Adaptive Control for Cancer Therapy based on Reinforcement Learning

Maria Inês de Mendonça Ferreira
Instituto Superior Técnico, Lisboa, Portugal

November 2021

In this work, having adopted an adaptive control strategy to fight cancer, 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.
By assuming that we do not know neither want to estimate the tumor volume model, we might use the Q-Learning technique to compute the optimal control policy that allows the minimization of the function without any knowledge on the tumor growth model that is influenced by the Immune System.

Keywords— Adaptive Control, Cancer, Reinforcement Learning, Q-Learning, Immune System.

I. Introduction

The main goal is to develop an adaptive optimal strategy for cancer therapy schedule based on state models, considering the Immune System response, which will be obtained through Reinforcement Learning alongside Adaptive Control.

A. Motivation

There are several external causes (namely from the environment) and internal causes (such as hormones, immunological conditions and genetic mutations), called factors that start the onset of cancer. This group of diseases is characterized by DNA mutations resulting in disorders that may or may not occur naturally. 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 distinct types of cancer [1].

Cancer is a malignant neoplasm, meaning that it is neither organized nor even aware of the limits. They do invade neighboring tissues and develop metastases – spread of the tumor over other parts of the body [2]. 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 a safe way.

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

B. Objectives

This master dissertation aims to present a preliminary feasibility study of the development of a state model of cancer evolution, including drug administration, with and without considering the Immune System influence, and the application of Reinforcement Learning techniques to the mathematical model of the tumor growth, in order to design personalised therapy schedules for cancer treatment.

C. Contributions

In this master dissertation, the analysis of the tumor growth mathematical model is proposed alongside with the Immune System response where, since there is a persistent failure with regards to the generalized cancer therapy, this study is improved by considering several Reinforcement Learning techniques in order to achieve an optimized treatment strategy based on Adaptive Control.

II. State of Art

This chapter provides a review over the literature presented to date regarding cancer treatment, cancer models, and several Reinforcement Learning algorithms that allow the control of the tumor growth model, as well as to optimize and personalize therapy.

A. 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 schedule cancer treatment to ensure effective and safe treatment [3]. According to the *World Health Organization*, it causes about 12.5% of all the deaths around the world [4].

There has been a notorious evolution with regards cancer treatment and also cancer knowledge itself. Therefore, in chronological order from the 18th century to the 20th century, there were several breakthroughs:

- In the 18th century, English surgeon Campbell De Morgan [5] speculated that cancer started locally and then spread over the body. This was only possible due to the first usage of the microscope;
- In the 19th century, doctors realized that a cleaner environment would contribute to less infections and to a safer recovery after surgery. By the end of this century, the radiation was discovered leading to the first cancer treatment without surgery involved; [4]
- 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 [4]. Nonetheless, since World War II cancer treatments have been improved although there are still some needed improvements, namely the non-existence of treatments for all types of cancer, and the non-generalization of the existing treatments.

B. Logistic Growth

The cancer dynamics must take into account the tumor growth model as well as the patient Immune System reaction to the tumor growth [3], since there is the need to choose the optimal treatment that ensures the minimization of the cancerous cells, without endangering the patient [6]. Up to date, there are several Tumor Growth models [7], as can be seen in table (1).

Table 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	$\frac{dV}{dt} = k_g * V * \ln(\frac{V_{max}}{V})$

B.1 Immune System

The Immune System is related to the capacity of the organism to react to external infections, and so it is spread over the body and has the ability to detect foreign tissue by comparing it to self-tissue, our own tissue. It also detects dead cells and expels them from our organism [8].

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. There are two major types of cells: B-cells and T-cells. The B-cells are related to the bone marrow whereas the T-cells concern the thymus. The development of all cells of the Immune System begins in the bone marrow with a blood-forming stem cell that, due to its capacity to generate the entire Immune System, is the most important cell in a cell transplant, specially to recover damaged tissues. Therefore, stem cells represent the basis of the entire Immune System [9].

C. Reinforcement Learning

One might be able to achieve the desired goal through methods such as Linear Quadratic Regulator [10] or Recursive Least-Squares approaches [11] 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 previous knowledge of the model.

