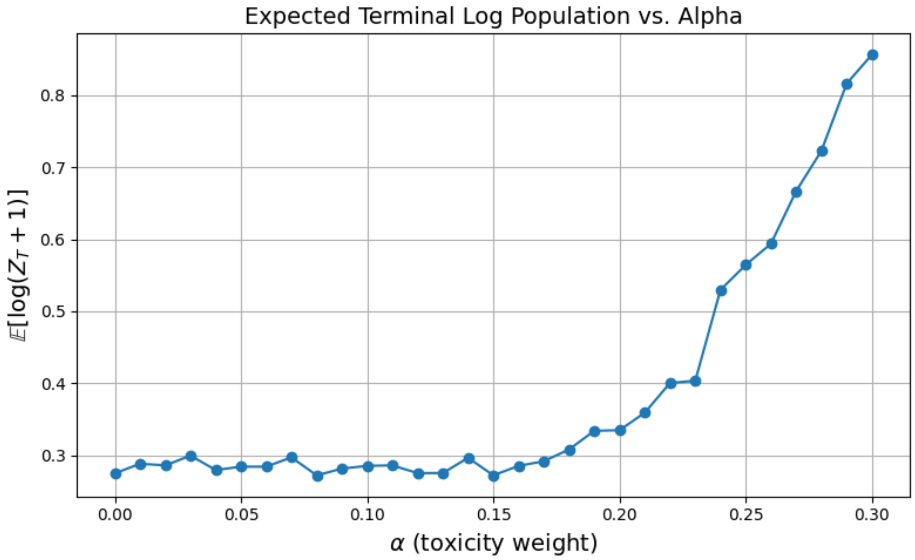


Figure 3.4: Illustration of the expected final cost as a function of α.

From the given policy, we can compute many trajectories and their average to see the evolution of the average population size over generations, as shown in Fig. 3.5. This figure indicates that when the process is uncontrolled (no drug is administered), the population grows exponentially with a large growth rate. Under the drug effect, we can have two scenarios: always administer a drug or follow the optimal policy. The figure shows that always giving a drug should be the best option, but it is not the case because it does not take into account the toxicity of the drug. So, the optimal policy appears to be better because we are sure to minimize the side effects of the drug accumulation and the minimization of the population size.

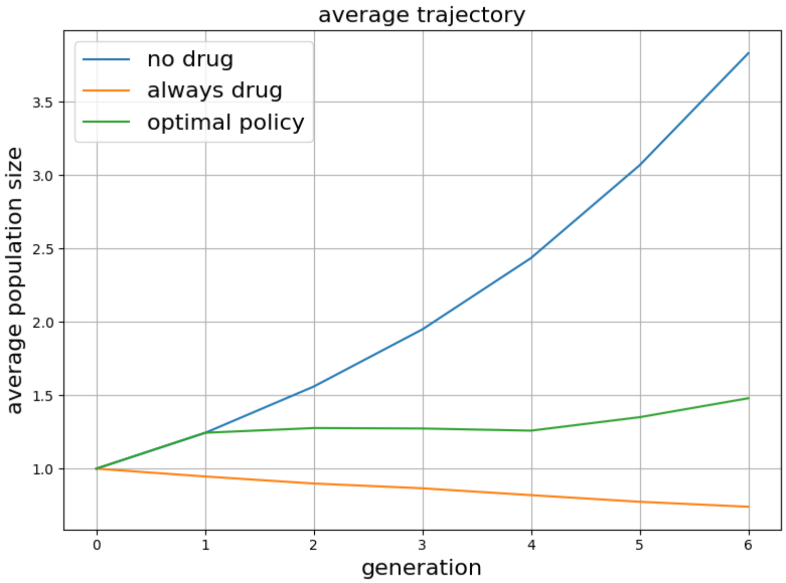


Figure 3.5: Average trajectory taken under specific policies as a function of generations, where the policies are as follows: a) curve in blue for never administering the drug, b) curve in green for the optimal policy for $\alpha = 0.27$, and c) curve in red for always administering the drug.

