AIMS Mathematics, 11(7): 21167–21189.
DOI: 10.3934/math.2026860
Received: 12 May 2026
Revised: 15 June 2026
Accepted: 29 June 2026
Published: 16 July 2026

http://www.aimspress.com/journal/Math

Research article

Adaptive tracking control with Lyapunov regularization for data-calibrated Takagi–Sugeno fuzzy Markov jump biological systems

Shirali Kadyrov[1,4,*], **Ardak Kashkynbayev**[3] **and Yershat Sapazhanov**[2,5]

[1] Department of Mathematics, New Uzbekistan University, Mirzo Ulugbek, Movarounnahr 1, Tashkent, Uzbekistan

[2] School of Digital Technologies, Narxoz University, 55 Zhandossov Street, Almaty 050035, Kazakhstan

[3] School of Computing and Artificial Intelligence, Nazarbayev University, 53 Kabanbay Batyr Avenue, Astana 010000, Kazakhstan

[4] Department of Mathematics, SDU University, Abylai Khan Street 1/1, Kaskelen 040900, Karasai District, Almaty Region, Kazakhstan

[5] Institute of Mathematics and Mathematical Modeling, Pushkin St. 125, 050010, Almaty, Kazakhstan

* **Correspondence:** Email: shiralikadyrov@gmail.com.

Abstract: Cancer–tumour–immune dynamics are nonlinear, uncertain, and subject to abrupt biological or treatment-induced regime changes. This paper develops a data-calibrated adaptive tracking-control framework for Takagi–Sugeno (T–S) fuzzy Markov jump biological systems with unknown local dynamics and unknown transition probabilities. The proposed method combines mode-dependent fitted Q iteration with a Lyapunov-penalized policy update, so that the learned controller uses transition data while remaining close to a bounded stabilizing reference action. For the sampled discounted problem, we establish Bellman contraction, well-posedness of the fitted regression update, an approximate fitted Q error bound, a quantitative Lyapunov-penalization bound, a practical tracking guarantee for the Lyapunov-regularized policy, and positivity preservation under projected states and dimensionless inputs. Public tumour-volume data associated with the FIR atezolizumab non-small-cell lung cancer study are used only to calibrate plausible tumour-growth ranges; a trajectory-level holdout split is applied to avoid in-sample reporting. In the resulting active-treatment simulation regime, the proposed controller gives the best overall balanced performance among the compared strategies, which now include proximal policy optimization (PPO), soft actor–critic (SAC), nominal model predictive control (MPC), local tracking feedback, reference-action control, fixed scheduling, non-switching fitted Q iteration, and no control. In the representative single-seed run, the Lyapunov-regularized fitted Q iteration (LR-FQI) method achieves low weighted cost (51.17), low tumour integral

squared error (ISE) (1.70), short threshold-exceedance time (4 days), and moderate total dimensionless input. A Monte Carlo evaluation over 50 independent Markov and noise trajectories gives the lowest mean weighted cost for the proposed controller (47.29 ± 19.91, mean ± standard deviation) and a consistent threshold-exceedance time of 4.00 ± 0.00 days. Assumption 4 is verified numerically over 5000 sampled domain points.

Keywords: adaptive tracking control; reinforcement learning; fitted Q iteration; Lyapunov regularization; Takagi–Sugeno fuzzy systems; Markovian jump systems; tumour–immune dynamics; data-driven control
Mathematics Subject Classification: 93C40, 93E35, 92C50, 68T05

1. Introduction

For clarity, all specialized terms used in the paper have the following meanings. A Takagi–Sugeno (T–S) fuzzy model is a weighted sum of simple local models, where the nonnegative weights are called membership functions and add to one. A Markov jump system is a system whose active regime is selected by a finite-state Markov chain. Data-driven control means that the controller is learned from simulated or observed transition data rather than from known system matrices. Fitted Q iteration is a reinforcement-learning method that approximates the discounted state-action value function, denoted by Q, from a finite data set. Lyapunov penalization means that policy selection includes a quadratic penalty that keeps the chosen input close to a stabilizing reference input. Throughout the paper, all control inputs are dimensionless variables in $[0, 1]$ and are not clinical doses.