That said, the main role of the Reinforcement Learning is to avoid the use of models. Therefore, the environment must be set and then the optimal control schedule must be chosen to fulfill the main goal [12]. To better illustrate the structure of the RL, a schema is presented in figure 1.

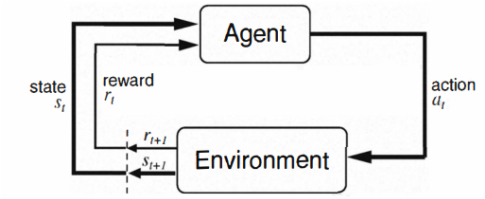

Figure 1: Reinforcement Learning structure.

From figure 1, we can see that there is 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. Then, 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 optimal control [13].

C.1 Q-Learning background

A Reinforcement Learning algorithm called Q-Learning aims the performance estimation of the controller without knowing the model and without

wanting to estimate it [14]. 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 dissertation. This method regulates the drug infusion to each patient even if there is no accurate knowledge about the patient system, non-linearity, cancer models, drug administration and its side effects, leading to improvements on the clinical goals [15]. There are also studies concerning learning optimal regimens from patient data created through clinical reinforcement trials [16].

Furthermore, there are several articles and papers with regards the application of the Q-Learning to control the tumor growth for example through chemotherapy [17], and to develop effective treatment regimes for individual patients [18].

Q-Learning technique stands for "Learning with quality" so, regarding the drug administration for cancer treatment based on Reinforcement Learning, the goal is then to reduce the tumor as much as possible, considering the tracking of a reference. This tracking is possible through the application of a controller based on how useful an action is going to be in gaining some future reward.

III. Cancer Models

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

Therefore, the cancer models can be described through a blocks diagram to better explain the input and output variables involved, as presented in figure 2.

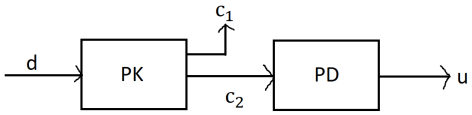

Figure 2: Block diagram of the cancer models.

A. Pharmacokinetics

Pharmacokinetics (PK) models are created to illustrate the transformations that a drug undergoes in an organism and the rules that determine its fate. Thus, it performs a main role regarding drug absorption, distribution, metabolism, and excretion [19].

In this case, 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, as described below:

$$\frac{dc}{dt} = Ac + bT. \tag{1}$$

Since we mentioned the concentration of each compartment, it is important to state that in this work a two compartment model such as the one used in Biomedical Engineering was considered at first sight, as described in figure 3.

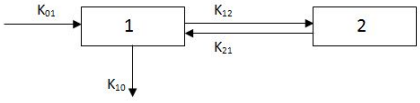

Figure 3: Quaternary model with two compartments.

In figure 3 we can see that 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 flows have a compensation task since when one has low amount of drug dosage, then the other compartment gives away that needed amount reaching an equilibrium.

Therefore, the state-space model presented in equation (1) can now be replaced by:

$$\begin{bmatrix} \frac{dc_1}{dt} \\ \frac{dc_2}{dt} \end{bmatrix} = \begin{bmatrix} \frac{-k_{12}-k_{10}}{V_1} & \frac{k_{21}}{V_1} \\ \frac{k_{12}}{V_2} & \frac{-k_{21}}{V_2} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{V_1} \\ 0 \end{bmatrix} u, \tag{2}$$

where the input signal u is 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 $y = c_2$.

B. Pharmacodynamics

Pharmacodynamics (PD) concerns the "What a drug does to a body". Then, we need to relate the plasma concentration c_2 with the drug effect denominated by u which translates into the *Hill* equation described as follows:

$$u(t) = u_{max} \frac{c_2^{\alpha}(t)}{c_{50}^{\alpha} + c_2^{\alpha}(t)}, \tag{3}$$

where the input is 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 effect of half the drug concentration.

IV. Tumor Growth Model

The Tumor Growth Model (TGM) evaluates the status of the tumor regarding its size and effect. 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 the tumor fight.

