# Reinforcement Learning Based Adaptive Control for Tumor Reduction

Inês Ferreira
*Instituto Superior Técnico*
Univ. Lisbon, Portugal
ines.m.ferreira@tecnico.ulisboa.pt

João M. Lemos
*INESC-ID*
*Instituto Superior Técnico*
Univ. Lisbon, Portugal
jlml@inesc-id.pt

Rita Cunha
*Institute for Systems and Robotics*
*Instituto Superior Técnico*
Univ. Lisbon, Portugal
rita.cunha@tecnico.ulisboa.pt

***Abstract*—This work addresses the design of cancer therapy for tumour reduction using adaptive optimal control based on reinforcement learning. The approach proposed consists of defining a decreasing reference trajectory for the tumour size, that drives it to zero with a convenient rate, together with a regulation algorithm that adjusts the drug dose so that the tumor size tracks this reference. The motivation to use adaptive methods stems from the high variability of biomedical dynamics, both inter and intra-patient, together with the aim of providing the regulation controller with the ability to tune to the optimal solution when the tumor size decreases. The adaptation mechanism uses Q-learning and a quadratic cost, resulting in a model-free linear quadratic controller. Directional forgetting recursive least squares is used to estimate the coefficients of the quality function. Simulation results, with a logistic tumor model that incorporates the the effect of immunotherapy are presented.**

***Index Terms*—adaptive control, reinforcement learning, Q-learning, linear quadratic, cancer therapy.**

## I. Introduction

### A. Motivation

The chemotherapy treatment of cancer tumors involves a balance between fast tumor reduction and the toxic effects of the treatment itself. Although, in clinical practice, the treatment dosing is made according to rules of art that stem from wide medical experience, it is increasingly recognised that it might be advantageous to make use of mathematical modelling and dynamic optimization. Formulating the above problem as an optimal control problem [1], [2] allows an objective decision on the amount and timing of drug dosing, adjusting the trade-offs involved by using a set of parameters.

The major difficulty of the above approach is, however, the high variability in dynamical behaviour that exists both inter- and intra-patients. The corresponding level of uncertainty might be such that it prevents the use of robust techniques, that would entail an unacceptable drop in control performance. To tackle such a problem, one might resort to adaptive control techniques, an approach in which the cancer therapy literature is scarce. In addition to other benefits, adaptive control paves the way to the increasingly important issue of personalized therapy [3]. Furthermore, it might be questioned whether a model-driven approach is appropriate to represent a system which is so complex that it is impossible to model from first-principles. This issue motivates data-driven approaches and, in particular, but not exclusively, the use of reinforcement learning [4].

Therefore, the above considerations motivate the design of tumor treatment techniques using adaptive optimal controllers based on reinforcement learning, which is the topic addressed in this work.

### B. Literature review

The use of optimal control to design cancer therapies has been the subject of a large bibliography. Some examples are [1], [5]. Although many works rely on Pontryagyn's principle, strategies based on model predictive control [7] and optimal impulsive control [2], [6] are also reported. The literature on the application of adaptive control to cancer therapy design is scarce. An example, where supervised multiple model adaptive control is used is given by [8].

Currently published articles on the application of reinforcement learning to tumor reduction do not include adaptation. Instead, a critic agent is trained with *a priori* data. Examples are given by [9] and [3], this last one using the OpenAI Gym framework. Another example is [10], that considers anti-angiogenic therapy. Other works on this topic are [11]–[13].

### C. Objectives and contributions

The objective of this work is the development of a model-free control algorithm that drives the size of a cancer tumor to a value close to zero, in a way that adapts to the patient and tumor characteristics.

The main contributions consists of a reinforcement learning adaptive feedback controller that incorporates integral effect to track a slowly changing reference, and its demonstration in a tumor dynamical model that incorporates the interaction with the immune system.

### D. Article structure

The article is organised as follows: after this introduction, the tumour model used for the simulations is described in section II; the reinforcement learning based adaptive control algorithm, as well as the control architecture, is explained in section III; simulation results, including a parameter sensitivity

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

study are shown in section IV; finally, section V draws conclusions.

## II. Tumour model

The aim of considering the tumor growth model is twofold: first to perform simulations of the effect of the drug dosage yielded by the adaptive controller (that is data-driven and, therefore, does not rely on a model) and, also, to provide insight on the dynamics of the system considered, that includes not only the tumor growth dynamics, but also the relation between the drug dose administered to the patient, and the interaction with other systems, like the immune system and angiogenesis [2].