3.2.2 Two-type branching process

The practical case we are working on is the case of cancer treatment, where we have normal and differentiated cancer stem cells. These two types of cancer represent type-0 and type-1 cells, respectively, used in Sect. 2.1.2. The possible transformations that we can observe are given in the tables 3.1 and 3.2

Outcome Description	**Offspring Tuple (i, j)**	**Probability Symbol**
Dies	$(0, 0)$	p_{00}
Remains quiescent (survives unchanged)	$(1, 0)$	p_{10}
Transforms into type-1	$(0, 1)$	p_{01}
Divides into 2 type-0 cells	$(2, 0)$	p_{20}

Table 3.1: Possible outcomes for type-0 cell division.

Constraints: $p_{00} + p_{10} + p_{01} + p_{20} = 1.$

Constraints: $q_{00} + q_{10} + q_{01} = 1.$

Outcome Description	**Offspring Tuple (i,j)**	**Probability Symbol**
Dies	$(0,0)$	q_{00}
Remains quiescent (survives unchanged)	$(0,1)$	q_{01}
Transforms into a type-0 cell	$(1,0)$	q_{10}

Table 3.2: Possible outcomes for type-1 cell division.

3.2.2.1 State

The state is no longer a single element, but a set of two elements (Z^0, Z^1), where Z^0 represents the number of type-0 cells and Z^1 the number of type-1 cells. Following the same idea of the single type, the two-type branching process should also satisfy the following condition: In generation t,

$$Z_t^0 + Z_t^1 \leq 2^t.$$

We then have:

⋆ In the generation $t = 0$, we start with up to one single cell in total, so

$$\mathbb{S}_0 = \begin{bmatrix} (0,0) & (0,1) \\ (1,0) & \times \end{bmatrix}.$$

⋆ In the generation $t = 1$, the possible outcomes for the number of cells go up to 2, so we have,

$$\mathbb{S}_1 = \begin{bmatrix} (0,0) & (0,1) & (0,2) \\ (1,0) & (1,1) & \times \\ (2,0) & \times & \times \end{bmatrix}$$

⋆...

With $Z_t = (Z_t^0, Z_t^1) \in \mathbb{S}_t$ in each generation.

3.2.2.2 Action

The main objective of the action remains to administer the drug or not. Instead of having one scenario, we have two different scenarios:

- Scenario 1: One drug that targets both cell types in the same way (for example, increasing the probability of death).

• Scenario 2: Two drugs targeting one type of cell each (again increasing the probability of death).

Let us define the scenarios:

⋆ $a_2 \in \{0,\, 1\}$ the drug that targets both cells,

⋆$a_0 \in \{0,\, 1\}$: the drug affects type-0 cells only,

⋆ $a_1 \in \{0,\, 1\}$: the drug affects type-1 cells only.

Then the probabilities become functions of the three control variables:

$$r^{(\beta)}_{ij}(a) = r^{(\beta)}_{ij}(0) + \sum_{\mu} \epsilon^{(\beta)}_{ij}(\mu)\, a_{\mu},$$

Where

⋆ (i, j) represents the couple of offspring: i represents the type-0 cell and j the type-1 cell,

⋆ $a = \{a_0, a_1, a_2\}$ the different scenarios for the action

⋆ $r^{\beta}_{ij}\epsilon(p_{ij},\, q_{ij})$, the initial probabilities of the process by type p_{ij} is type-0 and q_{ij} for type-1. So, $r^{\beta}_{ij} = p_{ij}$,

