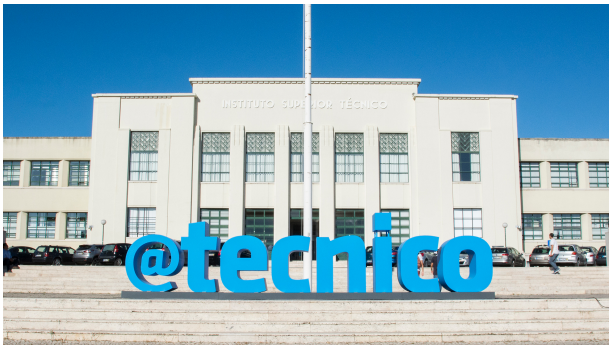
TÉCNICO LISBOA



Adaptive Control for Cancer Therapy based on Reinforcement Learning

Maria Inês de Mendonça Ferreira

Thesis to obtain the Master degree in
Electrical and Computer Engineering

Supervisor(s): Prof. Dr. João Manuel Lage de Miranda Lemos
Prof. Drª. Rita Maria Mendes de Almeida Correia da Cunha

Chairperson: Prof. Dr. João Fernando Cardoso Silva Sequeira
Advisor: Prof. Drª. Rita Maria Mendes de Almeida Correia da Cunha
Members of the Committee: Prof. Drª. Susana de Almeida Mendes Vinga Martins

November 2021

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Lisbon, 17th November 2021

Inês Ferreira

Acknowledgments

I would like to express my special thankful to my supervisors Professor João Miranda Lemos and Rita Mendes Cunha for all the tireless and remarkable support they provided me during the realization of this master's dissertation. It was a great pleasure to work with such educated and competent advisors throughout this journey and for that, thank you so much!

I would also like to express my greatest affectionate thanks to my family who always supported me in the best and worst moments during the completion of this dissertation, who always had comforting words, the best incentives and also the best assertiveness ever present.

In a way or another, with greater or lesser impact, several other people contributed a lot to the success of this dissertation, whether it was for personal, psychological or even calming support.

I can say for sure that, after the accomplishment of this master's dissertation, I feel like a much more developed person, either personally or professionally. It greatly contributed to the evolution of my way of thinking and of acting in situations of greater pressure without despair at first sight.

Although this dissertation is not something personified and does not speak for itself, I am grateful to all of those who supported me and to the theme of this dissertation itself, as the fight against cancer is a relentless battle that must be fought. And if there was anything I learned along this journey, it was that nothing wins at first, and you have to work hard every day to become better people, and this was transmitted to me by my family and supervisors, as such my sincere thanks!

Resumo

O processo de calendarização do tratamento do cancro é de extrema importância, uma vez que a evolução do tumor se trata de um processo dinâmico. Por outro lado, a questão essencial prende-se com a incerteza na dinâmica que motiva o uso de Controlo Adaptativo. Adicionalmente, sendo o objetivo principal o uso da adaptação baseada em dados, isto é, independente de um modelo, o uso de técnicas de Aprendizagem por Reforço é necessário.

Nesta tese, em primeira instância é considerado um modelo simplificado de crescimento do tumor, também conhecido como Modelo Logístico, em que o Sistema Imunitário e a Angiogénesis não são considerados. De seguida, estende-se este modelo aos blocos da Farmacodinâmica e Farmacocinética. Por fim, considera-se o modelo de crescimento de tumor que tem influência do Sistema Imunitário.

Tendo sido adotada uma estratégia de controlo adaptativa, as ações são baseadas em testes e nos correspondentes resultados alcançados, logo o objetivo é então melhorar e ajustar essas ações em relação à recompensa obtida durante este processo de aprendizagem, designado Aprendizagem por Reforço.

Assumindo que não sabemos nem queremos estimar o modelo, podemos usar a técnica *Q-Learning* para determinar a política de Controlo Ótimo que permite a minimização da função sem o conhecimento do modelo.

Assim, através de um aumento gradual do número de estados do modelo, os resultados e as comparações entre cada etapa do modelo são apresentados, na medida em que foram aplicados distintos controladores e algoritmos para obter os resultados mais otimizados relativos ao tratamento do cancro, isto é, resultados que permitissem a erradicação do cancro partindo de um valor inicial do volume do tumor, e que apresentassem um comportamento descendente e controlado em torno de uma referência imposta.

Palavras-chave: Controlo Adaptativo, Aprendizagem por Reforço, Modelo Logístico, *Q-Learning*, Controlo Ótimo, Tratamento do Cancro.

Abstract

The process of time-scheduling cancer treatment is extremely important since the evolution of the tumor is a dynamic process. Uncertainty in the dynamics is a major issue that motivates the use of Adaptive Control. This master dissertation focuses on the use of data-based adaptation, that is, regardless of a model, and relies on Reinforcement Learning techniques to achieve this goal of adaptation.

In this thesis, a simplified tumor growth model, known as the Logistic model, is first considered, not taking into account the Immune System (IS) and Angiogenesis (Angio). Then, this model is extended to include the Pharmacokinetics and Pharmacodynamics blocks. Finally, we consider a tumor growth model that is influenced by the IS.

Having adopted an adaptive control strategy, the actions are based on tests and corresponding results achieved, so the goal is then to improve and adjust those actions with respect to the reward function obtained during this learning process called Reinforcement Learning.

In particular, the Q-Learning technique is used to approach the optimal control policy that maximizes a reward function, without actually estimating the model that is assumed to be unknown.

Then, through an increase on the amount of states of the model, the results are presented and the comparisons between each stage of the model are yielded, in the sense that there are distinct controllers and techniques used, depending on the parameters and the states of the model, since we want to achieve the best results regarding Cancer Treatment, which means that the optimal results achieved would allow to eradicate the cancer by starting at an initial value of the tumor volume, and that show a descendant overall behavior controlled around the given reference.

Keywords: Adaptive Control, Reinforcement Learning, Logistic model, Q-Learning, Optimal control, Cancer Treatment.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The main goal of this master dissertation is to develop an adaptive optimal strategy for cancer therapy schedule based on state models, considering the Immune System response as well as the Angiogenesis subsystems. This optimal schedule will be obtained through Reinforcement Learning alongside Adaptive Control.

1.1 Motivation

Nowadays cancer is considered as one of the deadliest diseases. It does not have a single cause, Instead, there are several external causes (present in the environment) and internal causes (such as hormones, immunological conditions and genetic mutations), called factors that initiate the onset of cancer. It represents a group of many diseases characterized by DNA mutations resulting in disorders that may occur naturally or may be caused by environment interactions, such as very high radiation exposure. Between 80% and 90% of cancer cases are related to external causes, such as changes in the environment caused by man himself, habits and lifestyle, which can increase the risk of different types of cancer [1].

These changes regarding the environment modify the genetic structure (DNA) of cells, and are related to five types of environment: the environment in general (water and air quality), the social and cultural environment (lifestyle and habits), the working environment (chemical and related industries), and the consumption environment (food habits and administered drugs). Environmental risk factors for cancer are called carcinogens [1].

Cancer covers more than 100 different types of malignant diseases that have in common the disorderly growth of cells, which can invade adjacent tissues or close organs [2]. Given the uniqueness of the organism of each being, it becomes really difficult to track the spread of the disease among all the cells. According to the *Darwin's* selection process, only the aptest cell will survive and give rise to the tumor that is the so called cancer tumor [3]. Dividing rapidly, these cells tend to be uncontrollable and very aggressive, causing the formation of tumors [2].

A tumor corresponds to the increase in volume observed in any part of the body. When the tumor is due to an increase in the number of cells, it is called a neoplasm - which can be benign or malignant. Unlike cancer, which is a malignant neoplasm, benign neoplasms grow in an organized and generally slow manner, and have very clear limits. They do not invade neighboring tissues nor develop metastases – spread of the tumor over other parts of the body [4]. The main problem persists in the sense that it has not yet been possible to achieve an adequate treatment that allows the tumor to be destroyed as quickly as possible, in an effective and healthy way.

The year 2020 was marked by the coronavirus pandemic which had a massive impact on people with cancer, in addition to the constraints added to normal life. Despite that, cancer research has continued even at a slower pace. Behind closed doors, in socially distanced labs, researchers continued to start new cures for cancer and to make new discoveries. These breakthroughs represent a great incentive and are extremely motivating for the continuation of research in this scientific area so that new knowledge can be acquired with each passing day, hoping that one day the cure for cancer will be discovered.

The foremost motivations concern Optimal Control in the sense that we want to optimize the arrangement between therapeutic effects and its adverse effects, Adaptive Control to deal with intra and inter patient variability, and Reinforcement Learning as a way to connect the two previous approaches for non-linear systems in a way that is data-driven.

1.2 Objectives

This master dissertation aims to present a preliminary feasibility study of the development of cancer therapy strategies based on Optimal Adaptive Control for non-linear systems, through Reinforcement Learning techniques.

Therefore, there are two main objectives:

1. Development of a state model of cancer evolution, including drug administration, with and without considering the Immune System and the Angiogenesis influence. This object is subsidiary to the next one.
2. Application of Reinforcement Learning techniques to the mathematical model of the tumor growth, in order to design personalised therapy schedules for cancer treatment.

1.3 Contributions

In this master dissertation, the analysis of the tumor growth mathematical model is proposed alongside with the Immune System response as well as the Angiogenesis process.

Therefore, several algorithms are used to control the model and, since there is a persistent failure with regards to the cancer therapy, in the sense that it must be efficient and safe, this study is improved by:

1. Considering several Reinforcement Learning techniques;
2. Attempting to achieve an optimized treatment strategy;
3. Grounding the treatment strategy on Adaptive Control.

1.4 Thesis outline

In chapter 2, a brief review of the studies performed until now, concerning the cancer treatment, the tumor growth subsystems and the available Reinforcement Learning techniques, is made.

In chapter 3, the cancer models are discussed in detail in the sense that, better understanding and more effective control of tumor growth and spreading can be achieved by describing them as a dynamical system and monitoring their evolution. A review of the state of art is also provided in this chapter at the pharmacology level.

In chapter 4, the tumor volume growth model is mathematically explored, along with the logistic growth subsystems, culminating in the drug administration results.

In chapter 5, the control of the logistic model is analysed throughout a mathematical perspective, namely regarding the linearization and discretization of the system both considering and not considering the Immune System influence.

In chapter 6, the used algorithms for RL are presented, as well as the adaptive control techniques concerning the Q-Learning.

Then, in chapter 7, the results with respect to the *Matlab* programming regarding the system behavior, considering distinct dimensions, are yielded.

After that, in chapter 8 there is an analysis of the Q-Learning algorithm regarding several parameters changes and its influence.

Finally, in chapter 9 the conclusions are provided regarding the achievements and limitations of the work developed.

Chapter 2

State of Art

This chapter provides a review over the literature presented to date regarding cancer treatment alongside with several Reinforcement learning-based techniques that allow the control of the tumor growth model, as well as to optimize and personalize therapy.

2.1 Cancer treatment

The increasing threat of cancer to human life has lead to the need of improvement on the research in numerous fields, namely regarding the necessity to properly schedule cancer treatment to ensure effective and safe treatment [5]. According to the *World Health Organization*, it causes about 12.5% of all the deaths around the world [6].

Unfortunately, cancer has been around for many thousand years. As a matter of fact, the area of medicine that works with cancer came from the ancient Greek and Latin: the oncology [6]. The so called father of the modern medicine Hippocrates [7] started to draw and analyse tumors from outside of the body once the ancient Greeks did not believe in the study of the interior of death bodies [6]. Therefore, the bodies were studied in the sense that there existed four types of fluids in the body, and this theory was popular until the 19th century when the cells were discovered and so, the tumor could be at any part of the body, making the previous theory irrelevant [6]. In the 18th century, English surgeon Campbell De Morgan [8] speculated that cancer started locally and then spread in the body, by using the microscope for the first time. Later, in the 19th century, doctors realized that a cleaner environment would contribute to less infections as well as to a safer recovery of the patients submitted to surgery. By the end of this century, the radiation was discovered leading to the first cancer treatment without surgery involved [6]. Finally, in the 20th century, there was a developed study concerning the reasons why people were more or less likely to get different kinds of cancer [6]. It is still important to mention that since World War II the cancer treatments have been improved although there are still some improvements that need to be made such as the non-existence of treatments for all types of cancer, and also the non-generalization of the existing treatments.

More recently, in the year of 2020, despite the coronavirus pandemic situation, scientists from all over the world kept searching and developing new cancer treatment strategies, leading to 12 cancer research breakthroughs [9]:

1. University of Manchester discovered that tumors can lead to the growth of a type of cell found in our body that helps the tumor to hide from the immune system. Then, a special molecule can be produced in order to be blocked with targeted drugs, allowing the immune system to recognise and destroy the hidden cancer cells;

2. The European Institute of Oncology in Italy helped to discover a gene that could be used as a new target to design new breast cancer drugs, since it helps the growth and spread of breast cancer around the body to other organs;
3. University of Leicester discovered how a particular genetic mutation accelerates lung cancer spread in patients, leading to the identification of a new molecular mechanism that can be used to treat lung cancer;
4. The University of the Basque Country in Spain discovered the importance of a protein on the surface of cancer cells, affirming that blocking this protein could prevent cancer cells from undergoing some of the changes that lead to the spread over the body;
5. A large collaborative group of cancer researchers from around the world, showed that by gathering immunotherapy with tumor necrosis factor drug, the tumors that were before unaffected by immunotherapy, could then be eradicated;
6. Institute of Medical Research in Australia has been studying a new drug that could become one of the first targeted treatments for triple-negative breast cancer;
7. Scientists studied how chemotherapy works by looking at the DNA of microscope worms, that allowed to map out exactly how chemotherapy interacts with DNA, which one day might lead to a less aggressive and more effective chemotherapy;
8. Scientists from Australia are confident that they will be able to begin clinical trials on a new cancer vaccine by 2023;
9. Scientists in the Netherlands have made progress regarding a new treatment option for patients with bladder cancer, showing that combining two types of immunotherapy before surgery could be an effective way to prevent the recurrence of the cancer;
10. The UK team os researchers, discovered new compounds that inhibit a protein that plays a harmful role in tumor progression and metastasis in pancreatic cancer;
11. Scientists discovered that a cheap drug, commonly used to treat parasitic worm infection, could be a turning point regarding the treatment for prostate cancer;
12. VU University Medical Center in the Netherlands showed that a new drug could help overcome resistance to treatment in advanced stage cases of multiple myeloma, which is a cancer that forms in a type of white blood cell (plasma cell).

Depending on the advanced stage of the tumor and its type, there are several treatments such as surgery, chemotherapy, immunotherapy or radiotherapy [10]. The chemotherapy treatment for example, started at the beginning of the 20th century [11] in hopes of narrowing the administration of drugs due to their dangerous dose-limiting side effects [12]. At the cancer pharmacology level, the studies are guided by the development of more effective drugs as well as by the improvement of the drug administration by gathering specific pharmacokinetic and pharmacodynamic details of the drug grounded on available clinical trials [5].

2.2 Logistic Growth

The mathematical model for the cancer dynamics must take into account the tumor growth as well as the patient Immune System reaction to the tumor growth [5], considering that there is the need to choose the

optimal treatment that ensures the minimization of the cancerous cells, without endangering the patient [13]. Up to date, there were several Tumor Growth Models studied [14], being shortly presented in table (2.1).

Table 2.1: Tumor Growth Models.

Models	Equations
Linear growth	$\frac{dV}{dt} = k_g - d * V$
Exponential growth	$\frac{dV}{dt} = k_g * V - d * V$
Logistic growth	$\frac{dV}{dt} = k_g * V * (1 - \frac{V}{V_{max}})$
Gompertz growth	$\frac{dV}{dt} = k_g * V * \ln(\frac{V_{max}}{V})$

2.2.1 Immune System

The Immune System is related to the capacity of the organism to react to external infections. It is essential to each being survival, since without its presence, we would be susceptible to the attack from viruses or bacteria or diseases. The IS is spread over the body and has the ability to detect foreign tissue when compared to our tissue, the so called self-tissue. Dead cells are also recognized and expelled from our organism [15].

Cells travel through the bloodstream or in specialized vessels called lymphatics. Lymph nodes are small, round or bean-shaped clusters of cells. The spleen is an organ found in virtually all vertebrates. Both provide structures that facilitate cell-to-cell communication.

The bone marrow and thymus represent training grounds for two cells of the immune system: B-cells and T-cells, respectively. The development of all cells of the immune system begins in the bone marrow with a hematopoietic (blood-forming) stem cell, since all the other specialized cells arise from it. Due to its capacity to generate the entire Immune System, this is the most important cell in a cell transplant, specially to recover tissues damaged by diseases or traumas. Therefore, despite in most cases the development of each cell type is independent from other cell types, these stem cells represent the basis of the entire Immune System [16].

Even though all components of the immune system interact with each other, there are two categories usually considered: the Innate Immune System and the Adaptive Immune System [16].

The Innate Immune System relies on cells that require no additional training to perform their tasks. It is made of defenses, essentially barriers against infection that can be activated immediately once a pathogen attacks. These cells include neutrophils, monocytes, natural killer (NK) cells, and a set of proteins known as the complement proteins. Innate responses to infection occur rapidly and reliably, since anything that is identified as foreign or non-self is marked as a target for the innate immune response [17]. The Innate Immune System is composed by:

- Physical Barriers - skin, gastrointestinal and respiratory tract, nasopharynx, cilia, and other body hair.
- Defense Mechanisms - secretions, mucous, bile, gastric, saliva, acid, sweat, tears.
- General Immune Responses - inflammation, complement, non-specific cellular responses.