As shown in figure 1, the drug dosage administered to the patient is related to the concentration $c_2$ in the so called effect compartment, *i. e.* the region of the body in which the drug has a therapeutic effect, by the pharmacokinetic model (PK). The effect $u$ caused by the drug is then computed using the pharmacodynamic model (PD). In turn, the tumor volume $V$ is affected by the effect $u$, in a way represented by the tumor growth model. In this article, the influence of the immune system on tumor growth is also considered.

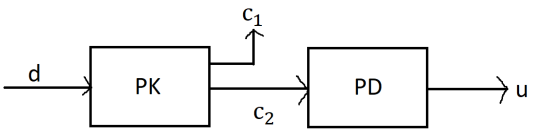


Fig. 1. Block diagram of the cancer models.

### A. Pharmacokinetics

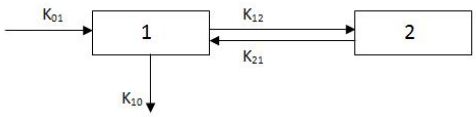


Fig. 2. Two compartment PK model.

The pharmacokinetics model [14] describes how the drug administered to the patient is distributed in the patient's body. This model assumes that the patient organism is divided in compartments, that correspond to one or more organs, which exchange drug among them, with a flow that depends on the concentration. Figure 2 shows a 2-compartment model, used in this article. Using the mass conservation principle, and denoting by $d$ the rate of drug administered to the patient, the concentrations $c_1$ and $c_2$ in the compartments are seen to satisfy the linear state model

$$\begin{bmatrix} \frac{dc_1}{dt} \\ \frac{dc_2}{dt} \end{bmatrix} = \begin{bmatrix} \frac{-k_{12}-k_{10}}{V_1} & \frac{k_{21}}{V_1} \\ \frac{k_{12}}{V_2} & \frac{-k_{21}}{V_2} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{V_1} \\ 0 \end{bmatrix} d, \qquad (1)$$

where $V_1$ and $V_2$ are parameters that represent the volume of each of the compartments 1 and 2, respectively and $k_{ij}$ drug dependent parameters. Assuming that 2 is the effect compartment, the output is then $y = c_2$.

### B. Pharmacodynamics

The pharmacodynamics (PD) model relates the concentration in the effect compartment, $c_2$, (input) to the drug effect, $u$. Opposite to the PK model, that is obtained from mass conservation, there is no first-principles way of obtaining the PD model in a compact way. Therefore, the model used relates its input and output variables in an aprioristic way, using the Hill equation, given by

$$u(t) = u_{max} \frac{c_2^{\alpha}(t)}{c_{50}^{\alpha} + c_2^{\alpha}(t)}, \qquad (2)$$

where $\alpha$, $u_{max}$, and $c_{50}$ are parameters to be estimated from data. The parameter $c_{50}$ represents the value of the concentration in the effect compartment for which the effect is half of the maximum $u_{max}$. These parameters may vary from patient to patient and also along time for a given patient, thereby justifying the use of adaptive control to compute the optimal drug dose.

### C. Logistic Growth Model

There are several models for tumor growth [15], [16]. Although the linear and exponential models have the advantage of simplicity, they are inadequate because they do not include a self-limitation of tumor growth, opposite to the logistic and Gompertz models. The model chosen to represent the tumor growth in this work is the logistic growth model, given by

$$\frac{dV}{dt} = aV(1 - \frac{V}{K}) - \beta uV - \theta_r Vr, \qquad (3)$$

where $V$ is the tumor volume, $u$ is the drug effect, and $r$ is the immunocompetent cell density, that establishes a relationship with the immune system. The parameters are assumed to the have the values $a = 0.1$, $\beta = 1$, $K = 5mm^3$, and $\theta_r = 1$.

In the absence of treatment and with no reaction from the immune system, the tumor volumes grows in time, according to a logistic equation, as a sigmoidal function.

### D. Interaction with the immune system

The interactions between the immune system (IS) and the cancer are extremely complex. Although more elaborate models can be considered, an approximation is provided by the the nonlinear dynamic system

$$\dot{r} = \alpha_2(1 - \beta_2 V)Vr + \gamma_2 - \delta_2 r. \qquad (4)$$

The numerical values of the constant parameters $\alpha_2$, $\beta_2$, $\gamma_2$, and $\delta_2$ are listed in table (I).