⋆ $\beta \in \{0, 1\}$ for the cell type,

⋆ $\epsilon^{\beta}_{ij}(\mu)$ are the effects of drugs,

⋆ $\mu \in \{0, 1, 2\}$ refers to the drug scenario,

⋆ Constraints: $\sum_{i,j} \epsilon^{\beta}_{ij}.(\mu) = 0,$ and $\sum_{i,j} r^{(\beta)}_{ij}.(a_0, a_1, a_2) = 1$

Examples :

$r^{(0)}_{00}(a_0, a_1, a_2) \Rightarrow r^{(0)}_{00}(0, 0, 0) + \epsilon^{(0)}_{ij}(0)\, a_0 + \epsilon^{(0)}_{ij}(2)\, a_2,$

$r^{(1)}_{10}(a_0, a_1, a_2) \Rightarrow q^{(1)}_{10}(0, 0, 0) + \epsilon_{ij}(1)\, a_1 + \epsilon^{(1)}_{ij}(2)\, a_2.$

3.2.2.3 Transition probability

Following the idea presented in a single type, the transition probability is based on the probability generating function of the total offspring produced by both types of cells. It is written as a function of Eq. (2.9).

The transition probability from state $Z_t = (Z^0_t, Z^1_t)$ to state $Z_{t+1} = (Z^0_{t+1}, Z^1_{t+1})$ is given by the product of the pgfs for each type raised to the number of individuals of that type in the given generation t: This reflects the independence of the branching process. The expression is then given by

$$H_t(s_0, s_1) = \big[h_0(s_0, s_1)^{z_0}.h_1(s_0, s_1)^{z_1}\big].$$

So, we can write

$$\mathbb{P}\big(Z_0(t+1) = k,\, Z_1(t+1) = l \,|\, Z_1(t) = z_1,\, Z_0(t) = z_0\,\big) = \big[s_0^k s_1^l\big] H_t(s_0, s_1).$$

Where $\big[s_0^k s_1^l\big]$ represents the coefficient $s_0^k s_1^l$ of $H_t(s_0, s_1)$.

3.2.2.4 Reward

The main goal remains to administer a drug to keep the number of cells constant or reduced as much as possible in the final generation.

Following the ideas of the single type and based on the fact that we have two types of cells,

⋆ The terminal state cost is written as $f(Z_T^0, Z_T^1) = \log(1 + \sum_\beta Z_T^\beta)$,

⋆ The step action cost gives $\alpha_\mu.a_\mu(t)$ for a single drug a_μ in generation t and $\alpha_{\mu,\eta}$ for the combined drugs a_μ and a_η in generation t.

α_μ is the dose modulation for one drug and $\alpha_{\mu,\eta}$ the combined drug modulation.

Combining both, the total final cost function is then

$$C(Z_T^1, Z_T^0) = \sum_{t=1}^{T} \big[\, \sum_\mu \alpha_\mu\, a_\mu(t) + \sum_{\mu,\eta} \alpha_{\mu,\eta}\, a_\mu(t)\, a_\eta(t) \,\big] + \log\big(1 + \sum_\beta Z_T^\beta\,\big).$$

With $\alpha_{\mu,\eta} = 0$ for $\mu = \eta$

For example,

⋆ In the scenario of two drugs, we have:

$$C(Z_T^1, Z_T^0) = \sum_{t=1}^{T} \big[\, \alpha_0\, a_0(t) + \alpha_1\, a_1(t) + \alpha_{01} a_1 a_0 \,\big] + \log\big(1 + Z_T^0 + Z_T^1\,\big).$$

⋆ In the scenario of one drug, we have:

$$C(Z_T^1, Z_T^0) = \sum_{t=1}^{T} \big[\, \alpha_2\, a_2(t) \,\big] + \log\big(1 + Z_T^0 + Z_T^1\,\big).$$

In the following, we will consider two cases:

⋆ Without plasticity ($q_{10} = 0$): differentiated cancer cells cannot dedifferentiate (reversible transformation for cancer stem cells to normal cancer stem cells.)

⋆ With plasticity, we allow cancer cells to dedifferentiate.