A. Logistic Growth Model

The model that represents the tumor growth used in this work is the Logistic Growth model, that can be mathematically represented by:

$$\frac{dV}{dt} = aV(1 - \frac{V}{K}) - \beta uV, \tag{4}$$

that 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). In this case, the parameters considered were $a = 0.1$, $\beta = 1$, $K = 5mm^3$ and $V(0) = 1mm^3$.

B. Logistic Growth Subsystem

The interactions between the IS and the cancer are extremely complex and traduced by a nonlinear dynamic such as:

$$\dot{r} = \alpha_2(1 - \beta_2 V)Vr + \gamma_2 - \delta_2 r, \tag{5}$$

where r is the immunocompetent cell density related to the triggered immune cells during the reaction, 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 (2).

Table 2: Immune System parameters.

Parameters	Value
α_2	0.00484
β_2	0.00264
γ_2	0.1181
δ_2	0.3998

C. 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, and to analyze the behavior of the tumor volume evolution through time with and without the IS influence, for each drug considered in this work it is important to present some drug administration results.

The parameters considered for each drug can be presented in table (3).

The results 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, is presented in figures 4 and 5.

From both figures, we can affirm that the curves without IS reach a higher maximum than the dashed curves standing for the tumor volume evolution over time with IS influence. By comparing figures 4 and 5, we can conclude that the Atezolizumab drug produces a higher reduction of the cancer tumor volume since the peak reached by both curves in figure 5 is smaller than the one with the Bevacizumab drug, leading then to a faster eradication of the tumor.

Table 3: 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

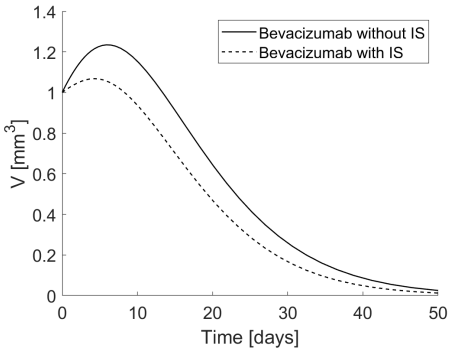

Figure 4: Tumor volume evolution with and without IS over time for Bevacizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

Another important aspect concern the IS response and the PK model behavior regarding both concentrations c_1 and c_2, as presented in figures 6 and 7, respectively.

From figure 6, we can affirm that the IS effect starts from a maximum value and then decreases until a constant value where it remains until the end of the simulation. Therefore, instead of presenting an inconsistent behavior, the IS adjusts the level of influence required considering the drug administration and the tumor volume evolution, presenting a constant help in fighting cancer.

Regarding figure 7, the Bevacizumab drug concentration increases and then reduces 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 also increases over time 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, meaning that the effect of the Atezolizumab drug on the pa-

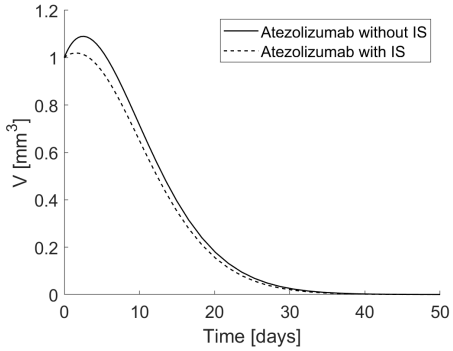

Figure 5: Tumor volume evolution with and without IS over time for Atezolizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

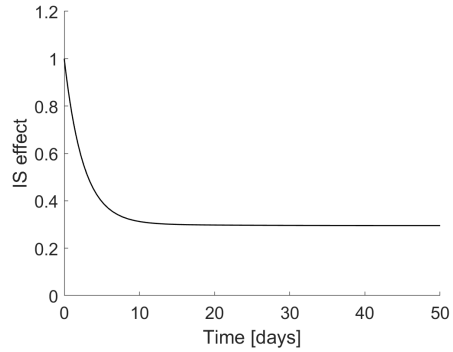

Figure 6: IS evolution over time for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