Mathematical oncology uses differential equations, stochastic processes, and optimization methods to describe tumour growth, immune response, and treatment effects. Classical tumour–immune models, including the Kuznetsov–Makalkin–Taylor–Perelson model and the Kirschner–Panetta model, show that tumour–immune interactions can exhibit immune stimulation, immune escape, dormancy, bistability, and treatment-dependent transitions [1,2]. These behaviours make feedback design difficult because a controller must handle nonlinearity, parameter uncertainty, positivity of biological variables, treatment constraints, and abrupt biological regime changes.

Optimal control has been used to design chemotherapy, immunotherapy, and combined treatment strategies. However, many cancer-control formulations remain model-based: they assume known tumour-growth parameters, immune recruitment rates, drug response coefficients, or an identified nonlinear model. In practice, these quantities are patient-dependent and difficult to estimate reliably. Biological regimes may also change because of immune suppression, drug resistance, treatment response, remission-like behaviour, or progression. Such transitions motivate Markovian jump modelling, while the nonlinear continuous dynamics can be represented by Takagi–Sugeno fuzzy rules.

Reinforcement learning (RL), adaptive dynamic programming, and fitted Q iteration provide data-driven methods for learning feedback policies without explicit model identification [3]. Existing RL methods for Markovian jump systems have addressed unknown transition probabilities and unknown dynamics, mostly in engineering-oriented settings [4–6]. Related AIMS Mathematics work has also studied tracking control for T–S fuzzy semi-Markovian jump systems, although in a model-based H_∞

setting rather than a data-driven biological setting [7]. Switched and constrained nonlinear systems pose closely related challenges: Zong et al. established adaptive fuzzy tracking control with input saturation for switched nonlinear systems [8]; Wang and Zong developed dynamic event-triggered adaptive fixed-time practical tracking control for nonlinear systems through a funnel function [9]; and Gao et al. studied finite-time H_∞ control of switched affine non-linear systems based on the Lie derivative and iteration technique [10]. These works motivate Lyapunov-based and constraint-aware designs, and the present paper extends their spirit to the data-driven Markov jump setting. In parallel, cancer-control studies have addressed adaptive tracking, robust control, positive T–S fuzzy control, and sampled-data tracking for tumour–immune models [11–14]. The intersection remains underdeveloped: there is limited work on data-driven adaptive tracking control for data-calibrated T–S fuzzy Markov jump biological systems.

The present paper develops such a framework. Its importance is twofold. From a systems-and-control perspective, it provides the first fully data-driven, Lyapunov-regularized fitted Q framework for T–S fuzzy Markov jump systems that is simultaneously applicable to biological regime switching, positivity constraints, and adaptive reference tracking without any model identification step. From a mathematical-oncology perspective, it demonstrates a reproducible workflow that connects public patient-response data to a rigorous computational control study, filling the methodological gap between data-driven RL and structured biological modelling. It is motivated by the objective of applying reinforcement learning from previously collected transition data to adaptive tracking control for cancer–tumour–immune systems and completely unknown T–S fuzzy Markovian jump systems. The public tumour data are used only to calibrate plausible growth ranges. The learned controls are normalized computational variables and are not clinical doses.

The main contributions are as follows.

(1) We establish a data-calibrated T–S fuzzy Markov jump formulation for tumour–immune tracking in which the local dynamics and transition probabilities are unknown to the controller. This formulation connects biological regime switching, fuzzy nonlinear approximation, positivity constraints, and adaptive tracking in a single control problem. Unlike model-based T–S fuzzy approaches [13, 14] and switched-system designs that require plant identification [8–10], the proposed formulation requires only a previously collected transition data set.