  - Inflammation - immune cells brought to the infection site through the increase of the blood flow to that area.
  - Complement - pathogens are marked for destruction through the creation of wholes in the pathogen cell membrane.

On the other hand, the Adaptive Immune System only relies on T-cells, which concern the thymus, and B-cells regarding the bone marrow, two cell types that require extra knowledge to learn not to attack our own cells. The advantages of the adaptive responses are their long-lived memory and the ability to adapt to new germs. Have you ever wondered how your recovery time decreases after a previously exposure to small infections, flu or common cold? The Adaptive Immune System, also known as Acquired Immunity, is responsible for that faster recovery, since it strategically sets an immune response [18]. Each one of these cells follows a different maturation and development process, more specifically:

- B cells - after formation and maturation in the bone marrow, the naive B cells move into the lymphatic system to circulate throughout the body. Once there, the naive B cells reach an antigen, which starts their maturation process. Each B cell has one among millions of distinctive surface antigen-specific receptors that are intrinsic to the organism's DNA.
- T cells - after formed in the bone marrow, T progenitor cells migrate to the thymus to mature and become T cells. Once there, the developing T cells start to express T cell receptors (TCRs) and other receptors (CD4 and CD8).

B cells also express a specialized receptor, called the B cell receptor (BCR), that assist with antigen binding, as well as internalization and processing of the antigen. All T cells express T cell receptors, and one of CD4 or CD8, not both, meaning that some T cells represent CD4 and others CD8 [18].

The white blood cells and the blood platelets are the main character of the Immune System. The first ones, also known as leukocytes, circulate over the body in blood and lymphatic vessels, parallel between veins and arteries, and are always aware of possible intruders in order to start the multiplication process when detecting a target, sending an alert to the other types of cells to do the same. The leukocytes are spread over the body, called lymphoid organs [15]. On the other hand, the blood platelets play a much more important role in our Immune System than previously thought. Besides their traditional role in blood clotting and wound healing, platelets act as first-rate attendants when a virus or bacteria enters the bloodstream, a role that was thought to be played by the white blood cells until then. Professor Eric Boilard, from Laval University in Canada explains that when a foreign body enters the organism, it induces the antibodies formation, meaning that the next time that the same foreign body enters the organism, these antibodies quickly agglomerate on the invader's surface, forming antibody complexes that set an inflammatory defense response. Since the platelets have sensors capable to detect those antibody complexes, the theory regarding the role of these blood cells begun [19].

An efficient immune response depends on the appropriate distribution and positioning of the immune cells over the dynamic tissue micro environments, which is controlled by the vascular network and its interactions with circulating immune cells [20]. Therefore, it is important to know the organs of the Immune System [21], as expressed in table 2.2.

2.2.2 Angiogenesis

Angiogenesis plays a key role regarding the growth of cancer cells and metastasis of tumors [29]. It can be considered both as an attractive target and a major challenge in the treatment of cancer, since it is the process of formation and remodelling new blood vessels and capillaries, from growth of pre-existing blood vessels [30].

Table 2.2: Immune System Organs.

Organs	Description
Thymus	Located in the upper chest. Immature lymphocytes leave the bone marrow and find their way to the thymus where they are taught to become mature T-lymphocytes [22].
Liver	Major organ responsible for synthesizing proteins of the complement system. Contains large numbers of phagocytic cells which ingest bacteria in the blood as it passes through the liver [23].
Bone Marrow	Sponge-like tissue found inside the bones, where most immune system cells are produced and multiplied. Where all cells begin their development from primitive stem cells [24].
Tonsils	Collections of lymphocytes in the throat [25].
Lymph Nodes	Collections of B-lymphocytes and T-lymphocytes throughout the body [26].
Spleen	Collection of T-lymphocytes, B-lymphocytes and monocytes. Helps to filter the blood and provides an interaction site for the Immune System organisms and cells [27].
Blood	Circulatory system that lead,s from one part of the body to another, cells and proteins of the Immune System [28].

2.3 Reinforcement Learning

Machine Learning is a part of Artificial Intelligence (AI) alongside Computer Science that allows the improvement of the learning process' accuracy through the use of data and algorithms [31]. Therefore, the Machine Learning hierarchy schema is present in figure 2.1.

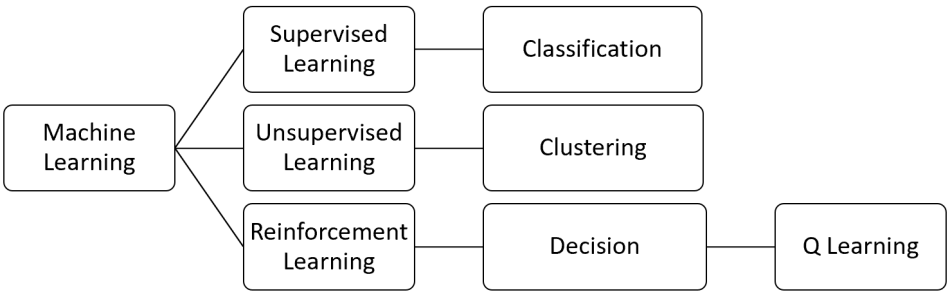


Figure 2.1: Machine Learning schema.

One might be able to achieve the desired goal through methods such as Linear Quadratic Regulator [32], Recursive Least-Squares approaches or even *Ricatti* equations [33] to calculate, online and offline, the feedback control gain, by estimating the model parameters at each iteration. Although, there are some drawbacks to the use of these methods since they require the previous knowledge of the model parameters.

The Linear Quadratic Regulator is a well-known design technique that provides practical feedback gains through the internal computation of the *Ricatti* equations [32] both for continuous and discrete-time systems. Recursive Least-Squares was discovered by Gauss but remained ignored until 1950 when Placket found the work developed earlier in 1821. It is an adaptive filter algorithm that recursively calculates the parameters of the filter that minimize a weighted linear least squares cost function, and through it converges to the desired solution extremely fast [34]. On the other hand, the feedback gain might also be calculated through the Generalized Policy Algorithm for Linear Quadratic Regulator aiming

to minimize the cost-to-go function that stands for the performance measure. The linear controller is obtained by solving a sequence of finite-horizon problems, where each problem is considered a static quadratic program, and through this the feedback gain alongside with the parameters of the model are estimated along time [32].

The main role of the Reinforcement Learning is to avoid the use of models, meaning that the environment must be set but then, the optimal control schedule must be chosen to fulfill the main goal [35]. To better illustrate the structure of the RL, a schema is presented in figure 2.2.

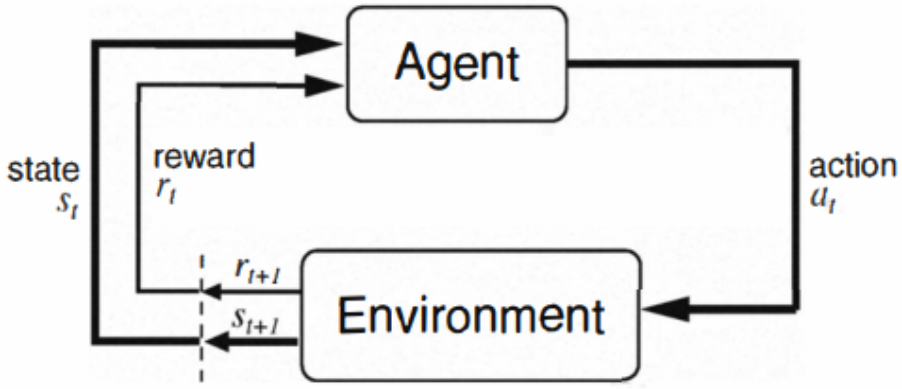


Figure 2.2: Reinforcement Learning structure.

From figure 2.2, we can state that there is always an agent and the environment, where given the state and the reward from the environment evaluation, the individualized optimal policies as a function of the state variables are estimated and the agent has to choose the action for the next state in order to apply it to the environment and create the loop system that allows the estimation of the overall optimal control [36].

The main objective of cancer treatment is to reduce the tumor volume and if possible eradicate it. Thus, to apply the same treatment to every patient, there is the need to generalize the used method without considering the specific ”constraints” of each patient, named model parameters in a mathematical language.

2.3.1 Q-Learning background

A Reinforcement Learning method called Q-Learning aims the estimation of the performance of the controller without knowing the model and without wanting to estimate it [37]. There are also studies concerning learning optimal regimens from patient data created through clinical reinforcement trials [38]. Therefore, this algorithm is used to develop a general framework granting the design of controllers for drug administration for cancer therapy, subject of interest in this master's dissertation. Therefore, this method regulates the drug infusion to each patient even if there is no accurate knowledge about the patient system, non-linearity, Pharmacokinetics, Pharmacodynamics, drug administration and its side effects, leading to improvements on the clinical goals [39].

In a simplified perspective, considering discrete states which is the opposite of this work, the Q-Learning technique can be seen as a mouse in a maze that wants to reach the cheese reward at the end. To do that, the mouse can choose between several paths. Each path as a reward that gives more energy to the mouse to keep the search for the cheese. Therefore, the mouse must analyse each possibility and choose the action that provides higher energy when added to the energy from the previous state. Thus, the optimal path is the one that leads to the cheese and, at the same time, allows the greatest accumulation of energy [40].

A visual interpretation of this example is presented in figure 2.3.

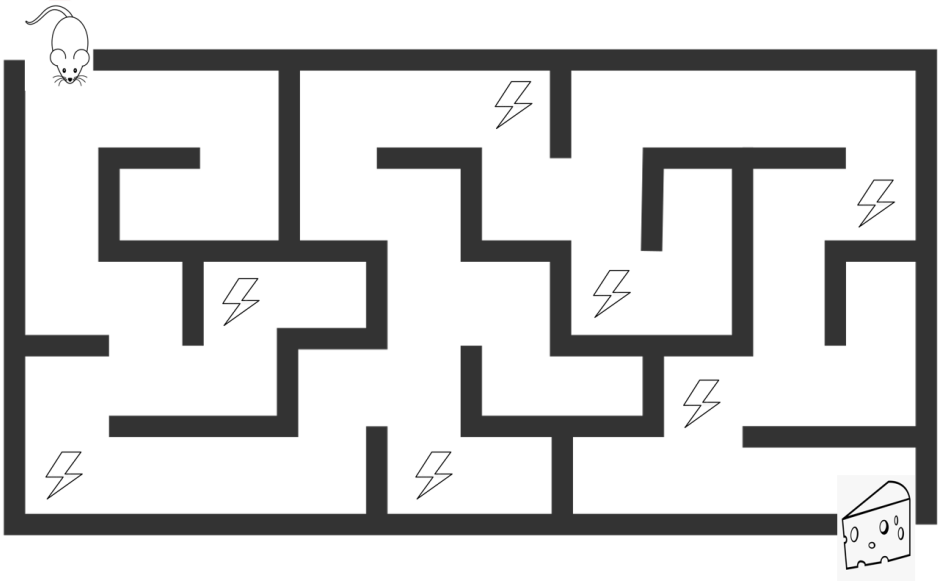


Figure 2.3: Mouse escaping from a maze.

This can also be recognized as a search algorithm of Artificial Intelligence (AI), in the sense that there is a cost function, several possible actions to take and a final objective. The actions must be taken considering the constraints of the problem, and for each action there is an associated reward that influences the cost function. This cost is added to the total cost function of the previous state, and the last is compared between each available action until there is no other action to apply. Finally, by learning the consequences of the distinct actions, it is possible to proceed until the goal of the problem and returning the optimal cost function [41].

Furthermore, there are several articles and papers with regards the application of the Q-Learning to control the tumor growth for example through chemotherapy [42], and to develop effective treatment regimes for individual patients [43].

Regarding the drug administration for cancer treatment based on Reinforcement Learning, the goal is to reduce the tumor as much as possible, considering the tracking of a reference through the application of a controller. The Q-Learning technique stands for ”Learning with quality”, and so, quality in this case represents how useful is an action going to be in gaining some future reward. Thus, the optimal policy at each state concerns the smallest difference between the reference to track and the tumor volume curves, such that the overall cost will be minimum and the behavior of the tumor growth will be the desired one.

Chapter 3

Cancer Models

This chapter describes the nonlinear state models for cancer, that will be used further to simulate the algorithms to develop in order to motivate the controllers structure.

Therefore, an important issue is to represent the cancer models through a blocks diagram to better explain the input and output variables involved, as presented in figure 3.1.

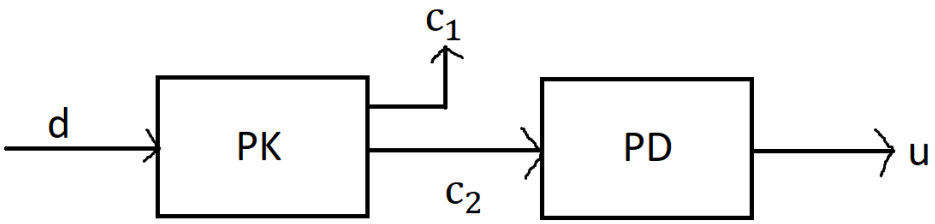


Figure 3.1: Block diagram of the cancer models.

3.1 Pharmacokinetics

Pharmacokinetics (PK) models are created to elucidate the transformations that a drug undergoes in an organism and the rules that determine this fate. So it performs a huge role in drug absorption, distribution, metabolism, and excretion [44].

For now, let us assume that the PK diagram block only receives as input parameter the drug infusion rate T and produces the effect concentration c that is a vector where in each entry there is the concentration in a given compartment. This PK model represents a state model where the state variables are represented by c and the model parameters are A and b, and so it can be described as follows:

$$\frac{dc}{dt} = Ac + bT. \tag{3.1}$$

Compartmental Models are linear approximated models that simplify mathematical modeling by translating the diffusion of a chemical species typically connected to diseases, and in the case that matters here, its state is made of the drug concentration of each compartment. Mass conservation and positivity are two main features of these models since for any system closed to transfers of matter and energy the system's mass must endure the same along time as well as the state variables to remain non-negative. The first feature imposes a structure on A.

In this report a two compartment model such as the one used in Biomedical Engineering is considered for first approach as described in figure 3.2.

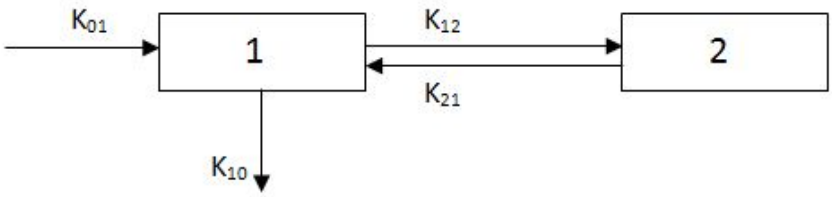


Figure 3.2: Quaternary model with two compartments.

In figure 3.2 there are two compartments connected with each other by two flow parameters k_{12} and k_{21} that determine the drug flow between compartment 1 and 2, and vice-versa. These drug flows work in the sense that when one has low amount of drug dosage, then the other compartment gives away that needed amount reaching an equilibrium.

Once the two compartmental model is used, one must realize that the state-space model performed on equation (3.1) can now be described with two state variables, a 2-by-2 matrix A and a 2-by-1 vector b, as well as through an output equation defined by a 1-by-2 matrix C that relates the output of the system to one of the state variables, in this case c_2. Therefore, the state-space model presented in equation (3.1) can now be replaced by:

$$\begin{bmatrix} \frac{dc_1}{dt} \\ \frac{dc_2}{dt} \end{bmatrix} = \begin{bmatrix} \frac{-k_{12}-k_{10}}{V_1} & \frac{k_{21}}{V_1} \\ \frac{k_{12}}{V_2} & \frac{-k_{21}}{V_2} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{V_1} \\ 0 \end{bmatrix} u, \tag{3.2}$$

where the input signal u is given by the first line of the identity matrix I, V_1 and V_2 represent the volume of each compartment 1 and 2, respectively, k_{ij} depend on the drug, and the output is then $y = c_2$.

3.2 Pharmacodynamics

Pharmacodynamics (PD) relates the drug concentration with its effect in the patient's organism. In order to do that, we need to relate the plasma concentration c_2 with the drug effect denominated by u which translates into a nonlinear equation known as *Hill* equation described as follows:

$$u(t) = u_{max} \frac{c_2^{\alpha}(t)}{c_{50}^{\alpha} + c_2^{\alpha}(t)}. \tag{3.3}$$

The PD block receives as input the effect concentration from the PK model and, depending on the administered drug, it returns the drug effect between 0 and u_{max}, value from which it saturates. In its turn, the parameter c_{50} represents the drug concentration that produces half of the effect.

Chapter 4

Tumor Growth Model

The Tumor Growth Model (TGM) evaluates the status of the tumor regarding its size and effect, but first, it is important to define the concept of tumor. A tumor is a swelling of a part of the body, generally without inflammation, caused by an abnormal growth of tissue, whether benign or malignant, in this exact case it is the accumulation of cancer cells. There are several tumor growth models aimed at testing growth theories as well as comparing them in practice to analyze the effect of drugs in combating tumor growth, in attenuating it and eventually in an attempt to decrease it.

4.1 Logistic Growth Model

One specific model that represents the tumor growth issue is the Logistic Growth model that can be mathematically represented by:

$$\frac{dV}{dt} = aV(1 - \frac{V}{K}) - \beta uV. \tag{4.1}$$

The Logistic Growth model is a nonlinear equation due to the quadratic term in V and the product between the drug effect u and the tumor volume V. At the same time, the parameters a and K depend on another constraints from the Immune System (IS) and the Angiogenesis diagram blocks.