The PK and PD parameters used for two different drugs are presented in table (II).

A comparison between the tumor volume evolution for each drug with and without the IS influence, is presented in figures 3 and 4. The immune system boosts the action of the drug, and help reducing the tumor size.

TABLE I
IMMUNE SYSTEM PARAMETERS

| Parameters | Value |
|---|---|
| $\alpha_2$ | 0.00484 |
| $\beta_2$ | 0.00264 |
| $\gamma_2$ | 0.1181 |
| $\delta_2$ | 0.3998 |

TABLE II
PK AND PD PARAMETERS FOR TWO DIFFERENT 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 |
| $\alpha$ | 1 | 1 |

## III. RL BASED ADAPTIVE CONTROL

The adaptive control algorithm used relies on a data-based strategy that directly approximates the cost using plant (*i. e.*, tumor growth dynamics) input and state data. This algorithm consists of a reinforcement learning type algorithm in which the so called *quality function*, or Q-function is approximated using linear regression [19], resulting in *approximate dynamic programming* [20] that yields a suboptimal control law.

### A. Q-Learning

Consider a plant represented by the nonlinear state model

$$x_{k+1} = f(x_k) + g(x_k)u_k. \tag{5}$$

The aim is to design a control law $u$ that minimizes the cost defined by

$$V_u(x_k) := \sum_{i=k}^{\infty} \gamma^{i-k} r(x_i, u_i), \tag{6}$$

with the instantaneous reward $r(x_k, u_k)$ given by

$$r(x_k, u_k) = x_k^\top Q_c x_k + u_k^\top R u_k, \tag{7}$$

and where $Q_c = Q_c^\top \succeq 0$, $R = R^\top \succ 0$ are weighting matrices. This cost verifies the Bellman equation

$$V(x_k) = r(x_k, u_k) + \gamma V(x_{k+1}). \tag{8}$$

The optimal control $u^*$ satisfies the discrete-time Hamilton-Jacobi-Bellman (HJB) equation, expressed in the optimal cost $V^*$ along an optimal state trajectory, and given by

$$V^*(x_k) = \min_u \left( r(x_k, u_k^*) + \gamma V^*(x_{k+1}) \right). \tag{9}$$

Although the optimal control can be obtained by solving the HJB equation, as is well known, the numerical complexity involved prevents the use of this approach in all but a few cases. Furthermore, the plant model is assumed to be known, ruling out pure data-driven methods. However, it is possible to consider approximations that yield a suboptimal control and that are suitable for real time implementation. These approximations are made by considering the current state trajectory, instead of obtaining the value function (or optimal cost) for all the possible initial states, and they rely on the Bellman equation 8.

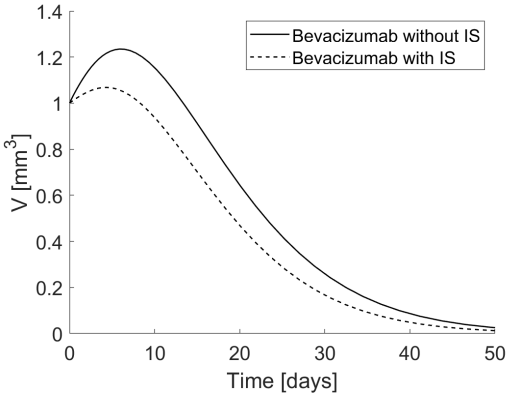


Fig. 3. Tumor volume evolution with and without IS over time for Bevacizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

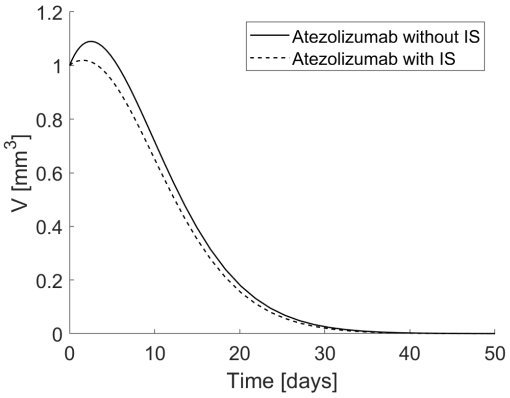


Fig. 4. Tumor volume evolution with and without IS over time for Atezolizumab for $\alpha = 1$, $u_{max} = 1$, $a = 0.1$, $K = 5$ and $V_i = 1$.