(2) We design a mode-dependent fitted Q iteration method for Markov jump biological systems using an augmented tracking state, fuzzy membership information, regime labels, and bounded dimensionless inputs. Unlike a single-mode learner, the proposed construction explicitly separates regime-dependent value functions while avoiding model identification. The control law follows the parallel distributed compensation (PDC) paradigm [15]: A shared policy structure is applied across fuzzy rules, with the Lyapunov penalty providing the coupling discipline that PDC designs ordinarily obtain from Lyapunov inequality conditions.

(3) We develop a Lyapunov-penalized policy update that couples data-driven value minimization with a stabilizing reference action. This mechanism controls the saturation behaviour often produced by purely greedy fitted Q policies and yields interpretable bounded inputs suitable for constrained biological simulations.

(4) We provide a theoretical analysis of the sampled discounted control problem, including contraction of the Bellman operator, well-posedness of the fitted regression step, an approximate fitted Q error bound, a quantitative bound on the effect of Lyapunov penalization, a practical

tracking guarantee for the Lyapunov-regularized policy, and positivity preservation under state and input projection. Assumption 4 (reference-action drift condition) is supported by a numerical verification over 5000 sampled domain points.

(5) We demonstrate the framework in a reproducible data-calibrated tumour–immune simulation based on public non-small-cell lung cancer target-lesion trajectories, using a trajectory-level train-test split. The proposed controller is evaluated together with eight comparison controllers: PPO [16], SAC [17], nominal MPC, local tracking feedback, reference-action control, fixed scheduling [12], non-switching fitted Q iteration [18], and no control. Thus, the revised numerical section reports nine controllers in total. A Monte Carlo evaluation over 50 seeds shows that the proposed LR-FQI achieves the lowest mean weighted cost and the shortest threshold-exceedance time jointly with the reference-action baseline. The proposed controller therefore provides the strongest overall balance across weighted cost, tumour ISE, threshold-exceedance time, immune-state behaviour, and input use within the stated simulator, rather than uniformly dominating every baseline on every individual metric.

(6) We identify the following limitations: (i) the drift condition of Assumption 4 is verified numerically on a compact sampled domain and not proved analytically for the full state space; (ii) the biological parameters are fixed in-silico quantities calibrated from public data, not personalized to individual patients; (iii) PPO and SAC are trained for only 15000 timesteps, which is sufficient for comparison but suboptimal; and (iv) the framework does not address partial observability of Markov modes or pharmacokinetic/pharmacodynamic constraints. These limitations delineate directions for future work.

2. Literature review and research gap

The present work sits at the intersection of mathematical oncology, fuzzy modelling, Markov jump systems, and reinforcement learning. The biological motivation starts from classical tumour–immune modelling. The Kuznetsov–Makalkin–Taylor–Perelson model demonstrated that cytotoxic lymphocyte response can generate tumour dormancy, escape, and threshold-like behaviours [1]. The Kirschner–Panetta model then provided a canonical tumour–immune–interleukin framework for analysing immunotherapy mechanisms [2]. Subsequent deterministic and stochastic studies extended these ideas to chemotherapy, immunotherapy, and optimal-control settings [19, 20]. Recent AIMS Mathematics work has also examined tumour–immune dynamics with fractional-type operators, further illustrating the journal-level interest in mathematical tumour–immune modelling [21]. In data-driven tumour modelling, Ghaffari Laleh et al. showed that even classical growth laws require careful validation on real patient-response data [22]. These studies motivate the use of real tumour measurements for calibration, while also highlighting that response data cannot by themselves validate a treatment policy.

The modelling component is rooted in the fuzzy-systems tradition initiated by Zadeh's theory of fuzzy sets [23]. Takagi and Sugeno later introduced the T–S fuzzy modelling paradigm, which represents nonlinear dynamics through locally valid rules blended by membership functions [15]. This paradigm has become a standard tool for nonlinear control because it allows nonlinear systems to be analysed through convex combinations of local models. In cancer-control applications, positive T–S fuzzy control and sampled-data output tracking have already been studied for tumour–immune

dynamics [13, 14]. However, these works are primarily model-based: they usually require a specified plant structure, identified coefficients, or Lyapunov inequalities built from known local dynamics.