After applying all these MDP characteristics and using Bellman Eq. 2.17, we obtain the following graphs

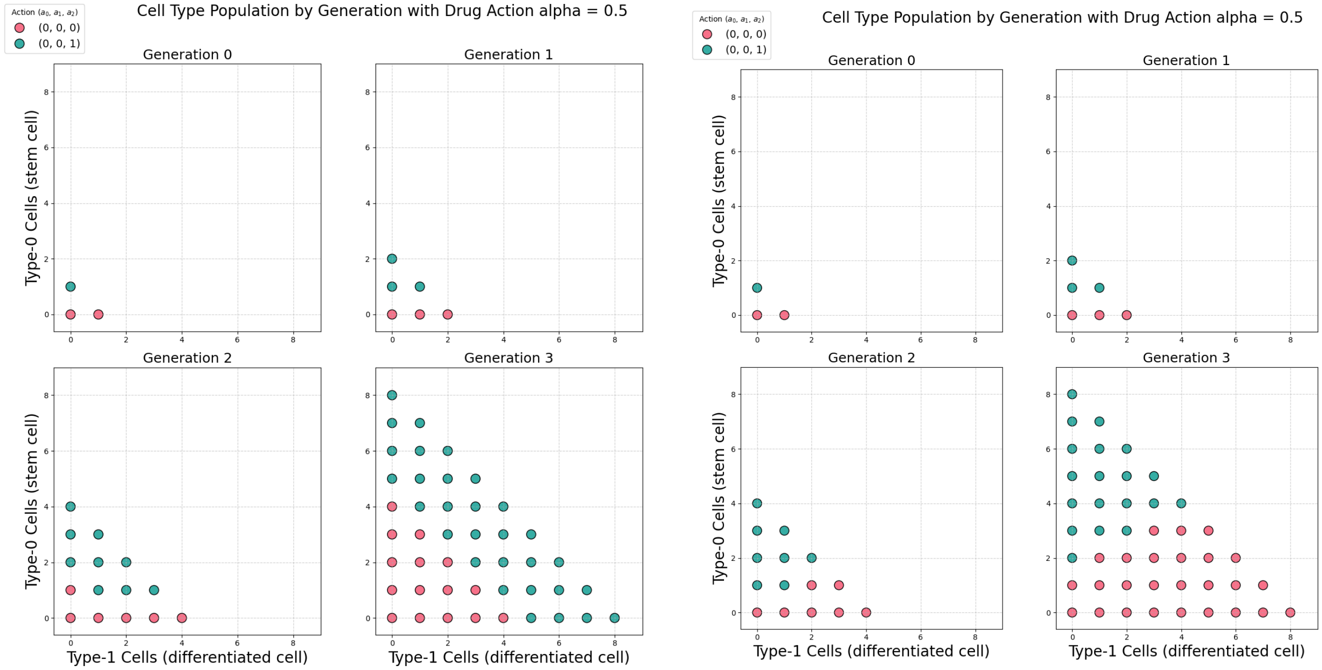


Figure 3.6: Optimal policy for a fixed dose of a drug without (left panel with four generations) and with plasticity (right panel with four generations); $\alpha = 0.5$.

3.2.2.5 First scenario: one single drug for both types of cells

The plots of Figure 3.6 illustrate the optimal action to take in each generation, depending on the number of stem cells and differentiated cells. The four panels on the left-hand side represent the optimal policy obtained in the case of one single drug that acts on two types of cells. We observe that in the first three generations (generation $t = 0, 1, 2$), the drug is almost always given when the stem cells are present in the medium. However, in generation $t = 4$, we observe that the drug is applied after a certain number of the total population size is reached ($Z_T = Z_T^1 + Z_T^0 = 4$), regardless of the cell size combination. Biologically, this suggests that rather than administering the drug continuously, it is more effective to apply it when there is a risk of expansion of stem cells.