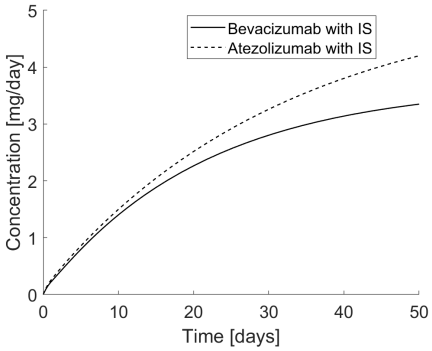

Figure 7: Drug concentration with IS over time for Bevacizumab and Atezolizumab, for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

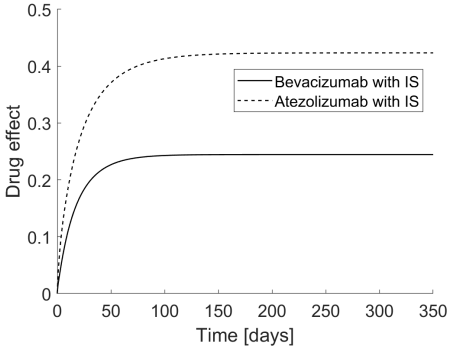

Figure 8: Drug effect with IS over time for Bevacizumab and Atezolizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

tient will be more significant, as can be seen in figure 8.

V. Control of the Logistic Model

By analysing the behaviour of the Logistic Model in previous chapter, we want to control this model to achieve better results regarding the tumor volume evolution. So, the controller applied to the TGM model can be described through a schema as described in figure 9.

A. Discretization with IS

We want to discretize the model using *Euler's* method [20]. 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:

$$\dot{V}(t) = aV(1 - \frac{V(t)}{K}) - \beta u(t)V(t) - \theta V(t)r(t), \quad (6)$$

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

Therefore, by assigning $f(V(t), r(t), u(t))$ to equation (6) and $g(V(t), r(t))$ to equation (5), we then have:

$$V(k+1) = V(k) + hf(V(k), r(k), u(k)). \quad (7)$$

$$r(k+1) = r(k) + hg(V(k), r(k)). \quad (8)$$

B. Linearization with IS

Then, we want to linearize the discretized model 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:

$$\Delta V(k+1) = \Delta V_1(k+1) + \Delta V_2(k+1) + \Delta V_3(k+1), \quad (9)$$

where:

$$\Delta V_1(k+1) = 1 + h(a - \frac{2a\bar{V}}{K} - \beta\bar{u} - \theta\bar{r})\Delta V(k). \quad (10)$$

$$\Delta V_2(k+1) = -h\theta\bar{V}\Delta r(k). \quad (11)$$

$$\Delta V_3(k+1) = -h\beta\bar{V}\Delta u(k), \quad (12)$$

and:

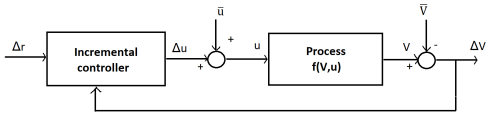

Figure 9: Logistic Model controller schema.

$$\Delta r(k+1) = \Delta r_1(k+1) + \Delta r_2(k+1), \quad (13)$$

where:

$$\Delta r_1(k+1) = h(\alpha_2 \bar{r} - 2\alpha_2\beta_2 \bar{V}\bar{r})\Delta V(k). \quad (14)$$

$$\Delta r_2(k+1) = 1 + h(\alpha_2 \bar{V} - \alpha_2\beta_2 \bar{V}^2 - \delta_2)\Delta r(k), \quad (15)$$

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

VI. Algorithms for RL

In this chapter, several algorithms for the application of Reinforcement Learning are presented, namely RLS with Exponential Forgetting, RLS with Directional Forgetting, Q-Learning and Velocity Algorithm.

The RLS with Exponential Forgetting concerns the forgetting factor that is related to the number of previous samples considered in the parameters estimation process. The RLS with Directional Forgetting is another variation of the normal RLS, where the transitions of the estimates by zero are not taken into account, leading to smoother 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, and so being a new controller it is important to explain it as well.

A. Q-Learning

The main RL algorithm presented in this master dissertation is the Q-Learning, where we do not know nor want to estimate the model parameters [10].