The switching component follows the Markovian jump-systems literature. Markov jump models were developed to represent systems whose dynamics change abruptly because of failures, environmental changes, mode switches, or unobserved regimes. The monographs of Costa, Fragoso, Marques, and coauthors provide a rigorous foundation for discrete- and continuous-time Markov jump linear systems, including stability, filtering, and optimal-control theory [24, 25]. These results are fundamental for engineering systems, but biological regimes such as therapy sensitivity, immune suppression, resistance, and recovery-like response create additional difficulties: States must remain nonnegative, transition probabilities are rarely known exactly, and safe exploration is constrained.

The control-learning component traces back to Bellman's dynamic programming principle [26]. Approximate dynamic programming and neuro-dynamic programming, developed prominently by Werbos and by Bertsekas, provided the conceptual bridge between dynamic programming, function approximation, and learning-based control [27, 28]. Modern reinforcement learning, synthesized by Sutton and Barto, gives a practical framework for learning value functions and policies from transition data [3]. In Markov jump systems, online reinforcement learning and off-policy adaptive optimization have been developed for unknown transition probabilities and unknown dynamics, mainly in engineering-oriented settings [4–6]. Related AIMS Mathematics work has studied tracking control for T–S fuzzy semi-Markovian jump systems, although in a model-based H_∞ setting rather than a data-driven biological setting [7]. On the switched-systems side, Zong et al. developed adaptive fuzzy tracking control for switched nonlinear systems with input saturation [8]; Wang and Zong addressed practical tracking control for constrained nonlinear systems with state-dependent switching [9]; and Gao et al. studied H_∞ control for switched affine nonlinear systems via switching-law design [10]. These works demonstrate the value of Lyapunov-based, constraint-aware designs for switched and constrained systems, and motivate the Lyapunov-penalized policy update developed in the present paper.

The research gap is therefore not the absence of any single ingredient. Rather, the gap is the absence of a computationally reproducible framework that combines: (i) T–S fuzzy approximation of tumour–immune nonlinearities, (ii) Markovian biological regime changes, (iii) fitted Q learning from previously collected transition data, (iv) adaptive tracking of a reference tumour–immune trajectory, (v) positivity and input constraints, and (vi) calibration from public tumour-volume data. This paper addresses this gap and shows that Lyapunov regularization is important for obtaining a robust and interpretable policy in the proposed biological simulation setting.

3. Problem formulation

Let $\mathbb{R}^4_+$ be the set of four-dimensional vectors with nonnegative components. Let $x(t) = [T(t), E(t), C(t), I(t)]^T \in \mathbb{R}^4_+$ denote the biological state at time t, where T is normalized tumour burden, E is normalized immune response, C is chemotherapy-like concentration, and I is immunotherapy-like concentration. The superscript T denotes transpose. Let $u(t) = [u_c(t), u_i(t)]^T \in [0, 1]^2$ denote dimensionless control inputs, where u_c is a chemotherapy-like input and u_i is an immunotherapy-like input. These inputs are not clinical doses.

Let $\sigma(t) \in \mathcal{S} = \{1, 2, 3\}$ denote a Markovian biological regime. The three modes are interpreted

as therapy-sensitive growth, resistant or immune-suppressed growth, and recovery-like response. The transition probabilities are unknown to the controller. The continuous-time T–S fuzzy Markovian jump representation is

$$\dot{x}(t) = \sum_{\ell=1}^{L} h_\ell(z(t))(A_{\ell,\sigma(t)}x(t) + B_{\ell,\sigma(t)}u(t) + a_{\ell,\sigma(t)}) + w(t). \tag{3.1}$$

Here L is the number of fuzzy rules, $z(t)$ is the premise variable used to evaluate the membership functions, $h_\ell(z(t)) \geq 0$, $\sum_{\ell=1}^{L} h_\ell(z(t)) = 1$, and $w(t)$ is a bounded disturbance. For rule ℓ and mode i, $A_{\ell,i}$ is the local state matrix, $B_{\ell,i}$ is the local input matrix, and $a_{\ell,i}$ is the local affine term. These matrices, affine terms, and transition probabilities are not used by the learning algorithm.