The control that optimizes the cost propagated by the Bellman equation 8 is given by

$$\frac{\partial}{\partial u_k}(x_k^T Q x_k + u_k^T R u_k + \gamma V(x_{k+1})) = 0, \tag{10}$$

from which

$$2u_k^T R + g(x_k)\frac{\partial V(x_{k+1})}{\partial x_{k+1}} = 0. \tag{11}$$

Solving this equation, yields $u_k$ as a function of $x_{k+1}$. Eliminating this non-causal dependency requires the use of the state model. To avoid this dependency, the quality or Q-function is defined by

$$Q(x_k, u_k) := r(x_k, u_k) + \gamma V(x_{k+1}), \tag{12}$$

and the optimal control is given by

$$\frac{\partial}{\partial u}(Q(x_k, u)) = 0. \tag{13}$$

The Q function satisfies the Bellman-like equation

$$Q(x_k, u_k) = r(x_k, u_k) + \gamma Q(x_{k+1}, u(x_{k+1})). \quad (14)$$

This equation is now used to define a linear regression model that can be used to approximate the Q function using least-squares. For that sake, assume a parametric approximation of the Q-function given by a linear combination of basis functions, as in

$$Q_h(x, u) = W^T \phi(x, u), \quad (15)$$

where $\phi(x, u)$ represents a set of functions. Actually, for linear quadratic problems, it is shown in [19] that using as basis functions 2nd order polynomials in the entries of the state and $u$ yields an exact representation of $Q$. For example, for a second-order, scalar system, the basis function polynomials are

$$\phi(x_k, u_k) = \begin{bmatrix} x_1^2(k) \\ x_2^2(k) \\ u^2(k) \\ x_1(k)x_2(k) \\ x_1(k)u(k) \\ x_2(k)u(k) \end{bmatrix}. \quad (16)$$

The vector of weights $W$ is a vector that can be estimated by the least-squares (LS) solution of the regression model

$$W_{j+1}^T(\phi(x_k, u_k) + \gamma\phi(x_{k+1}, u_{k+1})) = r(x_k, h_j(x_k)). \quad (17)$$

Performing this LS estimate forms the *learning step* of the Q-learning policy iteration algorithm

The above step is followed, in a recursive way, by the *policy improvement step* [19], defined by:

$$u_{j+1}(x_k) = \arg\min_{h(.)}(W_{j+1}^T(\phi(x_k, u))), \quad (18)$$

where the index $k$ refers to discrete time, and $j$ refers to the possibility of doing multiple iterations at each time.

In the case of the second-order plant mentioned above, with the regressor ordered as in (16), the control, according to (18), is computed as the linear state feedback

$$u(k) = -\frac{w_4}{2w_3}x_1(k) - \frac{w_5}{2w_3}x_2(k), \quad (19)$$

where the $w_i$ denote the estimates of the corresponding entries of $W$.

### B. Directional forgetting RLS

In order to get the estimates of $W$, several variants of the least-squares algorithm can be used, such as batch least-squares, recursive least-squares with exponential forgetting or recursive least-squares with directional forgetting (DF-RLS). In this work, $j = 1$ and DF-RLS is used.

Consider the general scalar linear regression model

$$y(k) = \varphi^\top(k)\theta + e(k), \quad (20)$$

where $k = 0, 1, 2, \ldots \in \mathbb{N}_o$, denotes discrete time, $y \in \mathbb{R}$ is the measured output data, $\varphi \in \mathbb{R}^{n_p}$ is measured regressor data, $e \in \mathbb{R}$ is a zero mean noise, unavailable to direct measurement, and $\theta \in \mathbb{R}^{n_p}$ is the vector of parameters to estimate. The directional forgetting algorithm is obtained by iterating the following equations

$$\beta(k) = 1 - \lambda + \frac{1 - \lambda}{\varphi^\top(k)P(k\ 1)\varphi(k)} \quad (21)$$

$$\mathbb{K}(k) = \frac{P(k-1)\varphi(k)}{1 + \varphi^\top P(k-1)\varphi(k)[1 - \beta(k)]} \quad (22)$$

$$P(k) = [I - \mathbb{K}(k)\varphi^\top(k)(1 - \beta(k)]P(k-1) \quad (23)$$

$$\hat{\theta}(k) = \hat{\theta}(k-1) + \mathbb{K}(k)[y(k) - \varphi^\top(k)\hat{\theta}(k-1)] \quad (24)$$

The expression for $\beta$ ensures that $P(k) \succ 0$ if $P(0) \succ 0$. The parameter $\lambda$ is the forgetting factor in the direction defined by the regressor for the incoming information.