Thus, we consider a model as:

$$x_{k+1} = f(x_k) + g(x_k)u_k, \quad (16)$$

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

$$V(x_k) = r(x_k, u_k) + \gamma V(x_{k+1}), \quad (17)$$

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

$$r(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k. \quad (18)$$

Since we want to determine the optimal control policy, the goal is to minimize the value function of equation (17) by performing its partial derivative considering the control policy u_k such that:

$$\frac{\partial}{\partial u_k}(x_k^T Q x_k + u_k^T R u_k + \gamma V(x_{k+1})) = 0. \quad (19)$$

From equation (19) we then get:

$$2u_k^T R + g(x_k)\frac{\partial V(x_{k+1})}{\partial x_{k+1}} = 0, \quad (20)$$

that depends on the state variables u_k and x_{k+1}.

The objective now is to follow a path where there is no previous knowledge about the system and the environment [14], and so all that matters is the estimation of the cost function to define an optimal control law that minimizes the reward estimate [15]. More precisely, the *Bellman Optimality* equation [10] can be expressed as:

$$V^*(x_k) = \min_u(Q^*(x_k, u)), \quad (21)$$

where the optimal Q function [10] is given by:

$$Q^*(x_k, u_k) = r(x_k, u_k) + \gamma V^*(x_{k+1}), \quad (22)$$

and the optimal control [10] is then:

$$h^*(x_k) = \arg\min_u(Q^*(x_k, u)). \quad (23)$$

Hereupon, to get the optimal control policy [10] we just have to derive the optimal Q function in order to the control policy u such that:

$$\frac{\partial}{\partial u}(Q^*(x_k, u)) = 0. \quad (24)$$

From equation (24) we do not have a dependency on the state variables, 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 [10] as:

$$Q_h(x_k, u_k) = r(x_k, u_k) + \gamma Q_h(x_{k+1}, h(x_{k+1})), \quad (25)$$

Furthermore, we would like to perform the Q function for RL using Policy Iteration algorithm. Hence, for non-linear systems we might assume a parametric approximation of polynomials or neural networks such that:

$$Q_h(x, u) = W^T \phi(x, u), \quad (26)$$

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

$$\phi(x_k, u_k) = \begin{bmatrix} x_k^2 \\ u_k^2 \\ x_k u_k \end{bmatrix}, \tag{27}$$

and W is a vector with 3 entries representing the LS solution to the policy evaluation step of the Q-Learning policy iteration algorithm [10] given by:

$$W_{j+1}^T(\phi(x_k, u_k) + \gamma\phi(x_{k+1}, u_{k+1})) = r(x_k, h_j(x_k)). \tag{28}$$

The policy improvement step [10] is then:

$$h_{j+1}(x_k) = \arg\min_{h(.)}(W_{j+1}^T(\phi(x_k, u)). \tag{29}$$

Later, to estimate the optimal policy, 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).

B. Velocity Algorithm

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.

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 10 there is a schema that allows to better understand it.

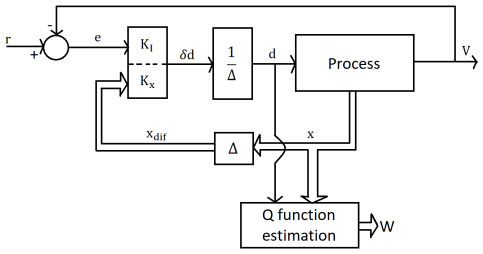

Figure 10: Velocity algorithm schema.

The algorithm presented in figure 10 can then be described through several difference equations:

$$\delta d(k) = -K_x x_{dif}(k) - K_I e(k) \tag{30}$$

$$d(k) = d(k-1) + \delta d(k) \tag{31}$$

$$e(k) = r(k) - V(k) \tag{32}$$

$$x_{dif}(k) = x(k) - x(k-1) \tag{33}$$

VII. Results

For the tumor volume as the output of the model and the drug effect its input, the first step is a model with a single state, the tumor volume. Then, the PK 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. Due to 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 (c_2 and V), called the full model. Finally, as the IS is crucial 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.

A. Approximated PK, PD and TGM model with IS

The most interesting results concern the final stage of the Q-Learning application to the model with three states: c_2, V and r. By using the previously presented Velocity Algorithm as control law, as well as the Directional RLS with exponential forgetting, we aim to control this system by using the LQ feedback gains for few instants and then the Q-Learning feedback gains that arise from the W estimates.