The four panels on the right-hand side still represent the optimal policy using the same drug dosage. However, in this case, differentiated cells can become stem cells. This increases the chance of having more stem cells in the medium. The policy does not prioritize differentiated cells, since stem cells can easily proliferate, whereas differentiated cells cannot grow independently, unless they originate from stem cell differentiation.

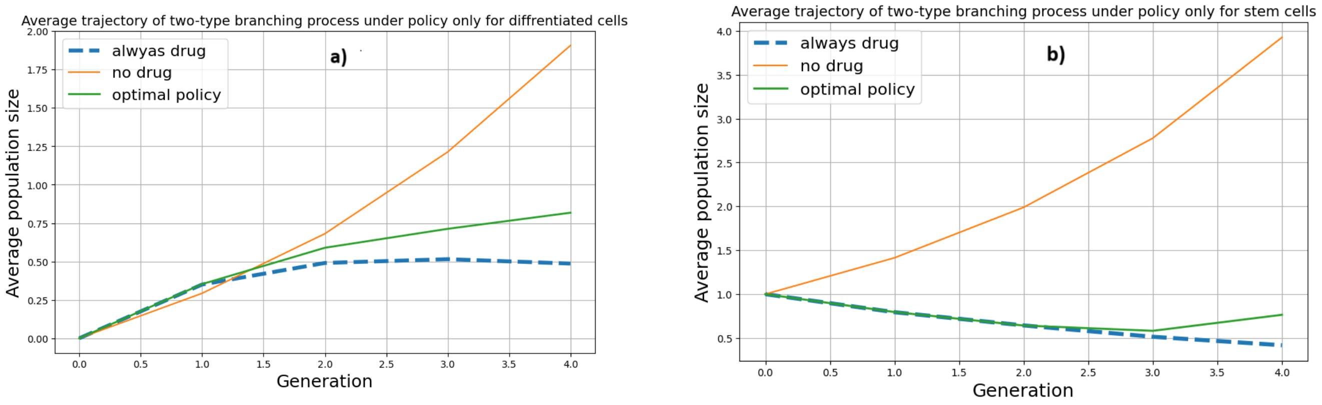


Figure 3.7: Average trajectory of cells as a function of generation under different policies with $\alpha = 0.5$, without plasticity.

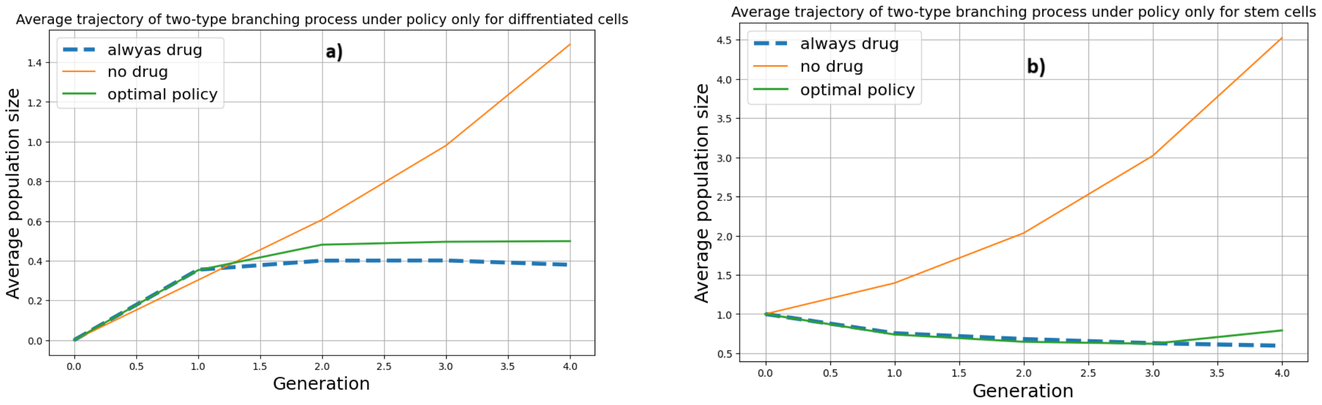


Figure 3.8: Average trajectory of cells as a function of generation under different policies with $\alpha = 0.5$, and with plasticity.