4.2 Logistic Growth Subsystems

As mentioned before, the Logistic Growth model depends on some parameters that arise from interactive subsystems, such as the Immune System (IS) and the Angiogenesis (Angio), which will be described down below [20].

4.2.1 Immune System

The Immune System is a host defense mechanism comprising many biological structures and processes within an organism aiming the protection against diseases. It is comprised by two types of cells: the secret agents and the warriors. The former are intended to understand whether a particular particle that appears in the plasma is part of the organism or not while the second act when secret agents discover that the particle is not self which is to say that it is not part of the organism, that is, it is not benign. The Immune System plays a very important role in the fight against cancer but not when analyzed exclusively, it only helps in the fight but it is not enough, and changes from individual to individual, making each patient a unique and separate case [20].

Therefore, these disorders of the IS can result in autoimmune diseases, inflammatory diseases and cancer. The interactions between the IS and the cancer are extremely complex and traduced by nonlinear dynamics such as:

$$\dot{r} = \alpha_2(1 - \beta_2 V)Vr + \gamma_2 - \delta_2 r. \tag{4.2}$$

While r is the immunocompetent cell density related to the triggered immune cells during the reaction, and once again V is the tumor volume, the parameters α_2, β_2, γ_2 and δ_2 are constant coefficients whose values are presented in table (4.1).

Table 4.1: Immune System parameters.

Parameters	Value
α_2	0.00484
β_2	0.00264
γ_2	0.1181
δ_2	0.3998

4.2.2 Angiogenesis

One way to fight against cancer is to prevent the Angiogenesis process which consists of the creation of blood vessels that supply food to the tumor allowing it to develop and grow from existing blood vessels, and so it pursues the growth of the vasculature by processes of splitting and sprouting. Angiogenesis is a normal and vital process in growth and development, as in wound healing and in the regeneration of the tissue. Nevertheless, it is also an essential step in the transition of tumors from a benign to a malignant state, leading to the use of angiogenesis inhibitors in the treatment of cancer. This blocking process is called antiangiogenesis. Instead of combating the multiplication of cancer cells, it is advisable to target the cells responsible for creating blood vessel walls, which is not the same as killing the tumor since the antiangiogenic treatment only attenuates the tumor by limiting its support [20].

4.3 Drug administration

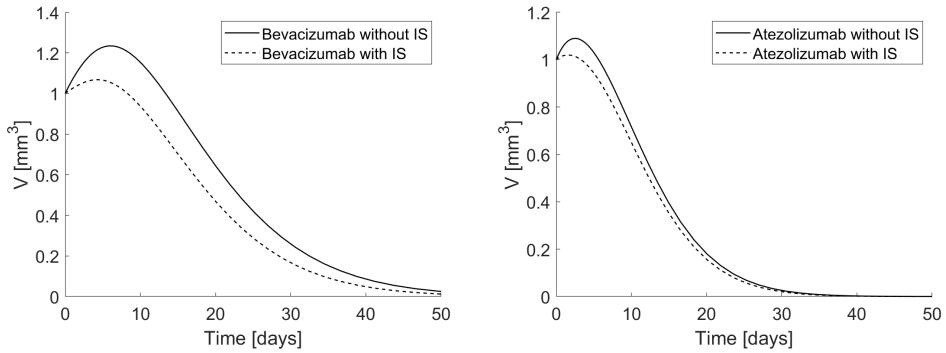
In order to better understand the meaning and the dependence between each one of the cancer model blocks previously presented in this chapter, a *Simulink* file was created together with a *Matlab* script with the initialization of the crucial parameters for that implementation. The *Matlab* script contains variables declaration for two different drugs: Bevacizumab and Atezolizumab. The parameters considered for each drug can be presented in table (4.2).

In this particular case, the results presented concern the IS influence in the system but first, a comparison between the tumor volume evolution for each drug with and without the IS influence, as described in figure 4.1.

From figure 4.1a, we can see that the curve without IS reaches a higher maximum than the dashed curve standing for the tumor volume evolution over time with IS influence. This means that the addition of the IS subsystem leads to a faster response to the cancer tumor, which is exactly what was expected since the IS is a beneficial factor. Regarding figure 4.1b, the exact same behavior is observed but now, due to the Atezolizumab drug parameters, the maximum value is lower than the one reached by the Bevacizumab drug. Another important comparison concerns the time of eradication of the tumor

Table 4.2: PK and PD parameters for Bevacizumab and Atezolizumab drugs.

PK	Bevacizumab	Atezolizumab
V_1	2660 ml	3110 ml
V_2	2660 ml	3110 ml
k_{12}	0.223 day^{-1}	0.3 day^{-1}
k_{21}	0.215 day^{-1}	0.2455 day^{-1}
k_{10}	0.0779 day^{-1}	0.0643 day^{-1}
PD	Bevacizumab	Atezolizumab
c_{50}	11.4274 mg/Kg	7.1903 mg/Kg
u_{max}	1	1
α	1	1



(a) (b)

Figure 4.1: (a) Tumor volume evolution with and without IS over time for Bevacizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$, and (b) tumor volume evolution with and without IS over time for Atezolizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

since after a month the tumor volume for the Atezolizumab drug administration is approximately 10% of the initial value whereas it is approximately 20% with IS and 30% without IS of the initial value for the Bevacizumab drug.

Regarding the *Simulink* tests performed with IS, it is necessary to illustrate the behavior of the IS evolution through time, as described in figure 4.2.

From figure 4.2, we managed to get that the IS effect starts from a maximum value and then starts to decrease until a constant value where it remains until the end of the simulation, meaning that instead of presenting an inconsistent behavior, the IS adjusts the level of influence required considering the drug administration and the tumor volume evolution. Thus, it is always helping to fight cancer.

Then, the results achieved for the concentration of Bevacizumab and Atezolizumab drugs over time, representing the PK model, are presented in figure 4.3.

By looking into figure 4.3 we can affirm that over time the Bevacizumab drug concentration increases and then begins to reduce the slope when it reaches the maximum of this drug concentration. On the other hand, when analyzing the dashed curve we can see that the drug concentration of Atezolizumab increases over time as well and achieves a higher maximum value due to this drug's parameters, but then it starts decreasing its slope even slower than the Bevacizumab drug, which means that the effect of the Atezolizumab drug on the patient will be more significant, as will be seen in figure 4.4.

Regarding figure 4.4a, for parameters such as $\alpha = 1$ and $u_{max} = 1$, the drug effect of the Bevacizumab drug reaches lower values than the effect of the Atezolizumab drug due to the longer-lasting presence of the concentration of the last drug mentioned. It is also important to state that the increasing

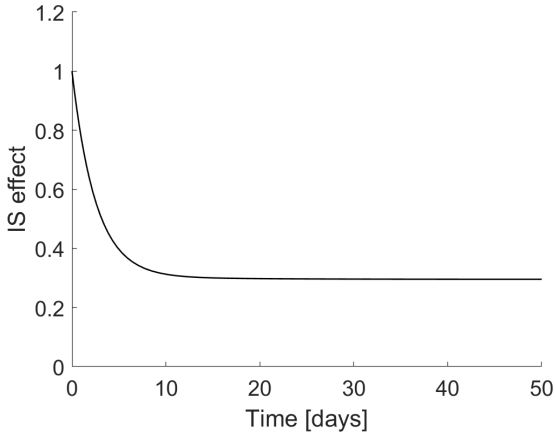


Figure 4.2: IS evolution over time for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

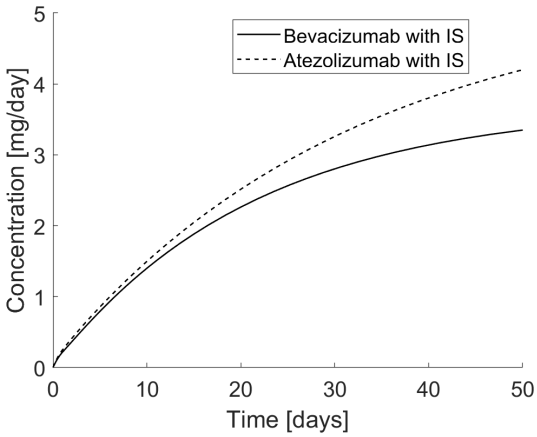


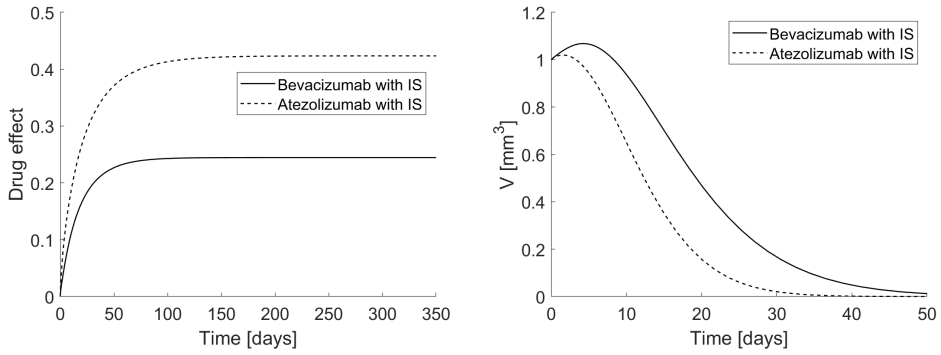
Figure 4.3: Drug concentration with IS over time for Bevacizumab and Atezolizumab, for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

slope of the Atezolizumab drug is more pronounced than the one of the Bevacizumab drug, which leads to a faster decrease of the tumor volume, as presented in figure 4.4b.

From figure 4.4b we can affirm that the tumor volume resulting from the Bevacizumab drug administration reaches the maximum value and then it starts decreasing slowly whereas regarding the Atezolizumab drug administration, the tumor volume starts to decline almost immediately, showing only a slight initial rise that quickly becomes a descendent behavior, which relates to what was mentioned earlier regarding the higher effect produced by the Atezolizumab drug. Thus, the Atezolizumab drug allows a faster eradication of the tumor than the Bevacizumab drug administration.

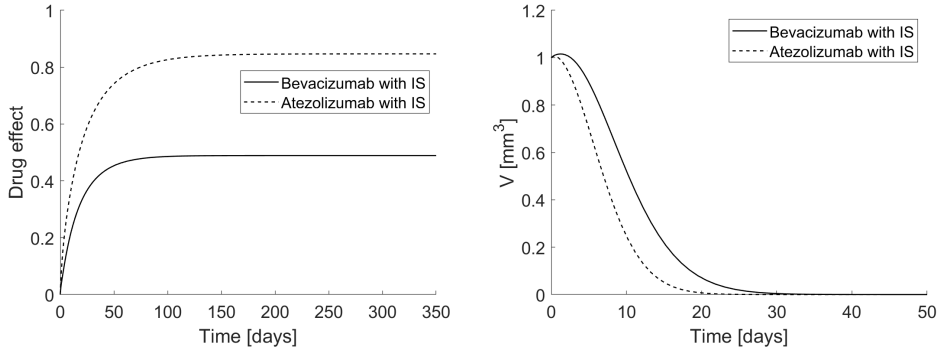
To check the influence of the aforementioned parameters, several simulations of the drug effect alongside the tumor volume evolution were performed for different values of α and u_{max} in order to demonstrate the functioning and influence of the PD model. That said, the curves representing the drug's effect and the tumor volume over time for $\alpha = 1$ and $u_{max} = 2$ are shown in figure 4.5.

By increasing the maximum effect value u_{max} twice, the effect of Bevacizumab drug (figure 4.5a) doubled and therefore the tumor volume growth (figure 4.5b) decreased in the same way, which means that producing the double of the effect leads to a halving of the tumor volume. Similarly, for the Atezolizumab drug represented with the dashed curves in figure 4.5, the output has the same behavior since only the initial drug parameters change, and it also eradicates the tumor by reducing it to zero over time, which means that increasing twice the Atezolizumab drug maximum effect leads to the destruction



(a) (b)

Figure 4.4: (a) Drug effect with IS over time for Bevacizumab and Atezolizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$, and (b) tumor volume evolution with IS over time for Bevacizumab and Atezolizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.



(a) (b)

Figure 4.5: (a) Drug effect with IS over time for Bevacizumab and Atezolizumab for $\alpha = 1$, $u_{max} = 2$, $a = 0.1$, $K = 5$ and $V_i = 1$, and (b) tumor volume evolution with IS over time for Bevacizumab and Atezolizumab for $\alpha = 1$, $u_{max} = 2$, $a = 0.1$, $K = 5$ and $V_i = 1$.

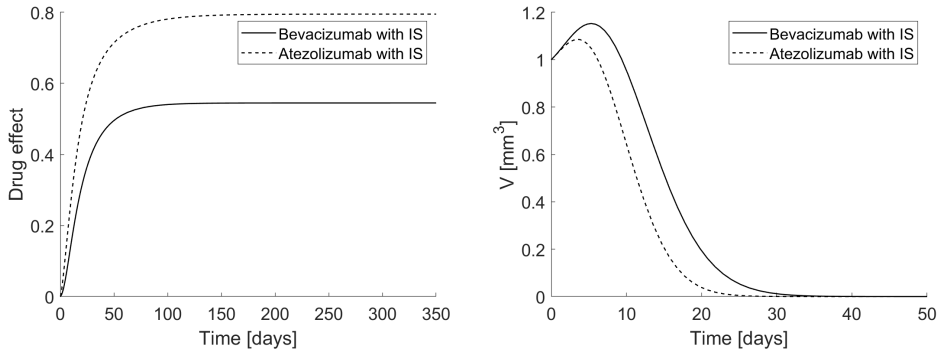
of the tumor while the drug concentration exists.

The same process was repeated for $\alpha = 2$ and $u_{max} = 1$ as described in figure 4.6.

By looking into figure 4.6a, initially both effect curves through time behave like a sigmoid since they start increasing slowly and then increase faster, less pronounced for the Atezolizumab drug, which leads to a faster reduction of the tumor volume regarding this drug as can be seen in figure 4.6b. From this figure, we can also see a slower increase of the tumor volume for the first 4 days approximately, and then the same descendant behavior occurs.

The last check performed concerns the tumor volume growth parameters namely the carrying capacity K and the intrinsic growth rate a, with fixed values for the PD model parameters $\alpha = 1$ and $u_{max} = 1$. The attempts consist of increasing the parameter K from 5 to 10 and decreasing the parameter a from 0.1 to 0.05, and then analyse the tumor growth volume increase or decline depending on these values regarding the steeper slope and of course for the two administered drugs, as presented in figure 4.7.

From the analysis of figure 4.7 it is important to mention that regardless of the changed parameter, the Atezolizumab drug always leads to a higher tumor volume decay. Despite that, it is also main to state that by increasing the carrying capacity parameter K, the maximum value of the tumor growth is exactly the same due to the presence of the IS. For instance, by comparing figure 4.7a to figure 4.7c we can affirm that a change in the intrinsic growth rate parameter a leads to a change of the slope of the curve since a decrease of this parameter implies a decrease on the same amount of the derivative of



(a) (b)

Figure 4.6: (a) Drug effect with IS over time for Bevacizumab and Atezolizumab for $\alpha = 2$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$, and (b) tumor volume evolution with IS over time for Bevacizumab and Atezolizumab for $\alpha = 2$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

the tumor volume value, which is mathematically correct since this parameter performs a multiplication in equation 4.1 and once it represents a negative slope which gets more intense by decreasing that parameter a, as expected. Therefore, when analyzing side by side figures 4.7a and 4.7c regarding the Atezolizumab dashed curve, a more complete conclusion must be taken since decreasing the a value a half, the tumor volume goes to zero over time, faster than the Bevacizumab influence allows to. Another important conclusion is related to the initial positive slope that enables the tumor to reach the maximum volume due to the fact that a lower value for a leads to a slower initial increment of the tumor volume growth. Lastly, by comparing figures 4.7b and 4.4b the same assumptions must me made once the initial slope is slower for a lower a value as mentioned before.

Another important aspect is related to the tumor growth volume after the drug effect disappears since there are cases where the tumor volume increases again to the initial value but more slowly, while there are other cases in which the tumor is completely extinct. In this case, since we want to analyse the system including the IS influence, this is not the case and so, due to the long lasting presence of the IS and so, if the tumor is eradicated by this amount of drug administered, then it remains null over long time. However, it is important to safeguard that this only happens due to full knowledge of the model and the parameters needed to eradicate the tumor, otherwise a new tumor enlargement may occur.