Hereupon, for a three state-space model the dimension is 10 combinations, and so the control law is:

$$u = -\frac{1}{2W_4}\begin{bmatrix} W_8 & W_9 & W_{10} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \tag{34}$$

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.

That said, the results with the Immune System influence are presented in figures 11, 12, 13, 14, 15 and 16.

Through figure 11, as expected, we can see that the dosage presents lower values since there is no need to apply as much dosage of the drug since the Immune System also contributes to the reduction of tumor.

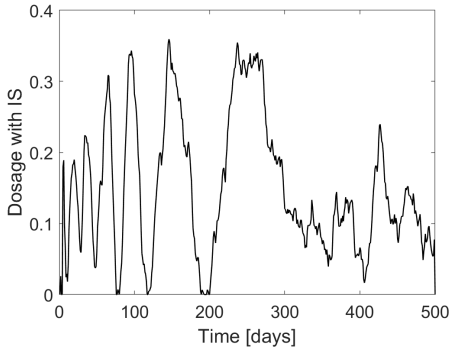

Figure 11: Dosage evolution comparing to reference through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

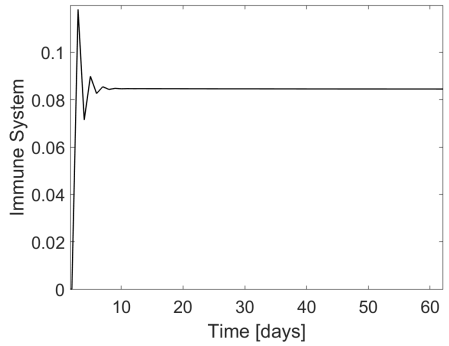

Figure 13: Immune system effect evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

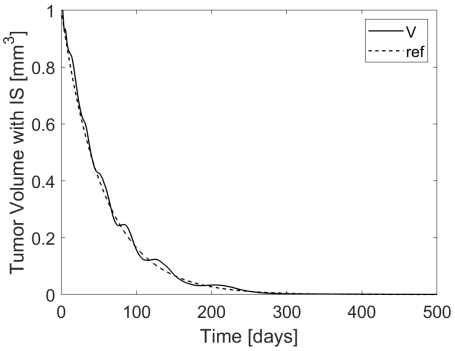

Figure 12: Tumor volume evolution comparing to reference through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

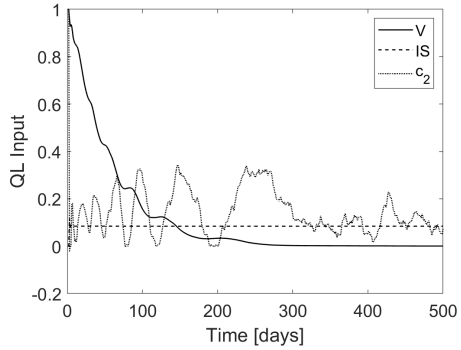

Figure 14: Q-Learning input signals evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

From figure 12 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 in the order of 10^{-4}.

From figure 13, after a zoom in to exactly observe the IS effect, we can conclude that it takes approximately 3 to 5 days to become constant, leading to the tumor volume decrease through time.

Thus, since now we have 3 states and the integrator state, from figure 14 we can see the three Q-Learning input states that allow the W estimation, as presented in figure 16, leading to the feedback gain of each state, as in figure 15.

Through figure 16, we might affirm that with the IS influence, the W estimates are constant from a specific instant, meaning that with the IS addition, we managed to be more accurate and precise in the tumor eradication process.

VIII. Q-Learning algorithm evaluation

Since the Q-Learning algorithm presented in this master dissertation uses Policy Iteration, which requires stabilizing feedback gains to produce good results, it is important to study the robustness and stability of this algorithm.

A. Robustness and Stability

Regarding the proof of robustness, we chose 2 parameters of the system to check the model's operating basin, where the optimal values used for this work were the ones presented in table (2). Thus, 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, as we can see from figure 17.