The use of DF-RLS is advantageous in situations in which there are poorly identifiable parameters, as is the case in the problem considered here, where there are more parameters than controller gains. According to DF-RLS, when the observed regressor does not carry information about one parameter, its estimate is frozen, thereby avoiding drifts that can lead to numerical problems.

### C. Velocity Algorithm

The reinforcement learning based adaptive controller described in section III addresses the computation of a state feedback gain to regulate the plant, *i. e.*, to drive the state close to zero. Hereafter, we consider the modification of the basic controller to track a slowly tracking reference.

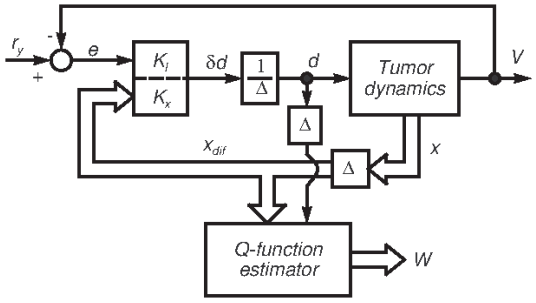


Fig. 5. Block diagram of the velocity algorithm.

Figure 5 shows the block diagram of the algorithm used to track a constant or slowly time-varying reference. The symbol $\Delta$ is defined as

$$\Delta := 1 - q^{-1},$$

where $q^{-1}$ is the backward shift operator. This structure includes an integrator that ensures small tracking error and a linear feedback of the state increment with respect to undisturbed state trajectory. Since the plant model is unknown, this increment is estimated by differentiating in time the state, obtaining

$$x_{dif}(k) := x(k) - x(k-1) \quad (25)$$

The manipulated variable is then computed by

$$\delta d(k) = -K_x x_{dif}(k) - K_I e(k), \quad (26)$$

where $e(k) = r_y(k) - V(k)$, $K_x$ is computed using the Q-learning algorithm, $K_I$ is an *a priori* chosen integral gain, and

$$d(k) = d(k-1) + \delta d(k). \tag{27}$$

This structure is known as the *velocity algorithm* and has been originally proposed in [22], being also described in [23] (pages 490-494) inm relation to gain-scheduling, to avoid the need to know the nominal state trajectory when using an incremental controller. In this paper, and as described in figure velocityAlgorithm, the velocity algorithm is complemented with extra blocks to estimate the parameters of the Q-function. In this respect, the key issue is to use incremental values of both the drug dose (control variable) and the state, to avoid biasing the estimates of the parameters that define $Q$.

## IV. Simulation results

These simulations consider 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 stabilizing feedback gains during an initial period and then the Q-Learning feedback gains that arise from the $W$ estimates.

For a third-order state model, the dimension of $W$ in (15) is 10, and the control law is

$$u = -\frac{1}{2w_4}\begin{bmatrix} w_8 & w_9 & w_{10} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \tag{28}$$

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 $d$.

The results, including the immune system influence are presented in figures 6, 7, 8, 9, and 10.

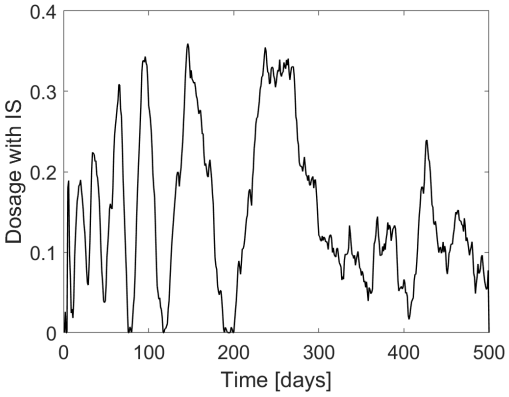


Fig. 6. Dosage evolution comparing to reference through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

Figure 11 shows the average quadratic cost that results when the controller obtained at convergence is used with model parameter values that are different from the ones with which the adaptive controller was originally simulated. The study