Remark 1 (PDC versus non-PDC design)**.** *The control architecture follows the spirit of* parallel distributed compensation *(PDC) [15]: The controller is constructed to be compatible with the same fuzzy-rule partition that represents the plant, and the learned value approximation uses the fuzzy-membership vector as part of the sampled learning state. PDC is preferable here to a non-PDC design because it preserves the convex blending structure of the T–S model and avoids introducing independently designed local controllers whose cross-rule compatibility would be difficult to guarantee when the local matrices $A_{\ell,i}$ and $B_{\ell,i}$ are unknown. The Lyapunov penalty in* (4.4) *provides the stabilizing discipline that model-based PDC designs ordinarily obtain from Lyapunov or LMI conditions, while remaining applicable in the present data-driven setting.*

The reference trajectory $x_r(t) = [T_r(t), E_r(t), 0, 0]^T$ specifies the desired tumour and immune trajectories:

$$x_r(t) = \left[0.22 + 0.68e^{-0.025t},\ 0.65 - 0.20e^{-0.020t},\ 0,\ 0\right]^T. \tag{3.2}$$

The augmented tracking state is

$$X(t) = \mathrm{col}(x(t) - x_r(t), x_r(t), 1) \in \mathbb{R}^9. \tag{3.3}$$

For implementation, the system is sampled with step size $\Delta t = 2$ days. Define $X_k = X(k\Delta t)$, $h_k = h(z(k\Delta t))$, $i_k = \sigma(k\Delta t)$, and $u_k \in \mathcal{A}$, where $\mathcal{A}$ is a finite grid on $[0, 1]^2$. The discounted tracking objective is

$$J^\mu(z_0) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k c(z_k, \mu(z_k)) \,\middle|\, z_0\right], \tag{3.4}$$

where μ is a feedback policy, $c(z_k, \mu(z_k))$ is the one-step tracking cost, $z_k = (X_k, h_k, i_k)$ is the sampled learning state, and $0 < \gamma < 1$ is the discount factor. In the numerical implementation, the one-step cost is

$$c(z_k, u_k) = q_T(T_k - T_{r,k})^2 + q_E(E_k - E_{r,k})^2 + r_c u_{c,k}^2 + r_i u_{i,k}^2, \tag{3.5}$$

where $q_T, q_E, r_c, r_i > 0$ are fixed weights, $T_{r,k} = T_r(k\Delta t)$, and $E_{r,k} = E_r(k\Delta t)$.

4. Lyapunov-penalized fitted Q iteration

For each mode $i \in \mathcal{S}$, approximate the state-action value function by

$$Q_i(X, h, u; \theta_i) = \phi(X, h, u)^T \theta_i, \tag{4.1}$$

where θ_i is the parameter vector to be learned and $\phi(X, h, u)$ contains all second-order monomials of $[X^T, h^T, u^T]^T$, including constant, linear, and quadratic terms. A standard greedy fitted Q policy is

$$\mu_i^Q(X, h) = \arg\min_{u \in \mathcal{A}} Q_i(X, h, u; \theta_i). \tag{4.2}$$

The single-mode comparison uses a single parameter vector θ for all modes. The proposed controller learns one parameter vector for each Markov mode and then performs a regularized policy update.

Let $\bar{u}(X, h, i)$ be a stabilizing reference action designed to reduce tumour tracking error while preserving input bounds:

$$\begin{aligned} \bar{u}_c &= \Pi_{[0,1]}\left(3.5(T - T_r)_+ + 0.2h_2 + 0.4\mathbf{1}_{\{i=2\}} + 0.8\mathbf{1}_{\{T>0.70\}}\right), \\ \bar{u}_i &= \Pi_{[0,0.4]}\left(0.05(E_r - E)_+ + 0.03(T - T_r)_+ + 0.02\mathbf{1}_{\{i=2\}}\right), \end{aligned} \tag{4.3}$$