To study the stability of the algorithm, 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

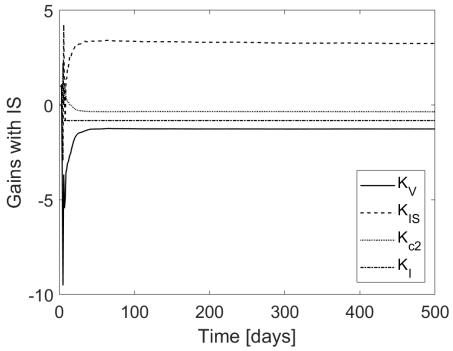

Figure 15: Gains evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

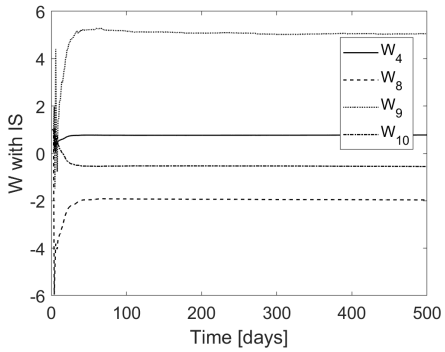

Figure 16: W estimates evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

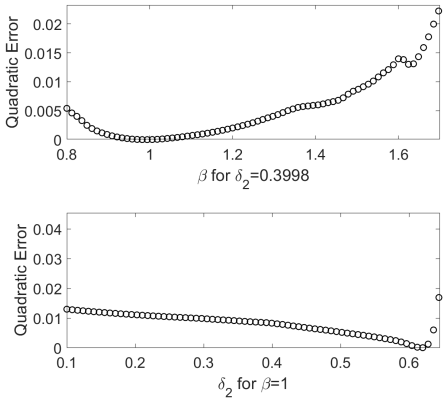

Figure 17: β and δ_2 individual evolution for $\delta_2 = 0.3998$ and $\beta = 1$, respectively, for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

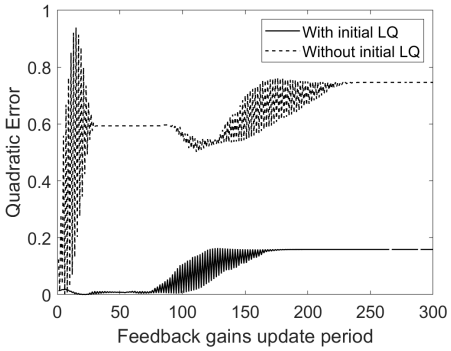

Figure 18: Tumor volume quadratic error evolution through increasing feedback gain update period, for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

instants and the use of the Q-Learning estimates from the beginning, as described in figure 18.

From figure 18, we can conclude that the tumor volume quadratic error for the dashed curve has an initial increase followed by a constant period. Then, there is 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 with initial LQ feedback gains is more accurate than the one without initial LQ values.

IX. Conclusion

Regarding the drug administration, 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, leading to a higher decrease of the tumor volume through time. With regards to the IS influence, we can affirm that it allows to almost instantly decrease the tumor volume.

Concerning the results, we can conclude that the IS response allowed to get a better performance regarding the tumor volume tracking of the reference, as well as the Q-Learning W estimates that kept constant until the end of the trial, after the stabilization interval. It is also important to mention that the Q-Learning was really challenging since initially it was not known which variables should be involved in the W estimates calculations, since there was the additional integrator state from the Velocity Algorithm, but 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.

Then, since the application of the Q-Learning as

a technique of the RL is the essential point of this work, through the analysis of 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 the chosen values for the variables were inside of the correct working range. Moreover, 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.

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.

References

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

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

[3] 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.

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

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

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

[7] 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.

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

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

[10] 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.

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

[12] 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.

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

[14] André Violante, "Simple Reinforcement Learning: Q-learning," 2019.

[15] Chathurangi Shyalika, "A Beginners Guide to Q-Learning," 2019.

[16] 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.

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

[18] C. Yu, J. Liu, "Reinforcement Learning in Healthcare: A Survey," *arXiv*, 2020.

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

[20] Wikipedia contributors, "Euler method — Wikipedia, the free encyclopedia," 2020.