Following the idea presented in Figure 3.6, the average trajectories shown in Figure 3.7 and Figure 3.8 illustrate the average trajectory of the population size as a function of generations. Without intervention (no drug), all cells grow exponentially (orange curves), indicating that the branching process is supercritical. When the drug is always administered (dashed line), the population size shrinks significantly, and the process becomes subcritical, eventually leading to extinction after certain generations. However, this strategy does not account for the drug toxicity, which is a significant concern; instead, it solely focuses on minimizing the population size.

In contrast, the optimal policy (green curves) for its part almost behaves similarly to the always administering a drug strategy, but balances both objectives: the toxicity of a drug and the limitation of population size. For this reason, the optimal policy sometimes recommends not administering a drug. We also observe that the maximum number of differentiated cells in Figure 3.7a is slightly higher than the average maximum number in Figure 3.7b. This difference is due to the plasticity of cancer cells. When there is no plasticity, we expect a higher number of differentiated cells compared to the case with plasticity.

3.2.2.6 Second scenario: two drugs, one for each type of cell

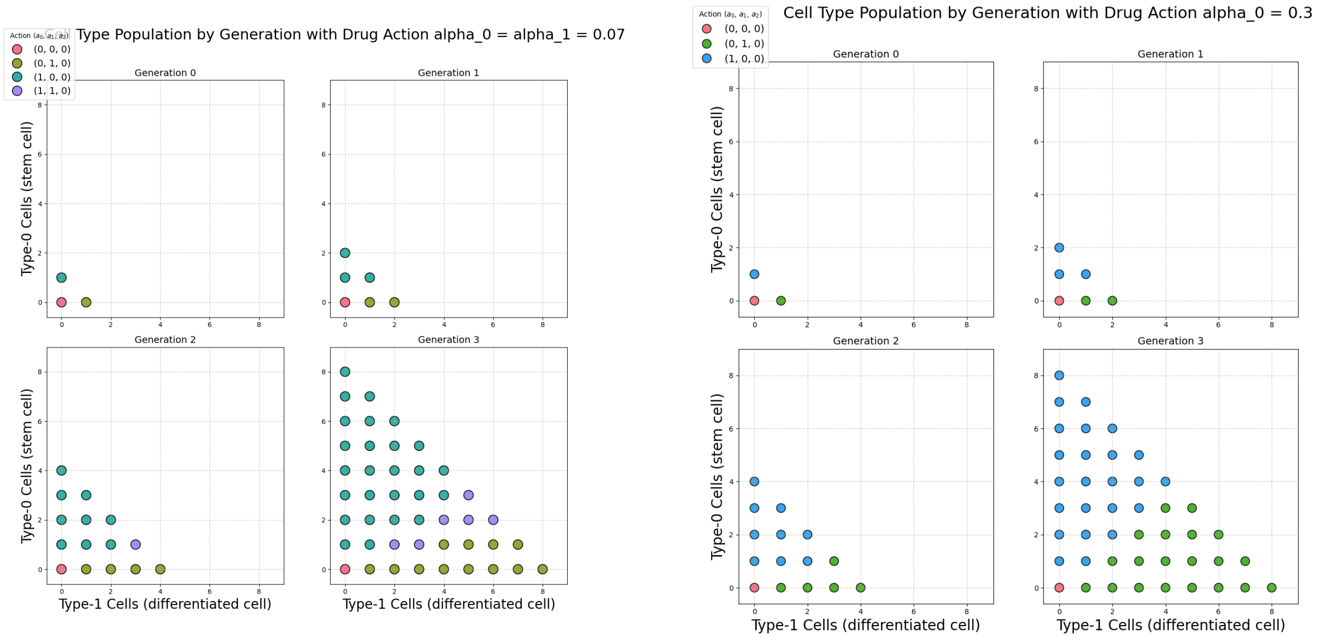


Figure 3.9: Optimal policy with two drugs, with and without plasticity. The left panel computes the policy with the same dose of drugs, while the right-hand side deals with different doses of the drugs.