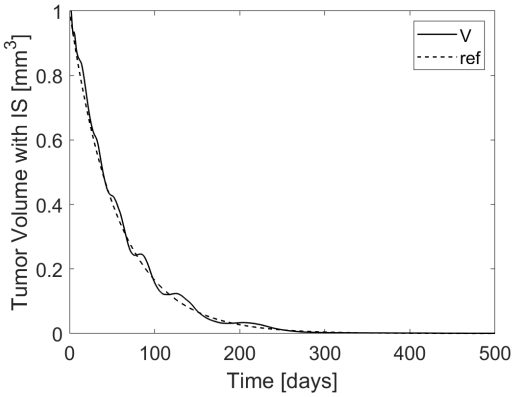


Fig. 7. Tumor volume evolution comparing to reference through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

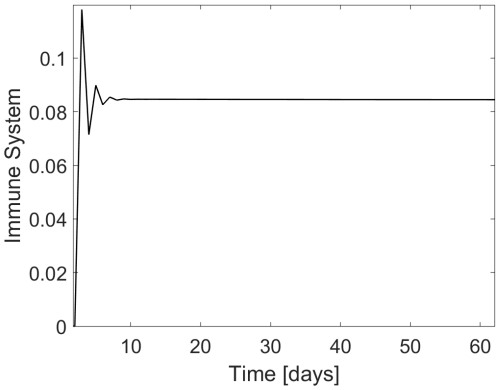


Fig. 8. Time evolution of the immunocompetent cell concentration

refers to model parameters $\beta$ and $\delta_2$ and provides an idea of the robustness properties of the controller obtained at convergence. The set of parameters for which the error surface is reasonably flat correspond to parameters for which the controller achieves insensitivity to model uncertainty.

## V. Conclusions

A model-free adaptive control algorithm to automatically adjust the drug dosage in order to reduce the size of cancer-tumors has been presented and demonstrated in simulation. The algorithm relies on Q-learning and least-squares identifiers to estimate an approximation of the quality function. The results obtained are comparable to the ones of other published reinforcement learning algorithms, that rely on different approaches.

Future work comprise other forms of incorporating in the controller *a priori* information on the patient, as well as a coordination with immune and anti-angiogenic therapies and tests with more detailed cancer models. A strong assumption made in this work is that the state is available. A significant problem is therefore to alleviate this requirement by using

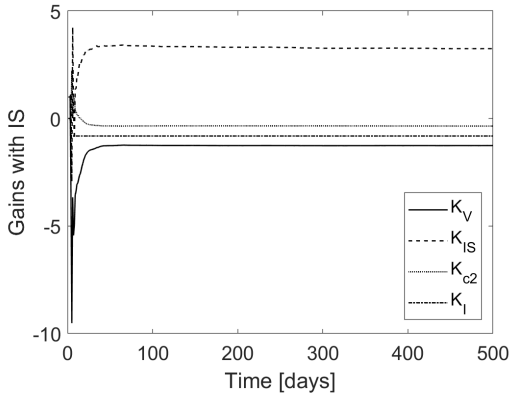


Fig. 9. Gains evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

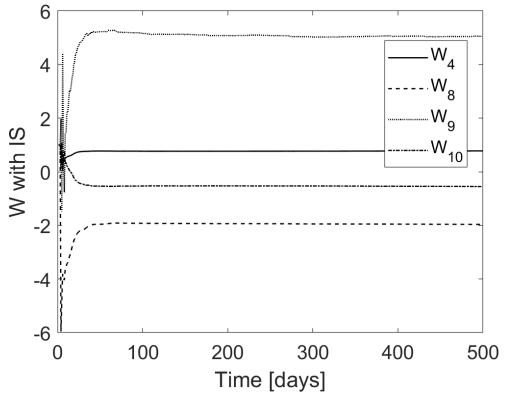


Fig. 10. $W$ estimates evolution through time for $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

algorithms that require less variables to be available for direct measure.

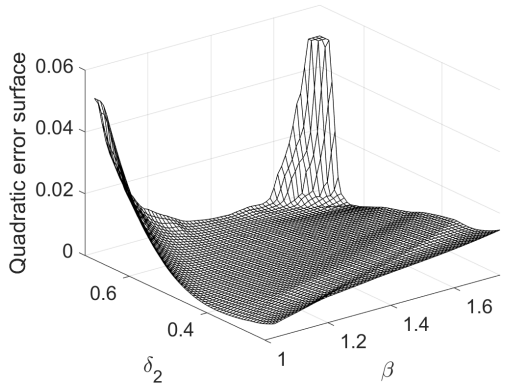


Fig. 11. The root-mean-square of the tracking error as a function of the model parameters $\beta$ and $\delta_2$. Reinforcement learning at convergence with $R = 0.1$, $\lambda = 0.995$ and $\gamma = 0.95$, using velocity algorithm as control law and RLS with directional forgetting for Q-Learning, with IS influence.