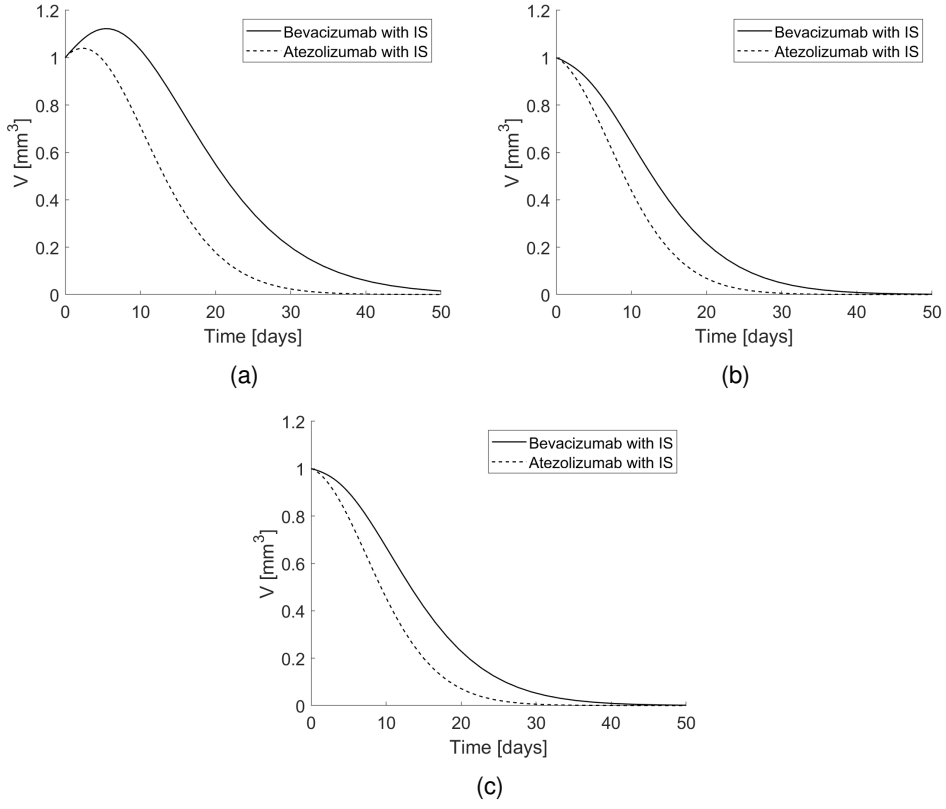


Figure 4.7: Tumor volume over time for (a) $\alpha = 1$, $u_{max} = 1$, $K = 10$ and $a = 0.1$, for (b) $\alpha = 1$, $u_{max} = 1$, $K = 5$ and $a = 0.05$, and for (c) $\alpha = 1$, $u_{max} = 1$, $K = 10$ and $a = 0.05$.

Chapter 5

Control of the Logistic Model

By analysing the behaviour of the Logistic Model in previous chapter, we aim to control this model in order to achieve better results regarding the tumor volume evolution. In order to illustrate the objective, the controller applied to the TGM model can be described through a schema as described in figure 5.1.

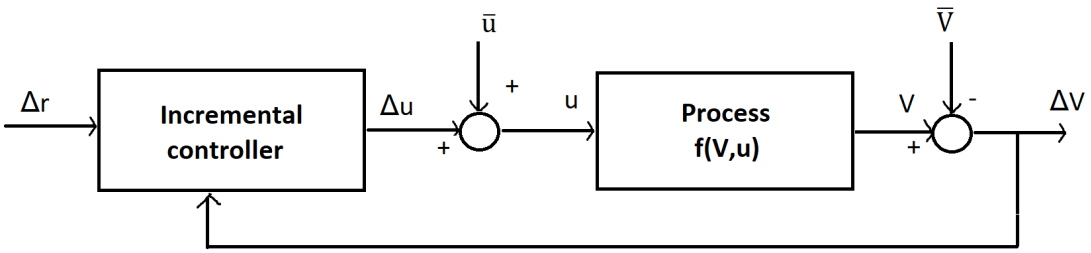


Figure 5.1: Logistic Model controller schema.

Therefore, the control law is given by:

$$u = -Ke, \tag{5.1}$$

where K is the feedback gain, u is the input and e is the error that represents the reference tracking. Thus, the error is given by:

$$e = \Delta r - \Delta V \tag{5.2}$$

It is also important to mention that in this case, the variable to control is the output V and the manipulated variable is the input u, since it is the only variable modified by the controller. In this case, there is also the addition of white noise to increase the excitation of the model and improve the performance regarding the tracking.

A parallel aspect that can be questioned is the fact that there is no reference to the discretization nor linearization of the PK and PD models, but the truth is that the PK model is linear, so it only needs to be discretized using a time constant, and the PD model is nothing more than a gain in series with the PK model.

On the other hand, since the TGM model is non-linear, we must discretize it by using *Euler's* method [45], and then linearize it for generic equilibrium points $\bar{V}$ and $\bar{u}$, meaning that in equation (4.1) we have $\beta = 1$ [46].

5.1 Discretization

By assigning $f(V(t), u(t))$ to equation (4.1), the discretized model of the logistic equation that we will consider from now on is:

$$V(k+1) = V(k) + hf(V(k), u(k)) = V(k) + h(aV(k)(1 - \frac{V(k)}{K}) - u(k)V(k)), \tag{5.3}$$

where h is the discretization term to convert continuous time t into discrete time k, and $a = 0.1$, $K = 5mm^3$.

5.2 Linearization

Once discretized, the model must be linearized around the equilibrium point $(\bar{V}, \bar{u})$ through the following expression:

$$\Delta V(k+1) = \frac{\partial f}{\partial V}(\bar{V}, \bar{u})\Delta V(k) + \frac{\partial f}{\partial u}(\bar{V}, \bar{u})\Delta u(k). \tag{5.4}$$

Finally, the incremental linearized discrete time model becomes:

$$\begin{cases} \Delta V(k+1) = A\Delta V(k) + b\Delta u(k) \\ \Delta y(k) = C\Delta V(k), \end{cases}, \tag{5.5}$$

where $C = 1$ while A and b are the model parameters. In this particular case of study we consider scalar values A and b given by:

$$\begin{cases} A = 1 + ha - \frac{2ah\bar{V}}{K} - h\bar{u} \\ b = -h\bar{V} \end{cases}. \tag{5.6}$$

In order to calculate the $\bar{u}$ as a function of $\bar{V}$, the discrete time equilibrium condition that must be verified is $V(k+1) = V(k)$. From here we can achieve that $\bar{u}$ is $a - a\bar{V}/K$ that can be replaced in the controller parameters system (5.6).

From equation (5.5), and also parameters A and b from equation (5.6), comes an incremental controller by applying the control law:

$$\Delta u(k) = -G\Delta V(k), \tag{5.7}$$

where G is the controller gain from the Linear-quadratic state-feedback regulator (LQR) for discrete-time state-space system.

5.3 Discretization with IS

Once again, we want to discretize the model using *Euler's* method [45]. Thus, with the addition of the IS state, we now have an extra state in the state-space TGM model, leading to the inclusion of a new term in the tumor volume equation:

$$\dot{V}(t) = aV(1 - \frac{V(t)}{K}) - \beta u(t)V(t) - \theta V(t)r(t), \tag{5.8}$$

where $a = 0.09$, $K = 10$, $\beta = 1$ and $\theta = 1$.

Therefore, by assigning $f(V(t), r(t), u(t))$ to equation (5.8) and $g(V(t), r(t))$ to equation (4.2), we then have:

$$V(k+1) = V(k) + hf(V(k), r(k), u(k)). \tag{5.9}$$

$$r(k+1) = r(k) + hg(V(k), r(k)). \tag{5.10}$$

5.4 Linearization with IS

After the discretization, we have to linearize it around the equilibrium points. Thus, since $\bar{V} = \frac{K}{10}$, where K is the carrying capacity of the tumor, by going backwards we get the equilibrium points of the input and the Immune System.

Then, the linearized system becomes:

$$\begin{bmatrix} \Delta V(k+1) \\ \Delta r(k+1) \end{bmatrix} = \begin{bmatrix} 1 + h(a - \frac{2a\bar{V}}{K} - \beta\bar{u} - \theta\bar{r}) & -h\theta\bar{V} \\ h(\alpha_2\bar{r} - 2\alpha_2\beta_2\bar{V}\bar{r}) & 1 + h(\alpha_2\bar{V} - \alpha_2\beta_2\bar{V}^2 - \delta_2) \end{bmatrix} \begin{bmatrix} \Delta V(k) \\ \Delta r(k) \end{bmatrix} + \begin{bmatrix} -h\beta\bar{V} \\ 0 \end{bmatrix} \Delta u(k), \tag{5.11}$$

where $\bar{V}$, $\bar{r}$ and $\bar{u}$ are the equilibrium points of V, r and u, respectively.

Chapter 6

Algorithms for RL

In this chapter, several algorithms for the application of Reinforcement Learning are presented, namely the RLS with Exponential Forgetting, Q-Learning, Velocity Algorithm and RLS with Directional Forgetting. The first one and the last one are RLS techniques characterized by distinct abilities to improve results. The Q-Learning is the main Reinforcement Learning algorithm presented in detail in this master dissertation. Lastly, the Velocity Algorithm allows to more precisely control higher order systems.

6.1 RLS with Exponential Forgetting

The Recursive Least-Squares (RLS) technique with Exponential Forgetting [47] allows the dynamic identification through recursively finding the coefficients that minimize a weighted linear least squares cost function concerning the input signals, recalculating the controller gain repeatedly in real time according to the input and output data of the system [34]. This process is known as the Dual Effect since we want to estimate the parameters as well as to control the system.

The forgetting factor λ that affects the equivalent memory, influences the number of previous samples considered, as follows:

$$t_{considered} = \frac{1}{1-\lambda}. \tag{6.1}$$

For a lower λ we have more fluctuation but at the same time it is more agile when it comes to follow variant patterns [48]. So, $\lambda = 0.95$ is already considered a low value since it only bases the estimation on 20 previous samples, whereas a $\lambda = 0.095$ considers 200 previous samples.

The model is then given by:

$$y(k+1) = \varphi'(k)\theta_0, \tag{6.2}$$

where φ is the regressor vector and θ_0 is a vector that contains the model parameters A and b of the real model (the patient model representation).

Therefore, the estimator equation is the following:

$$\hat{\theta}(k+1) = \hat{\theta}(k) + K(k+1)[y(k+1) - \hat{\theta}'(k)\varphi(k)], \tag{6.3}$$

where the regressor $\varphi(k)$ is given by $\begin{bmatrix} y(k) & u(k) \end{bmatrix}^T$.

Then, the next step is to compute the controller gain given by:

$$K(k+1) = \frac{P(k)\varphi(k)}{\lambda + \varphi'(k)P(k)\varphi(k)}. \tag{6.4}$$

Finally we must compute the solution matrix P of the *Ricatti* equation [32] as follows:

$$P(k+1) = \frac{[I - K(k+1)\varphi'(k)]P(k)}{\lambda}, \tag{6.5}$$

where in this case I is the identity matrix of dimension 2 and $P(k)$ is a diagonal matrix of dimension 2 as well, initialized at $1000I$ for this problem. After these computations at each iteration, we perform the LQR for discrete-time state-space system [49].

6.2 Q-Learning

Another algorithm that allows the learning of the value function, but now through Reinforcement Learning, is the Q-Learning where we do not know nor want to estimate the model parameters [32].

Thus, we consider a model as:

$$x_{k+1} = f(x_k) + g(x_k)u_k, \tag{6.6}$$

from which we achieve the optimal value and the optimal control as a function of the state:

$$V(x_k) = r(x_k, u_k) + \gamma V(x_{k+1}), \tag{6.7}$$

with $r(x_k, u_k)$ given by:

$$r(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k. \tag{6.8}$$

Since we want to determine the optimal control policy, the goal is to minimize the value function of equation (6.7) by performing its partial derivative considering the control policy u_k such that:

$$\frac{\partial}{\partial u_k}(x_k^T Q x_k + u_k^T R u_k + \gamma V(x_{k+1})) = 0. \tag{6.9}$$

From equation (6.9) we then get:

$$2u_k^T R + g(x_k)\frac{\partial V(x_{k+1})}{\partial x_{k+1}} = 0, \tag{6.10}$$

that depends on part of the dynamics $g(x_k)$, which assumed to be unknown.

The objective now is to follow a path where there is no previous knowledge about the system and the environment [37], and so all that matters is the estimation of the cost function to define an optimal control law that minimizes the reward estimate [39]. More precisely, the *Bellman Optimality* equation [32] can be expressed as:

$$V^*(x_k) = \min_u(Q^*(x_k, u)), \tag{6.11}$$

where the optimal Q function [32] is given by:

$$Q^*(x_k, u_k) = r(x_k, u_k) + \gamma V^*(x_{k+1}), \tag{6.12}$$

and the optimal control [32] is then:

$$h^*(x_k) = \arg\min_u (Q^*(x_k, u)). \tag{6.13}$$

Hereupon, to get the optimal control policy [32] we just have to derive the optimal Q function in order to the control policy u such that:

$$\frac{\partial}{\partial u}(Q^*(x_k, u)) = 0. \tag{6.14}$$

From equation (6.14) we now do not have a dependency on the system dynamics since the Q function is stored for all possible control actions carried out at each possible state concerning the pairs of values (x_k, u_k).

That said, the Q function must now be given by the *Bellman* equation [32] as:

$$Q_h(x_k, u_k) = r(x_k, u_k) + \gamma Q_h(x_{k+1}, h(x_{k+1})), \tag{6.15}$$

Furthermore, we would like to learn the Q function using RL and more specifically the Policy Iteration algorithm. Hence, for non-linear systems we might assume a parametric approximation of polynomials such that:

$$Q_h(x, u) = W^T \phi(x, u), \tag{6.16}$$

where $\phi(x, u)$ represents a set of functions. In this specific case, since the system is scalar, the polynomials are then:

$$\phi(x_k, u_k) = \begin{bmatrix} x_k^2 \\ u_k^2 \\ x_k u_k \end{bmatrix}, \tag{6.17}$$

and W is a vector with 3 entries representing the LS solution to the policy evaluation step of the Q-Learning policy iteration algorithm [32] given by:

$$W_{j+1}^T(\phi(x_k, u_k) + \gamma\phi(x_{k+1}, u_{k+1})) = r(x_k, h_j(x_k)). \tag{6.18}$$

The policy improvement step [32] is then:

$$h_{j+1}(x_k) = \arg\min_{h(.)} (W_{j+1}^T(\phi(x_k, u)). \tag{6.19}$$

In order to estimate the optimal policy, we must consider several iterations to allow its convergence in the sense that, during those iterations, we are only getting data to fill the buffer and so, the considered feedback gain must be close to the desired one. Later, we can apply several Least-Squares algorithms such as Batch Least-Squares, Recursive Least-Squares with exponential forgetting or even Recursive Least-Squares with directional forgetting, in order to get the estimates of W.

Therefore, it is relevant to present this algorithm step by step regarding its application to the TGM model, followed by a 2-states model with the TGM alongside an approximation of the PK model by a first-order system, and finally the complete model (PK+PD+TGM).

6.3 Velocity Algorithm

In the context of the adaptive control methods such as the Q-Learning, there was the need to change the line of thought regarding the reasoning of the control law. Therefore, this section presents an alternative algorithm that allows the control of multi-dimensional state-space systems such as the two systems mentioned last in the previous section.

Since the previously applied control law $u = -Kx$ only guarantees an accurate result for a system with one state, we need to introduce the Velocity algorithm.

This algorithm consists of an expansion of the state-space matrices A, b and C by adding a new state that is an integrator from the error between the output and the reference. In figure 6.1 there is a schema that allows to better understand it.

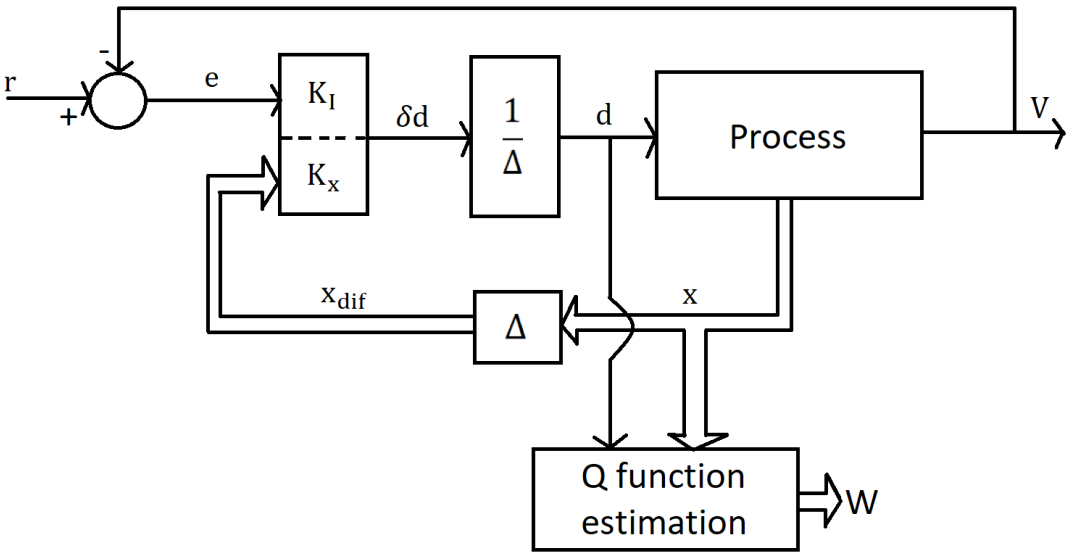


Figure 6.1: Velocity algorithm schema.

From figure 6.1, we can see that the difference between the reference and the process output is then multiplied by the integrator gain that is the added state mentioned before. Also, the state vector is differentiated and applied to the feedback gain of each state, giving rise to the input of the system that needs to be integrated before applied to the process.

This algorithm can be described through several difference equations:

$$\delta d(k) = -K_x x_{dif}(k) - K_I e(k) \tag{6.20}$$

$$d(k) = d(k-1) + \delta d(k) \tag{6.21}$$

$$e(k) = r(k) - V(k) \tag{6.22}$$

$$x_{dif}(k) = x(k) - x(k-1) \tag{6.23}$$

6.4 RLS with Directional Forgetting

Since there are several Recursive Least-Squares algorithms that allow to get the estimates for several algorithms, such as the Q-Learning in this particular case, it is important to mention that in this thesis, besides the normal RLS and the RLS with exponential forgetting, another RLS algorithm used was the RLS with directional forgetting.