In Figure 3.9, the color of each point indicates the optimal action at a given state when two drugs are available, with one targeting each cell type. On the left-hand side, we observe that the optimal policy switches depending on which cell type dominates the population. The drug targeting the stem cell (green points) is used more frequently because stem cells have a faster proliferation rate and are the main source of cancer. The drug that acts on differentiated cells (brown points) is only applied when necessary, when stem cells are almost non-existent in the population. When both types of cells are present in the population, the policy instructs us to administer both drugs. Therefore, the more heterogeneous the population, the more useful it is to combine both drugs.

The right-hand side shows that, when the dosage of a drug becomes important, the optimal policy focuses on one drug at a time. We then move from three strategies to two, applying either the drug for cancer stem cells or the drug for differentiated cancer stem cells. This helps us conclude that, when the fixed dose of the drug is low, the optimal policy considers combining both strategies, giving us three strategies. In the case of differentiated cancer cells without plasticity, the chosen set of parameters gives the same curve for the stem cells and the differentiated cells.

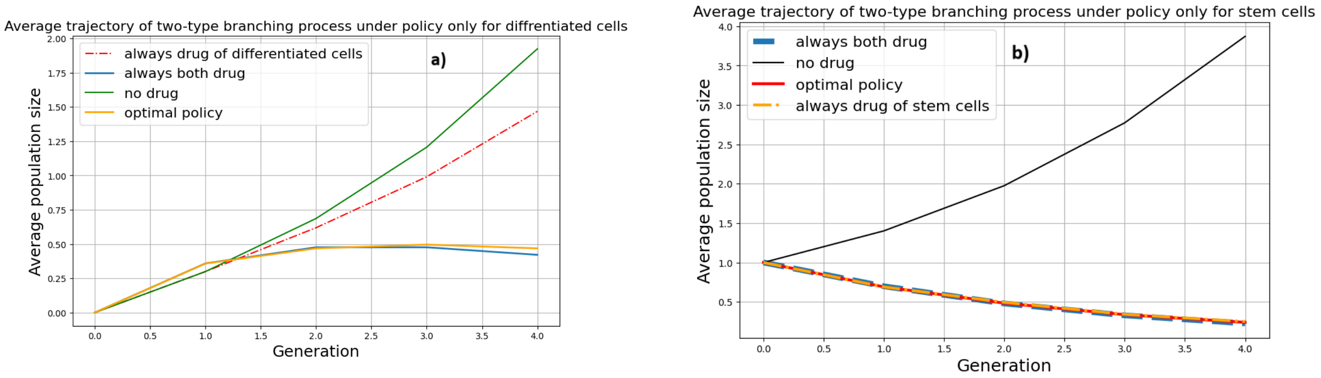


Figure 3.10: Average trajectory of cells as a function of generation under different policies with two drugs of different doses with and without cell plasticity.

As shown previously, the curves for no drug, the optimal policy and always administering a drug behave as explained in the figures showing the trajectories. However, it can also be seen in Fig.3.10 that applying only the drug to stem cells considerably reduces them, as does applying a drug that only acts on stem cells. The optimal policy remains the best approach, as it considers the toxicity of the drug.

Conclusion and Perspectives

The work presented in this thesis focuses on identifying the optimal strategy for administering a drug that kills cancer cells while considering its toxicity.