## REFERENCES


[1] H. Schättler and U. Ledzewicz. *Optimal control for mathematical models of cancer therapies*. Springer, 2010.

[2] J. P. Belfo and J. M. Lemos. *Optimal Impulsive Control for Cancer Therapy*. Springer, 2021.

[3] J. Tordesillas and J. Arbelaiz. Personalized Cancer Chemotherapy Schedule: a numerical comparison of performance and robustness in model-based and model-free scheduling methodologies, 2019, arXiv:1904.01200v3 [cs.LG]

[4] C. Yu, J. Liu, and S. Nemati. Reinforcement Learning in Healthcare: A Survey 2020, arXiv:1908.08796v4 [cs.LG]

[5] H. Sbeity and R. Younes. Review of optimization methods for cancer chemotherapy treatment planning. *Comput. Sci, Syst, Biol.*, 8:74-95, 2015.

[6] F. Cacace, V. Cusimano and V. Palumbo. Optimal impulsive control with application to antiangiogenic tumor therapy. *IEEE Trans. Control Systems Technol.*, 28(1):106-117, 2020.

[7] T, Chen, N. K. Kiirkby and R. Jena. Optimal dosing of cancer chemoteraphy using model predictive control and moving horizon state/parameter estimation. *Comp. Methods Progr. Biomed.*, 108(3): 973-983, 2012.

[8] F. F. Teles and J. M. Lemos. Cancer therapy optimization based on multiple model adaptive control. *Biomedical Signal Processing and Control*, 48: 255-264, 2019.

[9] R. Padmanabhan, N. Meskin and W. M. Haddad. Reinforcement-learning-based control of drug dosing for cancer therapy treatment. *Mathematical Biosciences*, 293: 11-20, 2017.

[10] P. Yazdjerdi, N. Meskin, M. Al-Naeimi, A. Moustafa, and L.Kovács. Reinforcement learning-based control of tumor growth under anti-angiogenic therapy. *Computer Methods and Programs in Biomedicine*, 173: 15-26, 2019.

[11] R. Padmanabhan, N, Meskin and W. M. Haddad. Reinforcement learning-based control of drug dosing with applications to anesthesia and cancer therapy. *Control Applications for Biomedical Engineering Sciences* (A. T. Azar, ed.), 251-297, Academic Press, 2020.

[12] Y. Zhao, M. Kosorok and D. Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294-315, 2009.

[13] Y. Zhao, M. Kosorok, D. Zeng, and M. Socinski. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4): 1422-1433, 2011.

[14] J. M. Gallo. Pharmacokinetics: Model structure and transport systems. *Clinical Research and Regulatory Affairs*, 18(3): 235-266, 2001.

[15] S. Benzekry, C. Lamont, A. Beheshti, A. Tracz, J. Ebos, L. Hlatky, and P. Hahnfeldt. Classical mathematical models for description and prediction of experimental tumor growth. PLOS Comput. Biol. 10(8):e1003800, 2014.

[16] J. J. Yin and D. Moes. A review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors. *CPT: Pharmacometrics and Systems Pharmacology*, 8(10): 720-737.

[17] A. Eladdadi, P. Kim, and D. Mallet (eds.). *Mathematical models of tumor-immune system dynamics*. Springer, 2014.

[18] K. Garber. Promising early results for immunotherapy-antiangiogenesis combination. *J. Natl. Cancer Inst.* 106(11): dju392, 2014.

[19] F. L. Lewis, D. Vrabie, and G. Vamvoudakis. Reinforcement learning and feedback control. *IEEE Control Systems Magazine*, 32(6): 76-105, 2012.

[20] R. S. Sutton and A. G. Barto. *Reinforcement Learning*, MIT Press, 2nd ed., 2018.

[21] R. Kulhavý. Restricted exponential forgetting in real-time identification. *Automatica*, 23(5):589-600, 1987.

[22] I. Kaminer, A. M. Pascoal, P. P. Khargonekar, and E. E. Coleman. A velocity algorithm for the implementation of gain scheduled controllers. *Automatica*, 31: 1185-1191, 1995.

[23] H. K. Khalil. *Nonlinear Systems*, 3rd ed., Prentice Hall, 2002.