It can be described by several equations such as the RLS with exponential forgetting, but now we have an extra equation which acts as a weight ensuring that the covariance matrix P is positive semi-definite if $P(0)$ is too:

$$\beta(k) = 1 - \lambda + \frac{1-\lambda}{\phi(k)'P(k-1)\phi(k)}. \tag{6.24}$$

$$K(k) = \frac{P(k-1)\phi(k)}{1 + \phi(k)'P(k-1)\phi(k)(1-\beta(k))}. \tag{6.25}$$

$$P(k) = [I - K(k)\phi(k)'(1-\beta(k))]P(k-1), \tag{6.26}$$

where I represents the identity matrix with the dimension of the amount of combinations for each state-space.

$$\hat{\theta}(k) = \hat{\theta}(k-1) + K(k)[r(k) - \phi(k)'\hat{\theta}(k-1)]. \tag{6.27}$$

This algorithm allowed to get better results regarding the peaks of the gains from the Q-Learning due to the zero transitions of the W values estimates.

Chapter 7

Results

The cancer evolution can be described through a child story, where the main character is the patient and the villain is the tumor. The last one is intended to adversely affect the patient, so it is necessary to adjust the parameters of the various algorithms in order to mitigate the consequences of this tumor. In this way, it becomes essential, implicitly, to represent the problem through a model, presented in stages whose evolution is dictated by the number of states involved.

Thus, being the tumor volume the output of the model and drug effect its input, the first step is a model with a single state, the tumor volume. Then, the pharmacokinetics state space model with two states alongside the drug effect and the tumor volume state is applied performing the complete model without the Is influence. Subsequently, and for the sake of simplification derived from the better achieved results, there is a focus on the plasma concentration c_2, leading to a PK model approximation by a first-order system, leaving a total system with two states, called the full model. Finally, as the IS is central to the fight against the tumor, this is also a state to be added to the full model, thus moving to a 3-state system.

Before moving on to the results presentation, it is important to mention that since the Q-Learning, for the models with more than a single state, was implemented using the velocity algorithm, which in addition to the states mentioned above is also constituted by the state of the integrator present in the figure 6.1, then the non-single state systems undergo the addition of this state to the process control law.

7.1 Non-linear TGM model

The first results concern the non-linear TGM model without considering the PK and PD models, meaning that we only have an output V and an input u, that represents the drug effect.

7.1.1 Batch Least-Squares

Hereupon, to get the estimates of W we used the Batch Least-Squares [32], [50] in the form:

$$\hat{W} = (\Phi^T \Phi)^{-1} \Phi^T \bar{y}, \tag{7.1}$$

where $\bar{y}$ stands for $r(x_k, h_j(x_k))$ and Φ is given by:

$$\Phi = \phi(x_k, u_k)^T + \gamma \phi(x_{k+1}, u_{k+1})^T. \tag{7.2}$$

Therefore, at each N iterations we compute equation (6.19) to obtain the optimal control policy applied to the system in the following iterations.

Since in this case we consider polynomials of the form of equation (6.17), from equation (6.19) the optimal control policy can then be achieved as follows:

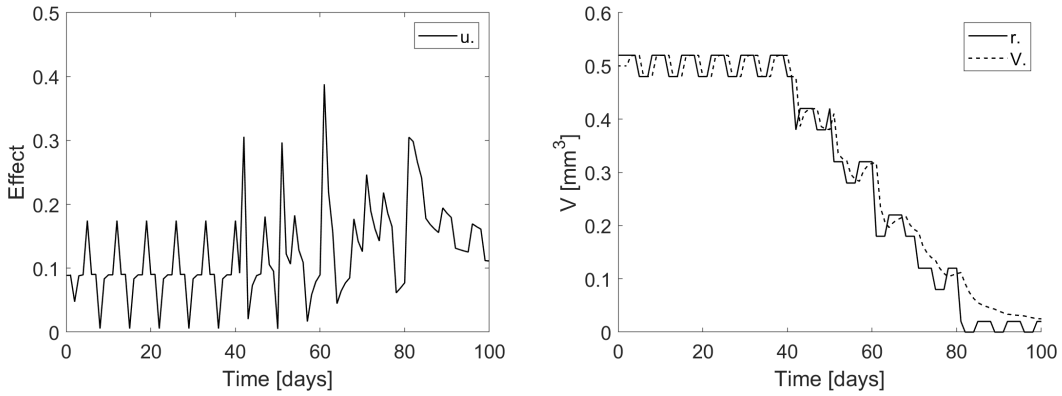
$$\nabla(W_1 x_k^2 + W_2 u_k^2 + W_3 x_k u_k) = 0 <=> 2W_2 u_k + W_3 x_k = 0, \tag{7.3}$$

leading to a feedback gain of:

$$u_k = -\frac{W_3}{2W_2} x_k \Rightarrow K = \frac{W_3}{2W_2}. \tag{7.4}$$

For an initial positive definite estimate $\hat{W}_0 = \begin{bmatrix} 0.1 & 1 & 0.2 \end{bmatrix}^T$, and a periodic square wave reference with amplitude 0.02, we want to estimate W for $N = 40$. Since we want the tumor volume to decrease until approximately zero over time, we must adjust the equilibrium points, namely the $\bar{V}$ and the corresponding $\bar{u}$.

Then, the non-linear model input signal and the output, alongside the reference, are presented in figure 7.1.



(a) (b)

Figure 7.1: (a) Input u, and (b) output V and reference r evolution for $A = 0.9910$ and $b = -0.4500$ for $R = 0.01$, for non-linear TGM model by using simple Batch LS for Q-Learning.

By looking at figure 7.1b, we can affirm that the system output V follows the reference almost all the time during the W estimation, and then when the reference starts to decrease to lead the tumor volume to zero, the follow-up becomes less accurate, but presents the overall desired behavior.

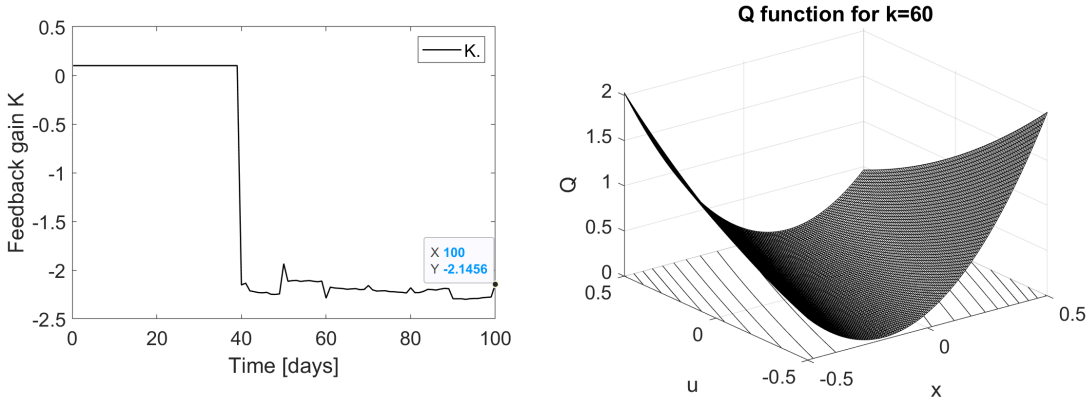
Then, the feedback gain K evolution and the Q function are presented in figure 7.2.

From figure 7.2a, for a gain of $K = -2.1030$, the attained gain is $K = -2.1456$, which is a close value to the desired one. In figure 7.2b, there is the Q function for a time instant $k = 60$, where we can see that there is a minimum that corresponds to the optimal control found.

Furthermore, by considering a reference resulting from the sum of three square waves with different amplitude, frequency and phase, we also want to derive the tumor volume to zero through time, and so the $\bar{V}$ as well as the $\bar{u}$ are adjusted. Therefore, the results for the tumor growth model are presented in figure 7.3.

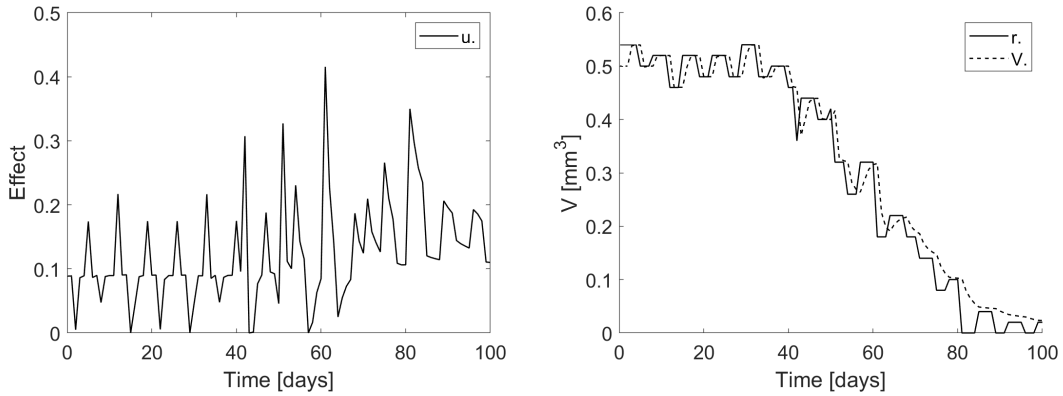
Thus, the gain K evolution through time k and the Q function are presented in figure 7.4.

By looking at figure 7.3b, we can affirm that, even though the reference has not a constant amplitude due to the square waves sum, the tumor volume follows the reference almost all the time, even better than the result presented in figure 7.1b, where the last days have a higher discrepancy regarding the



(a) (b)

Figure 7.2: (a) Feedback gain K evolution, and (b) Q function for $A = 0.9910$ and $b = -0.4500$ at day 60 for $R = 0.01$, for non-linear TGM model by using simple Batch LS for Q-Learning.



(a) (b)

Figure 7.3: (a) Input u, and (b) output V and reference r evolution for $A = 0.9910$ and $b = -0.4500$ for $R = 0.01$, for non-linear TGM model by using simple Batch LS for Q-Learning.

amplitude difference for the last three upper peaks presented.

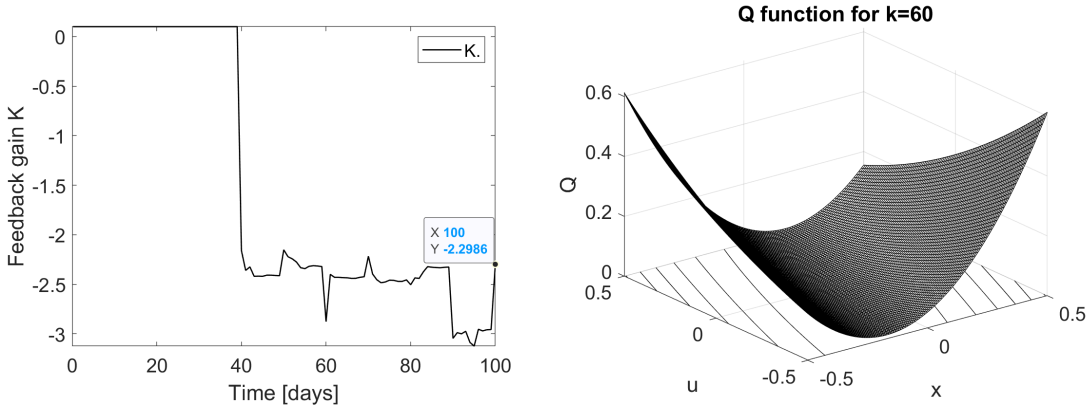
By comparing figures 7.2a and 7.4a, we might affirm that in the second one, the feedback gain K presents more notorious peaks when the equilibrium points are re-calculated, as mentioned before, than the first one. This happens due to the non-linearity of the system applied to a reference that changes slope more often than the single square wave reference. Therefore, for a desired value of $K = -2.1030$, the feedback gain attained through the three square waves reference was $K = -2.2986$.

In order to confirm the feedback gain evolution, we might compare the attained values for two random time instants, as presented in figure 7.5, with the theoretical ones as described in table 7.1.

Table 7.1: Theoretical gain values.

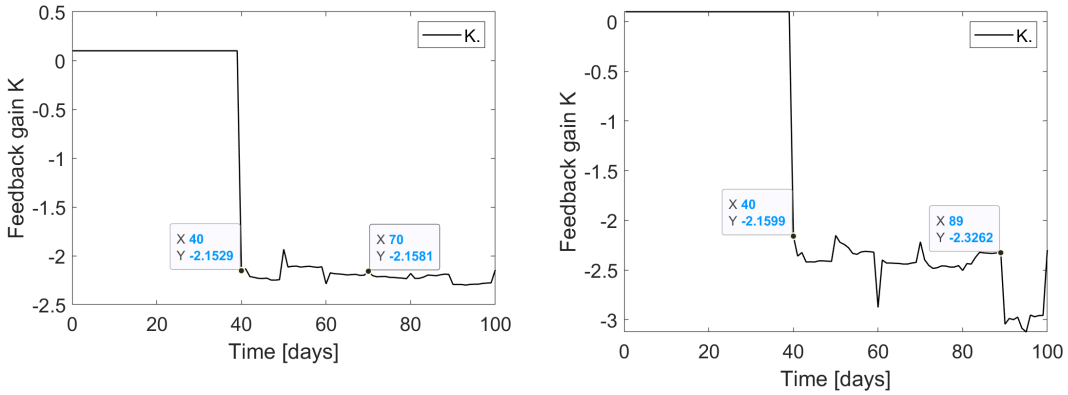
	Theoretical values	
One square wave	$K(40) = -2.1230$	$K(70) = -2.1986$
Three square waves	$K(40) = -2.1230$	$K(89) = -2.3285$

Therefore, by comparing figure 7.5 with the values from table (7.1), we can affirm that for both references, the accuracy is really high considering the small discrepancies between the theoretical and the attained gains for those two random instants.



(a) (b)

Figure 7.4: (a) Feedback gain K evolution, and (b) Q function for $A = 0.9910$ and $b = -0.4500$ at day 60 for $R = 0.01$, for non-linear TGM model by using simple Batch LS for Q-Learning.



(a) (b)

Figure 7.5: Experimental gain values for (a) one square wave reference at $k = 40$ and $k = 70$ and for (b) three square waves reference at $k = 40$ and $k = 89$, for non-linear TGM model by using simple Batch LS for Q-Learning.

Another relevant aspect to compare the accuracy of each reference when applied to the process is the quadratic error given by:

$$q_e = \sum_{k=1}^{n_k} (V(k) - r(k))^2, \tag{7.5}$$

where n_k is the number of samples acquired. These errors for both references are shown in table (7.2).

Table 7.2: TGM Quadratic error.

	Quadratic error
One square wave	0.0012
Three square waves	0.0013

From table (7.2), it pops up that both references present a close and low quadratic error, due to the good follow-up of the tumor volume.

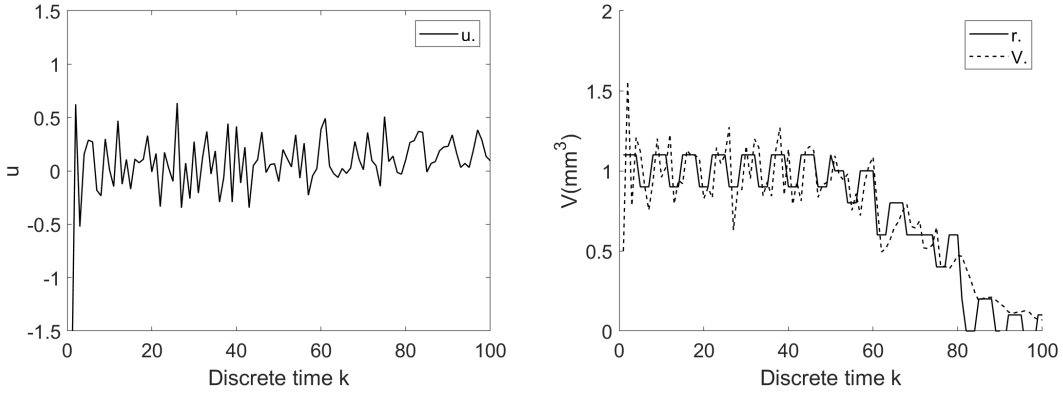
7.1.2 Batch Least-Squares with Prior

Another way to estimate the vector W is to use the Batch Least-Squares with Prior, which represents the computation of the new estimate considering the influence of the previous estimate:

$$\hat{W} = (\Phi^T \Phi + \alpha I)^{-1}(\Phi^T \bar{y} + \alpha \bar{W}), \tag{7.6}$$

where $\bar{W}$ represents the prior estimate of W and α measures the confidence one has in the prior.

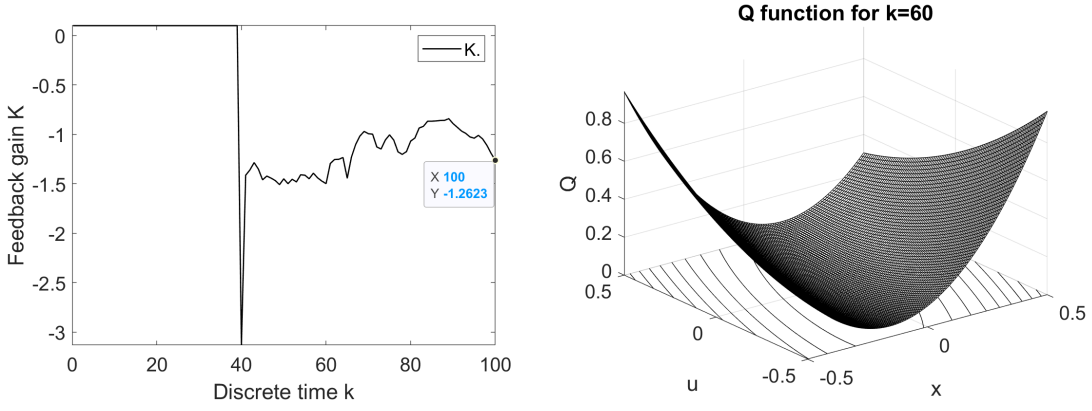
Similarly to what was done in the previous section, the first reference applied is a periodic square wave reference with amplitude 0.1, with adjusted equilibrium points as well. Thus, the results for the non-linear model with $A = 0.9910$ and $b = -0.4500$ for $N = 40$, regarding the input, output and reference signals, are represented in figure 7.6.