In the first chapter, we provided a general overview of branching processes and Markov decision processes. This helped us to understand that, due to their random reproduction, cancer cells can be modeled as a branching process, and that tools from decision-making theory can be used to solve such a problem.

In the final chapter, we mapped branching processes onto the Markov decision problem in order to solve the issue of drug strategy in cancer treatment in the case of cancer stem cells.

These studies reveal that when we consider a single-type branching process (i.e., cancer cells are only one type), the optimal strategy depends strongly on the drug dose and generation. When the dose is low, the drug is inexpensive, and the optimal strategy tells us to administer the drug at all times. When the drug is costly but not excessively so, we only administer it when necessary. Once the drug reaches a certain dosage level, the optimal policy tells us not to administer the drug and to allow the cancer to progress, because it is less costly not to treat the cancer than to administer a very costly drug.

The results obtained for the single type were also obtained for the two-type branching process (i.e. two types of cancer cell: cancer stem cell and differentiated cancer stem cell), but more factors were considered. We found that the strategy depends not only on the drug dosage, but also on how the drug acts on the cell. For this case, we considered two scenarios. The first scenario considered a single drug for both cell types and revealed that the optimal strategy was to administer the drug only when necessary, i.e. when plasticity was not being utilised. Once plasticity is used, however, the focus should be on stem cells, as they are the only source of cancer. In the second scenario, two different drugs were used, one for each type. In this scenario, we found that when the dosage is not costly and is the same for both drugs, the stem cell drug is administered only when there are almost no stem cells. The drug that acts on stem cells is

applied when there are stem cells. However, if we found two types of heterogeneity in the cell population, we applied both drugs.
This work allows us to conclude that the optimal strategy is not to administer or never administer the drug, but to administer it when necessary, depending on the dosage and generation. This strategy is called adaptive because it takes into account not just the size of the population, but also the toxicity of the drug.

Directions for further research include finding an optimal policy for chemotherapy cancer treatment under the same assumptions, using reinforcement learning (RL) tools. In real life, the tumor size is not directly observable; instead, we observe the blood biomarker level, which provides noisy information. So, the decision is based on a belief state of the population. This leads to a partially observable Markov decision process that captures real treatments.

Bibliography

[1] Louis Vermeulen and Hugo J Snippert. Stem cell dynamics in homeostasis and cancer of the intestine. *Nature Reviews Cancer*, 14(7):468–480, 2014.

[2] Henry William Watson and Francis Galton. On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144, 1875.

[3] R Durrett and S Moseley. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical population biology*, 77:42–48, 2010.

[4] Marek Kimmel and David E Axelrod. Multitype processes. In *Branching Processes in Biology*, pages 107–154. Springer, 2015.

[5] Niko Beerenwinkel, Tibor Antal, David Dingli, Arne Traulsen, Kenneth W Kinzler, Victor E Velculescu, Bert Vogelstein, and Martin A Nowak. Genetic progression and the waiting time to cancer. *PLoS computational biology*, 3(11):e225, 2007.

[6] Krishna B Athreya and Peter E Ney. *Branching processes*, volume 196. Springer Science & Business Media, 2012.

[7] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[8] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[9] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

[10] Ronald A Howard. Dynamic programming and markov processes. 1960.

[11] Deepti Mathur, Ethan Barnett, Howard I Scher, and Joao B Xavier. Optimizing the future: how mathematical models inform treatment schedules for cancer. *Trends in cancer*, 8(6):506–516, 2022.

[12] Nazila Bazrafshan and Mohammad Mehdi Lotfi. A finite-horizon markov decision process model for cancer chemotherapy treatment planning: an application to sequential treatment decision making in clinical trials. *Annals of Operations Research*, 295(1):483–502, 2020.

[13] Hans Clevers. The cancer stem cell: premises, promises and challenges. *Nature medicine*, 17(3):313–319, 2011.

[14] Linda JS Allen. *An introduction to stochastic processes with applications to biology*. CRC press, 2010.