where $(s)_+ = \max\{s, 0\}$, $\mathbf{1}_{\{\cdot\}}$ equals one when its condition is true and zero otherwise, and $\Pi_{[a,b]}(s) = \min\{b, \max\{a, s\}\}$ is a scalar projection onto $[a, b]$. The proposed Lyapunov-penalized policy is

$$\mu_i^{\mathrm{LR}}(X, h) = \arg\min_{u \in \mathcal{A}} \left\{ Q_i(X, h, u; \theta_i) + \lambda_L \|u - \bar{u}(X, h, i)\|^2 + \rho\|u\|^2 \right\}. \tag{4.4}$$

This is not a clinical rule. It is a mathematical regularizer that prevents unconstrained fitted Q iteration from selecting saturated actions that are inconsistent with the tracking objective (see Algorithm 1).

Algorithm 1 Lyapunov-penalized mode-dependent fitted Q iteration.

1: Choose an action grid $\mathcal{A} \subset [0, 1]^2$ and initialize $\theta_i^{(0)} = 0$, $i = 1, 2, 3$.
2: Generate previously collected transitions $\mathcal{D} = \{(X_k, h_k, u_k, i_k, c_k, X_{k+1}, h_{k+1}, i_{k+1})\}$ using a bounded exploratory behaviour policy.
3: **for** $r = 0, 1, \ldots, R - 1$ **do**
4: **for** each mode $i \in \mathcal{S}$ **do**
5: Form targets

$$y_k = c_k + \gamma \min_{v \in \mathcal{A}} Q_{i_{k+1}}(X_{k+1}, h_{k+1}, v; \theta_{i_{k+1}}^{(r)})$$

for all samples with $i_k = i$.
6: Solve the ridge regression

$$\theta_i^{(r+1)} = \arg\min_{\theta} \sum_{k: i_k = i} \left(\phi(X_k, h_k, u_k)^T \theta - y_k\right)^2 + \lambda\|\theta\|^2.$$

7: **end for**
8: **end for**
9: Return the Lyapunov-penalized policy (4.4).

5. Theoretical analysis

This section analyses the sampled discounted problem solved by Algorithm 1. The continuous biological model motivates the simulator, but the learning algorithm operates on sampled transition data with a finite action grid. The results below therefore establish well-posedness, approximation

properties, and conditional Lyapunov-style tracking guarantees for the sampled fitted-Q formulation. They do not assert clinical optimality and they do not constitute a continuous-time dosing theorem.

Let $z = (X, h, i) \in \mathcal{Z}$ denote the sampled learning state, where X is the augmented tracking state, h is the fuzzy-membership vector, and $i \in \mathcal{S}$ is the Markov mode. Let $\mathcal{A} \subset [0,1]^2$ be the finite action grid. The projected simulator induces a transition kernel $\mathbb{P}(\cdot \mid z, u)$ on $\mathcal{Z}$ for each $(z, u) \in \mathcal{Z} \times \mathcal{A}$. Thus the sampled problem is a discounted Markov decision process in the standard sense of dynamic programming and reinforcement learning [3,26,27,29]. Let $\mathcal{B}(\mathcal{Z}\times\mathcal{A})$ be the Banach space of bounded real-valued functions on $\mathcal{Z} \times \mathcal{A}$ with norm

$$\|Q\|_\infty = \sup_{(z,u)\in\mathcal{Z}\times\mathcal{A}} |Q(z,u)|.$$

The Bellman optimality operator is

$$(\mathcal{T}Q)(z,u) = c(z,u) + \gamma\mathbb{E}\left[\min_{v\in\mathcal{A}} Q(z^+, v) \mid z, u\right], \tag{5.1}$$

where $0 < \gamma < 1$ and z^+ is the next sampled state.