(a) (b)

Figure 7.6: (a) Input u, and (b) output ΔV and reference r evolution for $A = 0.9910$ and $b = -0.4500$ for $R = 0.2$, for non-linear model by using Batch LS with prior for Q-Learning.

By looking at figure 7.6b, we can affirm that the system output ΔV follows the reference almost all the time, but there are some peaks that arise from the sudden variation of the reference when turning from high to low values, and vice-versa, which when comparing to figure 7.1b are less pronounced due to the influence of the previous estimation at each iteration, meaning that the estimates are more accurate. Therefore, the feedback gain K evolution and the Q function are presented in figure 7.7.



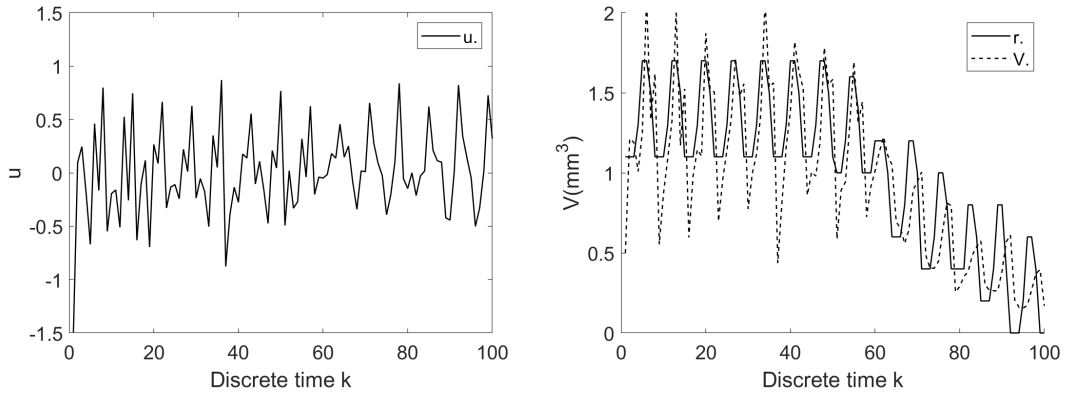
(a) (b)

Figure 7.7: (a) Feedback gain K evolution, and (b) Q function for $A = 0.9910$ and $b = -0.4500$ at day 60 for $R = 0.2$, for non-linear model by using Batch LS with prior for Q-Learning.

From figure 7.7a, we can affirm that in a general way, the feedback gain evolution through time is

more accurate than for the Q-Learning with simple Batch LS, since the attained gain $K = -1.2623$ is closer to the desired one $K = -1.3616$, than the corresponding gain using simple Batch LS, as shown in figure 7.2a. Regarding the Q function for time instant $k = 60$, we still acquire a global minimum, as stated before as well.

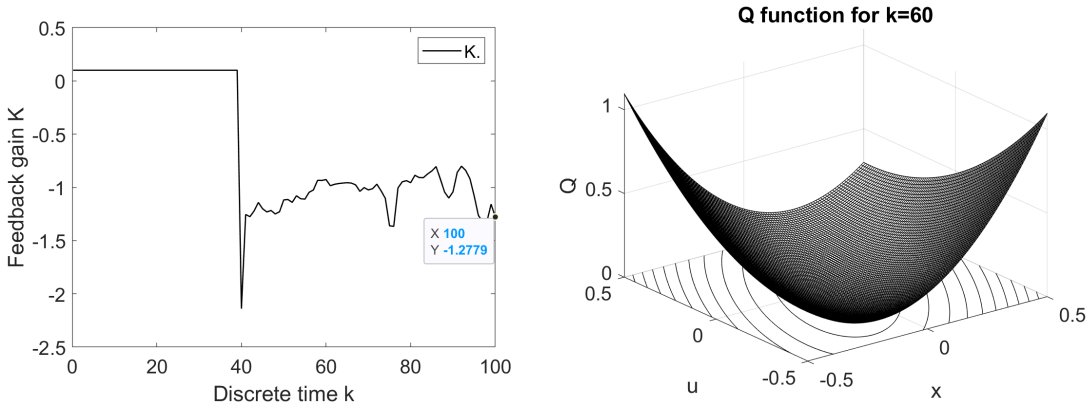
Nevertheless, by considering a reference resulting from the sum of three square waves with different amplitude, frequency and phase, and the adjusted equilibrium points, the results for the tumor growth model are presented in figure 7.8.



(a) (b)

Figure 7.8: (a) Input u, and (b) output ΔV and reference r evolution for $A = 0.9910$ and $b = -0.4500$ for $R = 0.2$, for non-linear model by using Batch LS with prior for Q-Learning.

Therefore, the gain K evolution and the Q function are presented in figure 7.9.



(a) (b)

Figure 7.9: (a) Feedback gain K evolution, and (b) Q function for $A = 0.9910$ and $b = -0.4500$ at day 60 for $R = 0.2$, for non-linear model by using Batch LS with prior for Q-Learning.

Comparing figures 7.3b and 7.8b, we can state that the results are approximately the same, meaning that the differences regarding the feedback gains for the same reference but using now Batch LS with prior derive from the distinct input signal randomly created using *Matlab*.

By crossing figures 7.4a and 7.9a, we might affirm that in the second one, for a desired value of $K = -1.3616$, the feedback gain K attains $K = -1.2779$, which is closer than the desired one, and the variations around the final value are smoother.

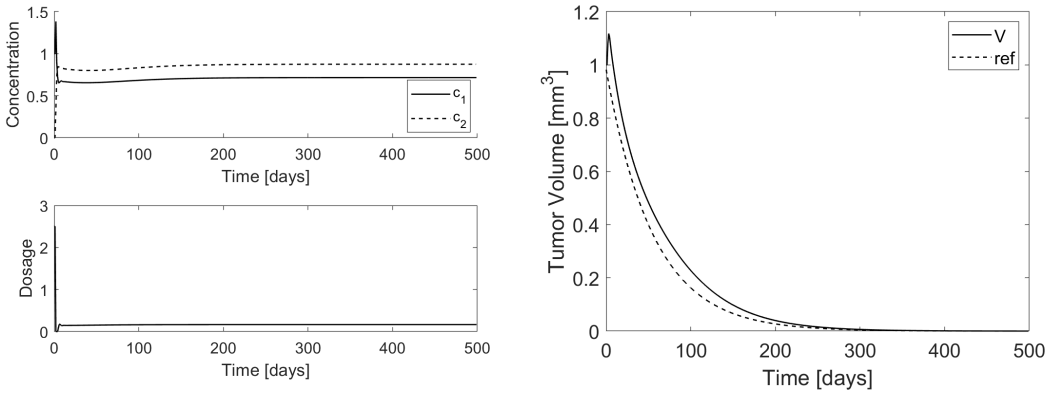
By comparing figures 7.7a and 7.9a, we can conclude that, since the feedback gain K attained is approximately the same using both references with the Batch LS with prior, then the only difference concerns the minimum value reached by each feedback gain evolution when the first W estimate is

acquired. In this case, for the reference with the three square waves, the minimum is approximately -2, and for the reference composed by the single square wave, the minimum is approximately -3. This means that with the three square waves, the feedback gain needs to adjust less than with the single square wave, to achieve a value close to the desired K.

7.2 Full PK model, PD and TGM without IS

The next stage of the Velocity algorithm and Q-Learning application is the full system with the two state-space Pharmacokinetic model, the gain from the Pharmacodynamic block and the non-linear Tumor Growth system.

In order to achieve this goal, we have to be careful regarding the variables and the parameters combinations, as well as the initial conditions of the distinct states, otherwise the amount of errors can be disastrous. That said, the first step to overcome is the application of the Velocity algorithm to the full process, with the fixed state gains from the LQ, without the Q-Learning algorithm. The results regarding this test are presented in figure 7.10.



(a) (b)

Figure 7.10: (a) PK concentration c_1 and c_2, dosage, and (b) tumor volume evolution comparing to reference, for $R = 0.1$ through time, using velocity algorithm as control law.

From figure 7.10, we can affirm that the tumor volume follows the reference most of the time despite the initial elevation due to the carrying capacity of this non-linear model. Therefore, the Velocity algorithm applied to this full process is accurate.

7.3 Approximated PK, PD and TGM model without IS

The next step is then to apply the Q-Learning to the full process, which means that the gains are no longer known from the LQ but from the Q-Learning estimates of the W parameters. Unfortunately, since the addition of another state to the PK model lead to worse results, from now on the process, considered as full model, is the serie of the PK model approximated by a first order system, multiplied by the PD gain, with the non-linear TGM model.

Then, considering an approximated first-order model for the PK block given by:

$$c_2(k+1) = -0.05c_2(k) + d(k), \tag{7.7}$$

and the same non-linear TGM model mentioned before, we move on to the application of the Q-Learning to a two state-space system.

Considering the two state-space system composed by the state mentioned in equation (7.7) and the tumor volume state, now the polynomial vector for the ϕ increases the dimension from 3 to 6:

$$\phi(x_1, x_2, u) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ u^2 \\ x_1 x_2 \\ x_1 u \\ x_2 u \end{bmatrix}, \tag{7.8}$$

where x_1 is the tumor volume V, x_2 is the approximated plasma concentration c_2 and u is the process dosage input d.

Similarly to what was done in equation (7.3), we now have a 6-dimension W estimates so:

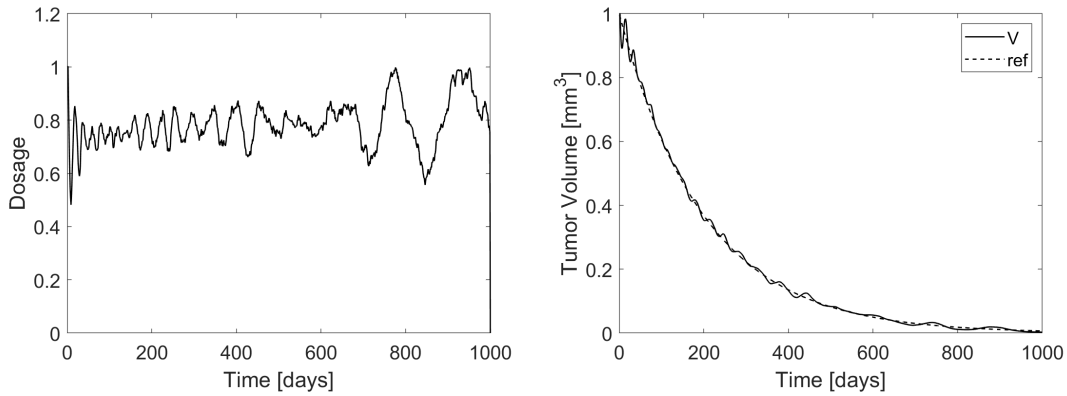
$$\nabla(W_1 x_1^2 + W_2 x_2^2 + W_3 u^2 W_4 x_1 x_2 + W_5 x_1 u + W_6 x_2 u = 0, \tag{7.9}$$

leading to the following control law:

$$u = -\frac{1}{2W_3} \begin{bmatrix} W_5 & W_6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{7.10}$$

Therefore, by using the Recursive Least-Squares with directional forgetting presented before, we want to lead the tumor volume to zero by following an implied reference. In order to achieve this goal, after several attempts with distinct input parameters for the Q-Learning, as well as distinct RLS algorithms applied, the best results come from fixing the integrator gain, defined by the tracking error, in the LQ gain during all moments, together with using the known states gains of the LQ for brief moments and then switching to the gains of Q-Learning arising from W estimates.

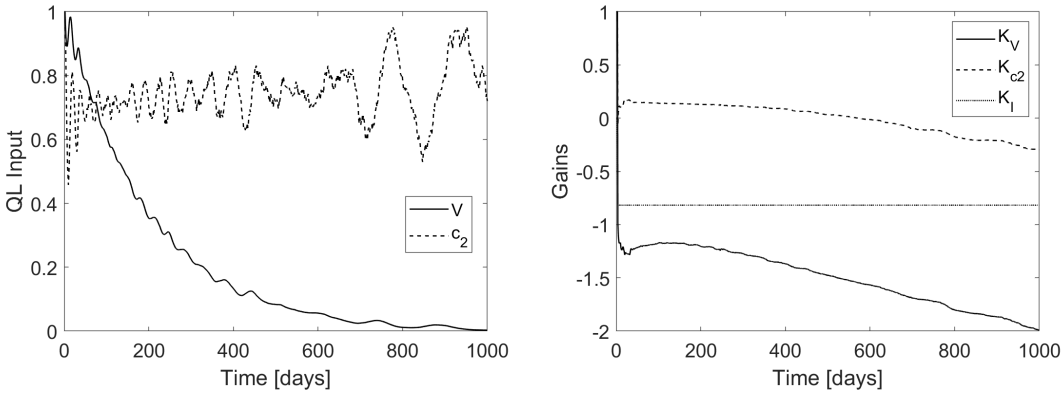
Thus, the results from the Velocity algorithm and Q-Learning applied to the process with a negative exponential reference are presented in figures 7.11, 7.12 and 7.13.



(a) (b)

Figure 7.11: (a) Dosage, and (b) tumor volume evolution comparing to reference through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning.

From figure 7.11a we can affirm that there is a small oscillation at low frequencies due to the peak value that the tumor volume reaches at the beginning, as shown by figure 7.11b, which leads to a higher



(a) (b)

Figure 7.12: (a) Q-Learning input signals, and (b) gains evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning.

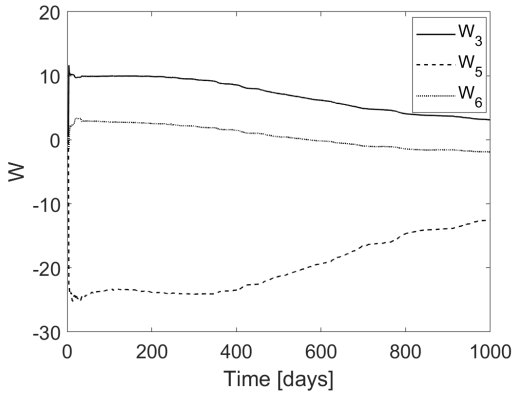


Figure 7.13: W estimates evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning.

difference between this signal and the reference, producing a bigger discrepancy of the initial dosage. Despite that, the remaining samples show a correct tracking of the reference.

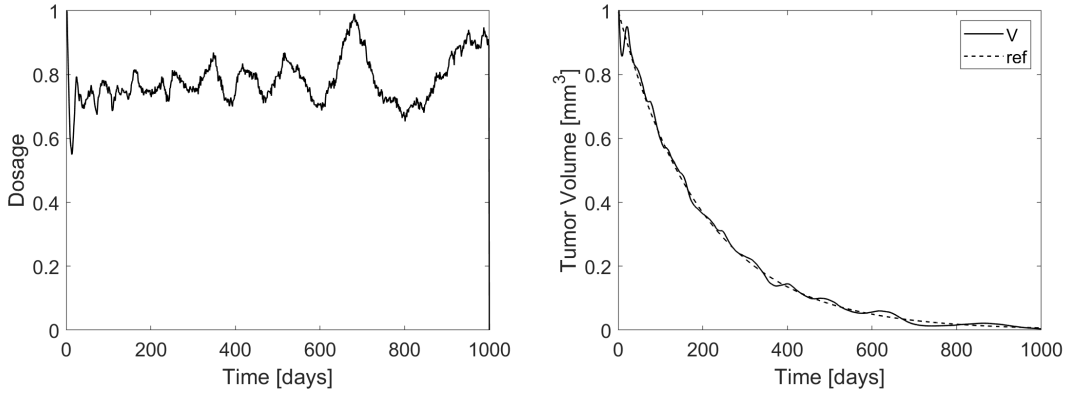
On figure 7.12a we can see the input state signals that contribute for the Q-Learning computation regarding the estimates of W, presented in figure 7.13, that lead to the gains of the states, since the integrator gain is fixed as mentioned before, as shown in figure 7.12b.

Since these algorithms involve the combination of several parameters, it is necessary to use the quadratic error given by (7.5) to evaluate the performance of each combination. In order to avoid a massive number of figures, the only results presented are the best combinations found considering both the metric results and the tumor volume behavior with regards the reference tracking. Consequently, the comparison is given by table (7.3).

Table 7.3: Quadratic error for the approximated PK and non-linear TGM model.

	Quadratic error
$R = 0.1$, $\lambda = 0.995$, $\gamma = 0.95$	0.0022
$R = 1$, $\lambda = 0.995$, $\gamma = 0.95$	0.0025
$R = 2$, $\lambda = 0.995$, $\gamma = 0.95$	0.0033
$R = 0.1$, $\lambda = 0.995$, $\gamma = 0.995$	0.0024

From table (7.3) we are able to verify that for the first presented combination, the one discussed earlier, the quadratic error is smaller than for the remaining ones. Although the third combination is the one with the most accentuated quadratic error, it is important to show the results, regarding the tumor volume evolution and the dosage input, in order to refute this combination, as shown in figure 7.14.



(a) (b)

Figure 7.14: (a) Dosage, and (b) tumor volume evolution comparing to reference through time, for $R = 2$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law, and RLS with directional forgetting for the Q-Learning.

As we can see from figure 7.14a, the dosage for $R = 2$ presents slower value changes, meaning that it takes more time to adapt to the values from the feedback gains, leading to a slower behavior regarding the tumor volume as well, as can be seen in figure 7.14b since each piece of a sudden change takes longer to respond and to track the reference at a higher performance.

7.4 Full model with IS

By using the previously presented Velocity Algorithm as control law, as well as the Directional RLS with exponential forgetting, we now want to control a three state-space system by using the LQ feedback gains for few instants and then the Q-Learning feedback gains that arise from the W estimates.

Therefore, by including the influence of the Immune System, one expects the dosage to reduce once there is no need to apply as much dosage of the drug since the Immune System also contributes to the reduction of tumor.

Hereupon, for a three state-space model the dimension increases from 6 to 10 combinations, meaning that the Q-Learning polynomial vector ϕ is now given by:

$$\phi(x_1, x_2, x_3, u) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \\ u^2 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_3 \\ x_1 u \\ x_2 u \\ x_3 u \end{bmatrix}, \tag{7.11}$$

where x_1 is the tumor volume V, x_2 is the Immune System influence r, x_3 is the approximated plasma concentration c_2 and u is the process dosage input d.

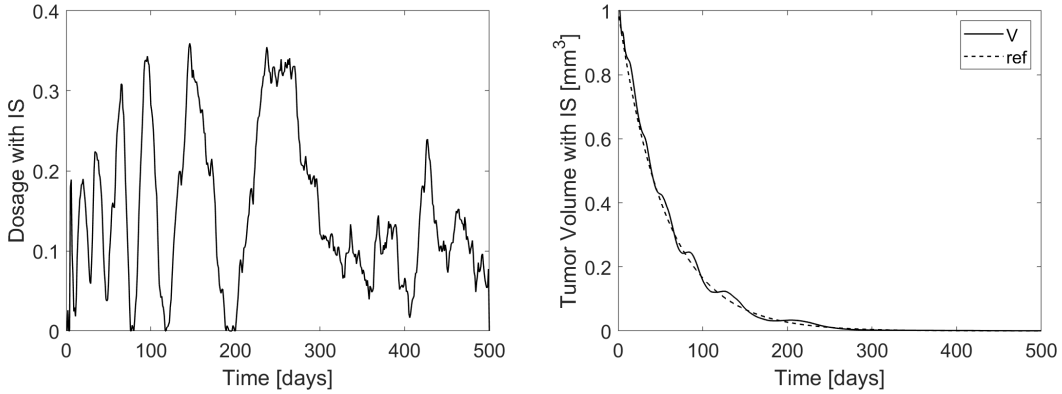
Similarly to what was done in equation (7.9), we now have a 10-dimension W estimates so:

$$\nabla(W_1x_1^2+W_2x_2^2+W_3x_3^2+W_4u^2+W_5x_1x_2+W_6x_1x_3+W_7x_2x_3+W_8x_1u+W_9x_2u+W_{10}x_3u)=0, \quad (7.12)$$

leading to the control law:

$$u=-\frac{1}{2W_4}\begin{bmatrix}W_8 & W_9 & W_{10}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}. \quad (7.13)$$

That said, the results with the Immune System influence are presented in figures 7.15, 7.16, 7.17 and 7.18.



(a) (b)

Figure 7.15: (a) Dosage, and (b) tumor volume evolution comparing to reference through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

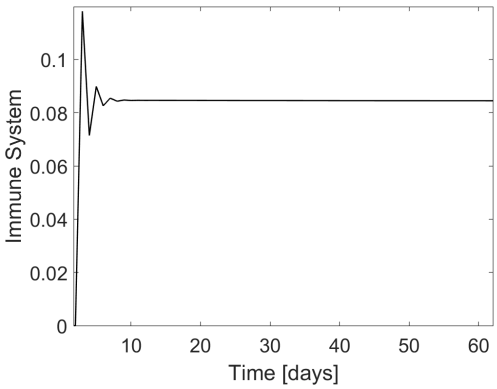
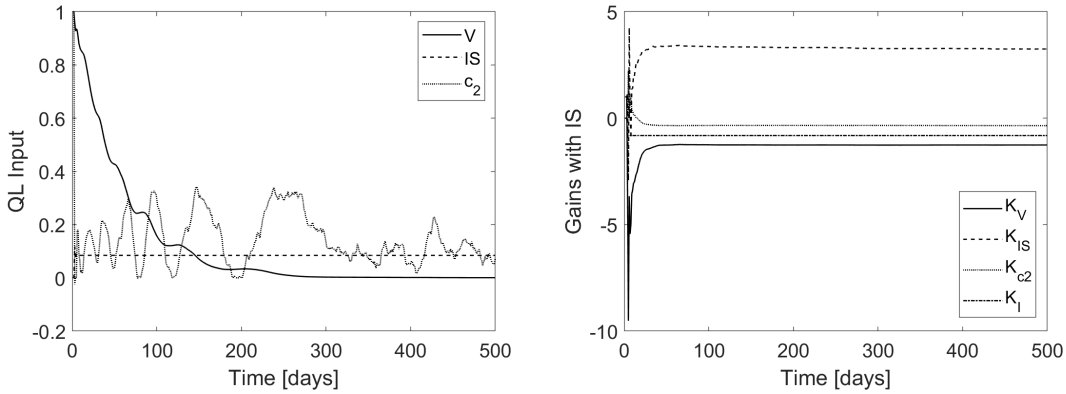


Figure 7.16: Immune system effect evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

By comparing figures 7.11b and 7.15b, we can affirm that with the IS influence, there is no oscillatory behaviour of the tumor volume at low frequencies. Instead, the tumor volume starts by decreasing and then stabilizing at the error tracking, following the reference as close as possible, with a quadratic error



(a) (b)

Figure 7.17: (a) Q-Learning input signals, and (b) gains evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

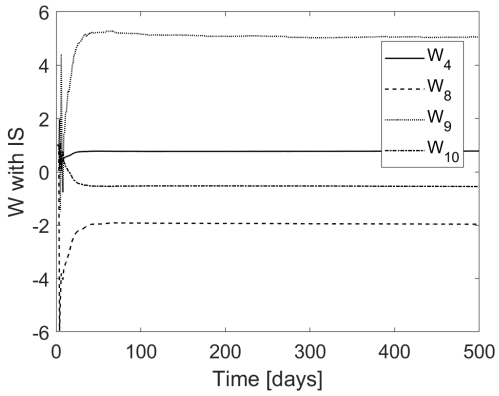


Figure 7.18: W estimates evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

in the order of 10^{-4}, which when comparing to the previously presented quadratic error in table (7.3) in the first row, is way smaller.

From figure 7.16, after a zoom in to exactly observe the IS effect, we can conclude that it takes approximately 3 to 5 days to become constant, which corresponds to the time it takes to start fighting the tumor, leading to its decrease through time.

Thus, since now we have 3 states and the integrator state, from figure 7.17a we can see the 3 Q-Learning input states that allow the W estimation, as presented in figure 7.18, leading to the feedback gain of each state, as in figure 7.17b.

Through figures 7.13 and 7.18, we might affirm that with the IS influence, the W estimates are constant from a specific instant whereas before they kept increasing or decreasing, meaning that with the IS addition, we managed to be more accurate and precise in the tumor eradication process.

Chapter 8

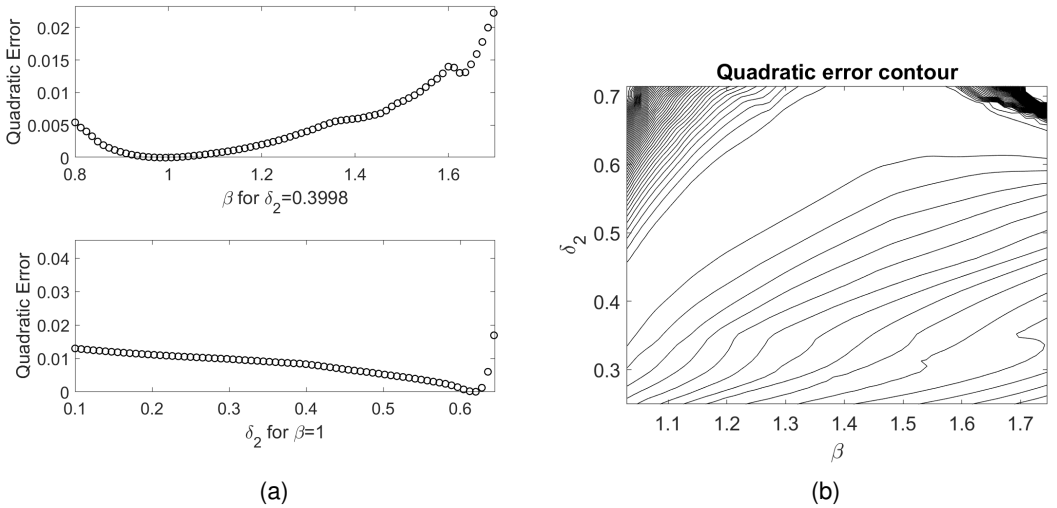
Q-Learning algorithm evaluation

Since the Q-Learning algorithm implemented in this master dissertation is related to Policy Iteration that requires stabilizing feedback gains to produce good results, it is important to study the robustness and stability of this algorithm, as well as the quadratic error and the decay of the gains, through the evolution of several parameters.

8.1 Robustness and stability

Regarding the proof of robustness of the Q-Learning algorithm, we chose 2 parameters of the system to check the model's operating basin depending on the evolution of those parameters selected for testing. It is important to stress out that the optimal values used for this work were the ones presented in table (4.1).

Then, after several trials we concluded that the working ranges of the parameters β and δ_2 were $[0.8; 1.72]$ and $[0.1; 0.65]$, respectively. For better understanding, we set a grid with those parameters' ranges and performed the quadratic error for each combination, leading to figures 8.1a and 8.1b.



(a) (b)

Figure 8.1: (a) β and δ_2 individual evolution for $\delta_2 = 0.3998$ and $\beta = 1$, respectively, and (b) β and δ_2 contour through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

Another way to analyse the behavior of these parameters' is the surface representation that translates

the optimal range through the upturned concavity, presented in figure 8.2.

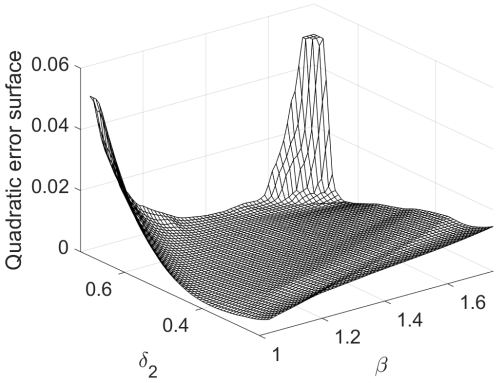


Figure 8.2: β and δ_2 surface through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

At the same time, to study the stability of the algorithm we changed the way we were doing the update of the feedback gains. More specifically, we performed the computation of the feedback gains using the Q-Learning with RLS with directional forgetting as before, but now we changed the update time from each instant to an update period between 1 and 300. Thus, considering the IS influence on the full system, it is important to show the results considering both the use of LQ initialization for a few instants, meaning that the first gain values come from the Linear-quadratic (LQ) regulator, and the use of the Q-Learning estimates from the beginning, as described in figure 8.3.

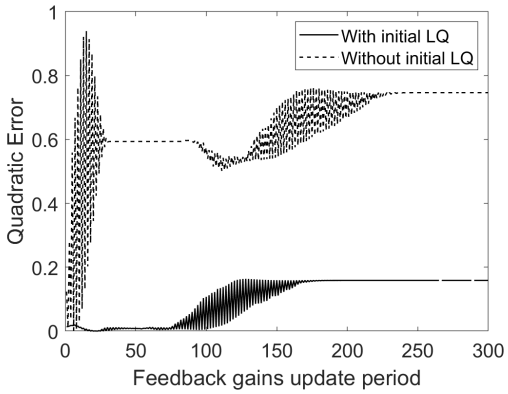


Figure 8.3: Tumor volume quadratic error evolution through increasing feedback gain update period, for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

From figure 8.3, we can conclude that the tumor volume quadratic error for the dashed curve has an initial increase followed by a constant period and a new increase until a constant value again. Therefore, without the initial LQ feedback gains, the ”U” shape curve is more notorious than the one representing the test with initial LQ feedback gains. On the other hand, the full curve presents a constant value followed by an increase until another stabilization. Since we want to achieve results with lower quadratic error, the overall test using initial LQ feedback gains is better than the one without initial LQ values.

8.2 LQ feedback gains through LQ parameter R evolution

By changing the R parameter of the Linear-quadratic (LQ) regulator from 0 to 100, there is an interesting behavior of the feedback gains that must be presented, as can be seen in figure 8.4.

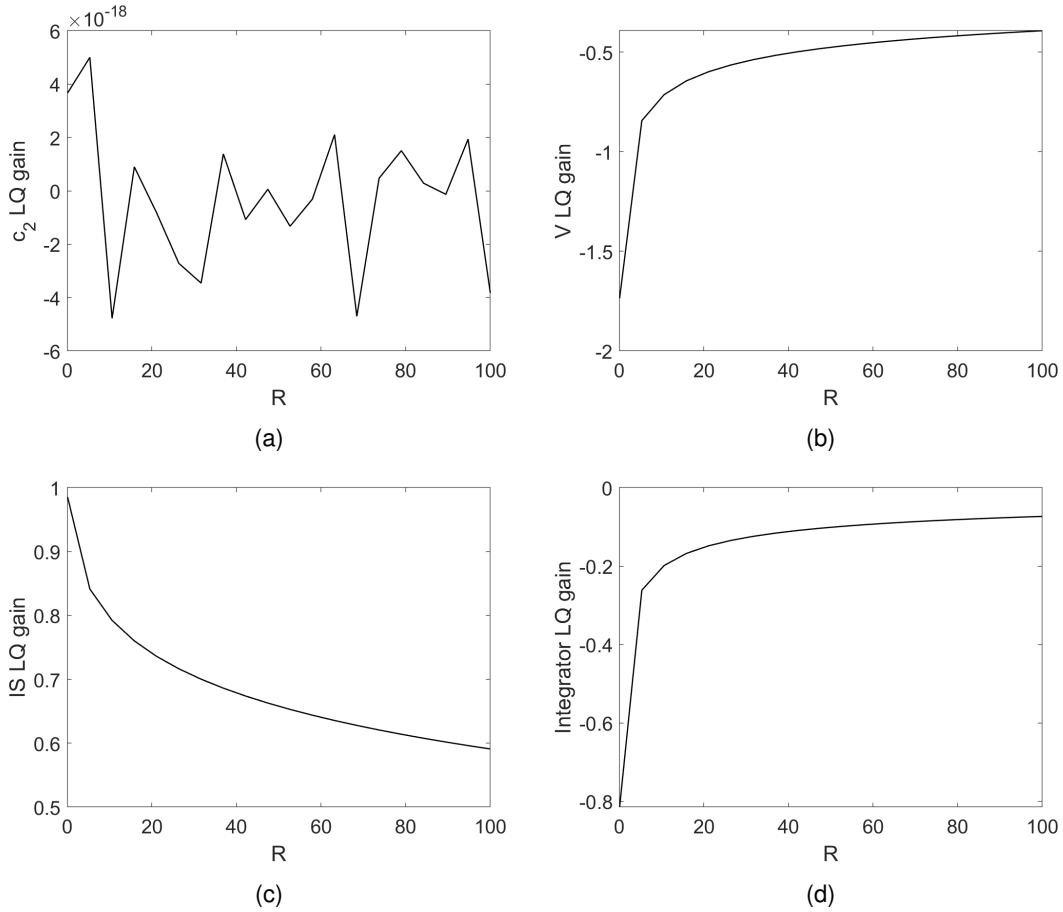


Figure 8.4: (a) c_2 gain, (b) V gain, (c) r gain, and (d) integrator gain through Linear-quadratic (LQ) regulator R evolution, for $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

From figure 8.4 we can affirm that some gains increase whereas others decrease, but there is a common factor: all of them approach zero. This means that, with the increase of the R parameter that determines the feedback gains used for the further LQ iterations required for the Q-Learning gains estimates calculations, the LQ gains decrease in absolute value, which stands for an expected and accurate conclusion. More specifically, from figure 8.4a we can state that the c_2 gain evolution presents an oscillatory behavior whose average value tends to zero. The remaining figures 8.4b, 8.4c and 8.4d reflect a clearer idea of what was mentioned above in relation to the trend towards zero.

8.3 Q-Learning gains through increase of LQ initial iterations

Considering a distinct amount of initial LQ iterations, it is interesting to check the Q-Learning gains estimates evolution, as described in figure 8.5.

By looking at figure 8.5, we can check that the overall gains present the same behavior in the sense that all the curves without any initial LQ iterations tend to a constant value different from all the others. In other words, while the 2 and 6 initial LQ iterations curves tend to approximated values, for each state

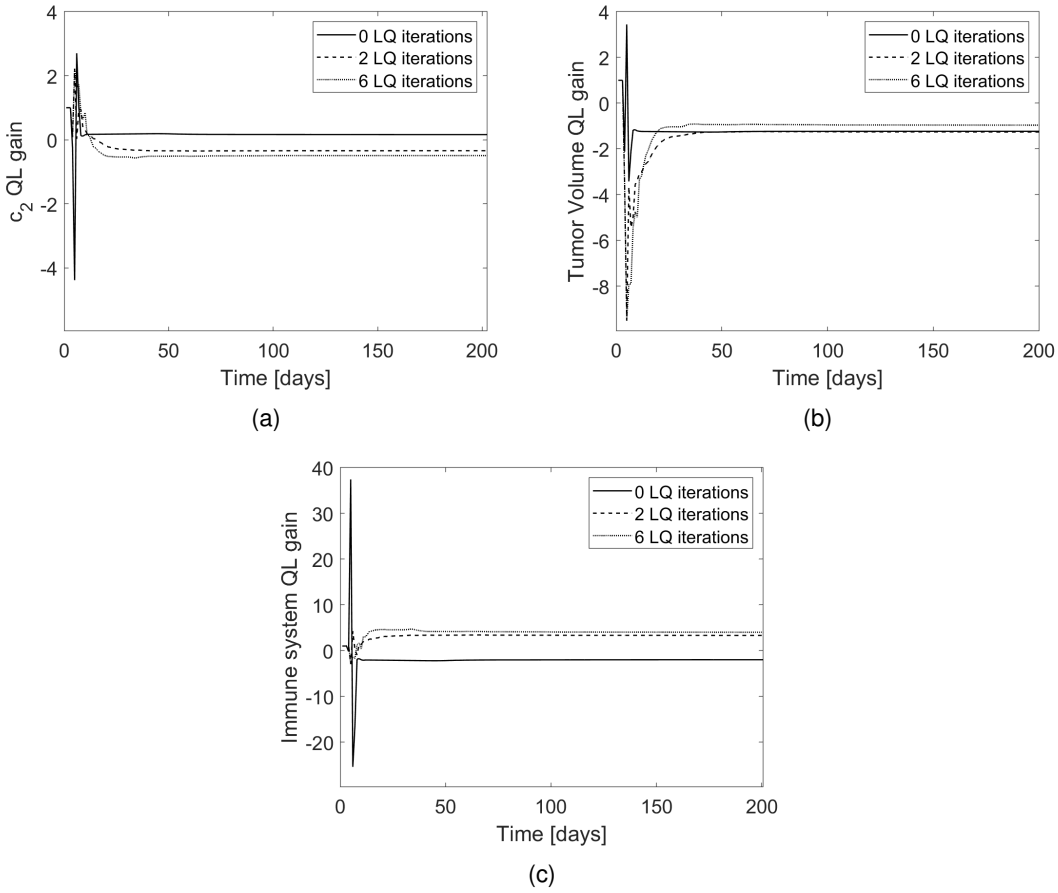


Figure 8.5: (a) c_2 gain, (b) V gain, and (c) r gain decay, through Linear-quadratic (LQ) regulator initial iterations evolution, for $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

gain, the result without any initial LQ iteration tends to a different value, even though a constant one as well.

In this concrete case, the comparison between the LQ gains and the ones estimated through the Q-Learning calculations is illustrated in table (8.1).

Table 8.1: State gains convergence: theoretical, without initial LQ and with 2 initial LQ.

Gains	LQ	QL without initial LQ	QL with 2 initial LQ
K_{c_2} gain	$3E^{-18}$	0.1514	-0.3463
K_V gain	-1.7357	-1.2251	-1.2624
K_r gain	0.9853	-1.8952	3.2715

Due to the number of states of the system alongside the integrator gain influence, which we considered as a fixed gain for the Velocity algorithm, the gains presented in table (8.1) regarding the third and fourth columns are not quite close to the theoretical ones in the second column.

The aforementioned conclusions go accordingly to the fact that, regarding the Q-Learning algorithm with Policy Iteration, which requires a start from a point with stabilizable gains produces, the produced results for the Q-Learning estimates calculations are more precise.

8.4 RLS λ parameter effect on results

In order to check the sensibility of the system to the Recursive Least-Squares with directional forgetting, we used the λ parameter, that stands for the forgetting factor, to explore the tumor volume and dosage quadratic errors evolution through these parameter changes.

Therefore, for an evolution of λ from 0.5 to 1, the results are presented in figure 8.6.

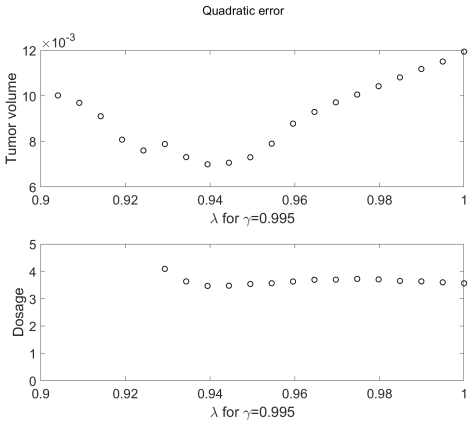


Figure 8.6: Volume and dosage through RLS λ evolution for $R = 0.1$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

From figure 8.6 we can analyse the quadratic error of the tumor volume, which is calculated through the difference between the reference and this state V, leading to small errors between 0 and 0.02. On the other hand, regarding the dosage quadratic error, since there is no reference level for this calculations, the attained errors are higher than the previously mentioned ones.

We can affirm that there is a ”U” shape curve with a minimum that contains the currently used value for λ for the entirely work, which traduces an accurate choice.

8.5 Effect of discount factor γ

Since the Q-Learning algorithm also relies on some parameters, another interesting test is to check the influence of the γ evolution, that is a discount factor for the optimal control calculations, in the tumor volume and dosage quadratic errors.

Similarly to what was done in the previous section, for an evolution of γ from 0.5 to 1, the results achieved are presented in figure 8.7.

From figure 8.7, we can see that once again the quadratic error for the dosage was higher than the quadratic error for the tumor volume. Regarding the quadratic error, we can see that it increases until 0.02, that traduces a non-static result, and so, the used value for this parameter for the entire work is $\gamma = 0.995$, which belongs to the range of the increasing quadratic error, so the influenced region of this parameter.

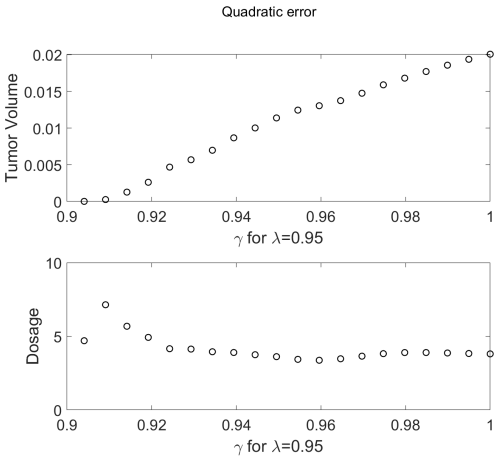


Figure 8.7: Volume and dosage through QL γ evolution for $R = 0.1$ and $\lambda = 0.995$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

Chapter 9

Conclusion

Regarding the drug administration, through all the results achieved in section 3 of chapter 4, as expected different drugs produce distinct effects and so, different changes in the tumor volume. With respect to the tested drugs, we can conclude that the Atezolizumab drug produces a higher effect due to the lasting presence of this drug concentration through time. Similarly, with regard to the tumor volume, the Atezolizumab drug allows a higher decreasing of the tumor through time due to the same stated reason. It is also important to state that the IS influence allows to almost instantly decrease the tumor volume whereas without the IS influence, there is an initial elevation and then the decreasing behavior. Finally, it is also important to mention that these conclusions persist regardless the variations of the cancer models parameters.

Concerning the control of the logistic model, through the discretization and corresponding linearization in chapter 5, with and without the IS, as well as the definition of a controller, we were able to apply several algorithms for the Reinforcement Learning application, as described in chapter 6. The application of these algorithms, and the performance of several simulations, allowed to conclude that between the Recursive Least-Squares with Exponential Forgetting and the Recursive Least-Squares with Directional Forgetting, the second one produced better results concerning the gains evolution for the Q-Learning, since it assumed that the transitions of the W estimates by zero should not be translated into peaks. In turn, the velocity algorithm proved to be better than the initial incremental controller when applied to higher order systems, regarding the reference tracking. Lastly, the Q-Learning was more challenging in the sense that initially it was not known which variables should be involved in calculating the W estimates, since there was an additional state that was the integrator from the velocity algorithm, and in the end the integrator state turned out to be unnecessary to calculate these estimates, but essential as a weight for the error between the reference and the tumor volume.

In chapter 7 the results were presented and through the several comparisons and analysis between all the phases of the work, from the simplest to the most complex model, we can conclude that the best results concerning the full model are the ones with RLS with directional forgetting and the Q-Learning application with Velocity algorithm, since these results represent the full model with IS influence and the tumor volume follows the reference almost all the time, with a low error and an accurate tumor eradication process. This conclusion is most related to the fact that using the Batch Least-Squares, with and without Prior, applied to the non-linear TGM model, through several simulations there was always the need to use external disturbances to provide enough excitation to the model alongside the change of the reference from an initial square wave to a sum of square waves, meaning that a repeated behavior lead to the loss of tracking. Then, with the cancer models (PK and PD) involved, and latter with the IS influence, the results improved significantly, allowing the application of a decreasing exponential reference that more closely resembles the true behavior of the tumor volume evolution through time.

Then, since the application of the Q-Learning as a technique of the RL is the essential point of this work, an analysis of this algorithm was performed in chapter 8, leading to some conclusions regarding the parameters influence and working ranges. Regarding the robustness and stability of this algorithm, we can conclude that the algorithm is robust since the quadratic error evolution for the chosen ranges of β and δ_2 is really small and also due to the presence of a minimum in the surface plot created from the grid between the ranges of these two parameters. On the other hand, concerning the stability of the algorithm we can conclude that it is stable once it converges to constant values after several iterations, with or without initial LQ iterations, which is the same as saying, with or without knowing the gains of the linearized model. Then, from the results with regards the influence of the LQ parameter R, we can state that the gain of each state involved in the controller, even the integrator one, tends to approach zero as the value of R parameter increases. As mentioned before, the LQ initial iterations represent the knowledge of the gains of the linearized model, so we can confirm that since the applied algorithm is the Q-Learning with Policy Iteration, which requires a stabilizable start, then the use of more initial LQ iterations helped to get more precise results regarding the W estimates of the Q-Learning. The forgetting factor λ of the RLS is also a determinant parameter in respect of the tumor volume and dosage quadratic errors, so from the results we can conclude that, despite the overall zero quadratic errors, when it begins to get reasonable values, there is a minimum in the "U" shape curve that is the best value to use, and that is the one used in this work. Finally, from the analysis of the effect of the discount factor γ, similarly to what happened with the λ influence, the results are also almost all null, except when starting a sudden ascent, where the value of γ used during this work is included.

Nonetheless, there is also further work to be developed, namely the inclusion of the Angiogenesis subsystem to the model, which presupposes the addition of a new state, and the application of this work, with the necessary adjustments, to neural networks instead of the combination of polynomials that was used throughout this entire dissertation.

Despite that, since cancer represents a constant struggle, there are always new studies to improve the work already developed in order to make it less fallible in the face of all the generalization of patients.

Bibliography

[1] National Cancer Institute, "What causes cancer?," *National Cancer Institute*, 2018.

[2] National Cancer Institute, "What is cancer?," *National Cancer Institute*, 2020.

[3] J. Lennox, "Darwinism," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, fall 2019 ed., 2019.

[4] National Cancer Institute, "Is every tumor cancer?," *National Cancer Institute*, 2018.

[5] R. Padmanabhana, N. Meskina, W. M. Haddad, "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment," *Mathematical Biosciences*, vol. 293, no. 10, pp. 11–20, 2017.

[6] Wikipedia, "Cancer — Wikipedia, the free encyclopedia," 2020.

[7] Wikipédia, "Hippocrates — Wikipedia, the free encyclopedia," 2020.

[8] Wikipedia contributors, "Campbell De Morgan — Wikipedia, the free encyclopedia," 2020.

[9] Worldwide Cancer Research, "Twelve cancer research breakthroughs we made last year," *Worldwide Cancer Research*, 2021.

[10] Wikipedia contributors, "Treatment of cancer — Wikipedia, the free encyclopedia," 2020.

[11] V. T. DeVita and E. Chu, "A history of cancer chemotherapy," *Cancer Research*, vol. 68, no. 21, pp. 8643–8653, 2008.

[12] F. Worden, P. Anthony J. Perissinotti, and P. Bernard L. Marini, *Cancer Pharmacology and Pharmacotherapy Review: Study Guide for Oncology Boards and MOC Exams*. Springer Publishing Company, 2016.

[13] A. Petrovski, J. McCall, "Multi-objective Optimisation of Cancer Chemotherapy Using Evolutionary Algorithms," *Lecture Notes in Computer Science*, vol. 1993, 2001.

[14] J. J. H. A. Yin, D. Moes, "A review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors," vol. 8, no. 10, pp. 720–737, 2019.

[15] Tim Newman, "How the immune system works," *Medical News Today*, 2018.

[16] Immune Deficiency Foundation, "The Immune System and Primary Immunodeficiency," *Immune Deficiency Foundation*, 2020.

[17] Khan Academy, "Innate Immunity," *Khan Academy*, 2016.

[18] Khan Academy, "Adaptive Immunity," *Khan Academy*, 2015.

[19] Health Diary, “Essential role of platelets in the immune system discovered,” *Health Diary*, 2018.

[20] C. Stockmann, D. Schadendorf, R. Klose, I. Helfrich, “The Impact of the Immune System on Tumor: Angiogenesis and Vascular Remodeling,” *Frontiers in oncology*, vol. 4, p. 69, 2014.

[21] Institute for Quality and Efficiency in Health Care, “What are the organs of the immune system?,” *National Center for Biotechnology Information*, 2020.

[22] Wikipedia contributors, “Thymus — Wikipedia, the free encyclopedia,” 2021.

[23] Wikipedia contributors, “Liver — Wikipedia, the free encyclopedia,” 2021.

[24] Wikipedia contributors, “Bone marrow — Wikipedia, the free encyclopedia,” 2021.

[25] Wikipedia contributors, “Tonsil — Wikipedia, the free encyclopedia,” 2021.

[26] Wikipedia contributors, “Lymph node — Wikipedia, the free encyclopedia,” 2021.

[27] Wikipedia contributors, “Spleen — Wikipedia, the free encyclopedia,” 2021.

[28] Wikipedia contributors, “Blood — Wikipedia, the free encyclopedia,” 2021.

[29] Laura Elizabeth Lansdowne, “Angiogenesis in Cancer,” *Technology Networks*, 2019.

[30] Springer Nature, “Angiogenesis,” *Nature Portfolio*, 2021.

[31] IBM Cloud Education, “Machine Learning,” *IBM*, 2020.

[32] F. Lewis and D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control,” *Circuits and Systems Magazine, IEEE*, vol. 9, pp. 32 – 50, 01 2009.

[33] IEEE Circuits and Systems Magazine, “Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control,” *IEEE Xplore*, vol. 9, pp. 32–50, 2009.

[34] P. C. Young, *Recursive Least Squares Estimation*, pp. 29–46. Springer Berlin Heidelberg, 2011.

[35] R. Padmanabhan, N. Meskin, W. M. Haddad, “9 - reinforcement learning-based control of drug dosing with applications to anesthesia and cancer therapy,” in *Control Applications for Biomedical Engineering Systems* (A. T. Azar, ed.), pp. 251–297, Academic Press, 2020.

[36] Y. Zhao, M. Kosorok, D. Zeng, “Reinforcement learning design for cancer clinical trials,” *Statistics in medicine*, vol. 28, 2009.

[37] André Violante, “Simple Reinforcement Learning: Q-learning,” 2019.

[38] Y. Zhao, M. Kosorok, D. Zeng, M. Socinski, “Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer,” *Biometrics*, vol. 67, no. 4, 2011.

[39] Chathurangi Shyalika, “A Beginners Guide to Q-Learning,” 2019.

[40] ADL, “An introduction to Q-Learning: reinforcement learning,” *freeCodeCamp*, 2018.

[41] D. Poole, A. Mackworth, “Artificial Intelligence: Foundations of Computational Agents,” *Artinfo*, 2017.

[42] A. Hassani, M. Naghihi, “Reinforcement Learning Based Control of Tumor Growth with Chemotherapy,” *2010 International Coriference on System Science and Engineering*, 2010.

[43] C. Yu, J. Liu, “Reinforcement Learning in Healthcare: A Survey,” *arXiv*, 2020.

[44] J. M. Gallo, “Pharmacokinetics: Model structure and transport systems*,” *Clinical Research and Regulatory Affairs*, vol. 18, no. 3, pp. 235–266, 2001.

[45] Wikipedia contributors, “Euler method — Wikipedia, the free encyclopedia,” 2020.

[46] Wikipedia contributors, “Linear approximation — Wikipedia, the free encyclopedia,” 2020.

[47] J. Jiang and Y. Zhang, “A revisit to block and recursive least squares for parameter estimation,” *Computers Electrical Engineering*, vol. 30, no. 5, pp. 403 – 416, 2004.

[48] C. Paleologu, J. Benesty, and S. Ciochina, “A robust variable forgetting factor recursive least-squares algorithm for system identification,” *Signal Processing Letters, IEEE*, vol. 15, pp. 597 – 600, 02 2008.

[49] Mathworks, “Linear-quadratic (lq) state-feedback regulator for discrete-time state-space system,” 2020.

[50] Mathematics StackExchange, “Linear Algebra Batch Least Squares,” 2019.