Assumption 1 (Bounded sampled problem)**.** *The sampled state set used for learning and evaluation is bounded by projection, the action set $\mathcal{A}$ is finite, and the one-step cost is bounded: $0 \le c(z,u) \le c_{\max} < \infty$ for all $(z,u) \in \mathcal{Z} \times \mathcal{A}$.*

Lemma 1 (Well-defined Bellman operator)**.** *Under Assumption 1, $\mathcal{T}$ maps $\mathcal{B}(\mathcal{Z} \times \mathcal{A})$ into itself. Moreover, every discounted cost-to-go is bounded above by $c_{\max}/(1-\gamma)$.*

Proof. Let $Q \in \mathcal{B}(\mathcal{Z}\times\mathcal{A})$. Since $\mathcal{A}$ is finite, the map $z^+ \mapsto \min_{v\in\mathcal{A}} Q(z^+, v)$ is measurable and bounded by $\|Q\|_\infty$. Therefore,

$$|(\mathcal{T}Q)(z,u)| \le c_{\max} + \gamma\|Q\|_\infty,$$

so $\mathcal{T}Q$ is bounded. For any policy, the discounted accumulated cost is bounded by $\sum_{k=0}^{\infty} \gamma^k c_{\max} = c_{\max}/(1-\gamma)$. □

Theorem 1 (Bellman contraction and optimality)**.** *Under Assumption 1, the Bellman operator* (5.1) *is a γ-contraction on $\mathcal{B}(\mathcal{Z}\times\mathcal{A})$ under the supremum norm. Consequently, it has a unique fixed point Q^*. Any greedy policy*

$$\mu^*(z) \in \arg\min_{u\in\mathcal{A}} Q^*(z,u)$$

is optimal for the sampled discounted problem (3.4)*.*

Proof. For bounded Q_1, Q_2 and any (z,u), the inequality

$$\left|\min_v a_v - \min_v b_v\right| \le \max_v |a_v - b_v|$$

gives

$$\begin{aligned}|\mathcal{T}Q_1(z,u) - \mathcal{T}Q_2(z,u)| &\le \gamma\mathbb{E}\left[\max_{v\in\mathcal{A}} |Q_1(z^+,v) - Q_2(z^+,v)| \mid z, u\right] \\ &\le \gamma\|Q_1 - Q_2\|_\infty.\end{aligned}$$

Taking the supremum over (z,u) proves contraction. The Banach fixed-point theorem gives existence and uniqueness of Q^*. The standard optimality theorem for discounted Markov decision processes then implies that every policy greedy with respect to Q^* is optimal [26, 27, 29]. □

Assumption 2 (Regularized regression well-posedness)**.** *For each mode i, let Φ_i be the matrix whose rows are the feature vectors $\phi(X_k, h_k, u_k)^T$ for samples satisfying $i_k = i$. Either the ridge parameter satisfies $\lambda > 0$, or $\lambda = 0$ and Φ_i has full column rank.*

Theorem 2 (Uniqueness of the fitted-Q update)**.** *Under Assumption 2, each mode-wise fitted-Q update in Algorithm 1 has a unique solution. If $\lambda > 0$, uniqueness holds even when Φ_i is rank deficient. The update depends only on observed states, fuzzy memberships, actions, modes, costs, and next states; it does not require the matrices $A_{\ell,i}$, $B_{\ell,i}$, $a_{\ell,i}$, or the Markov transition matrix.*

Proof. At a fixed fitted-Q iteration, the targets y_k are determined by the previous value approximation. For mode i, the update solves

$$\min_{\theta} \|\Phi_i\theta - y_i\|_2^2 + \lambda\|\theta\|_2^2.$$

The Hessian is $2(\Phi_i^T\Phi_i + \lambda I)$. If $\lambda > 0$, this matrix is positive definite for every Φ_i. If $\lambda = 0$, positive definiteness follows from the full-column-rank condition. Hence the objective is strictly convex and has a unique minimizer,

$$\theta_i = (\Phi_i^T\Phi_i + \lambda I)^{-1}\Phi_i^T y_i,$$

whenever the displayed inverse exists. All quantities in this expression are formed from the transition data and the selected feature map. No model matrices or transition probabilities enter the computation. □

Theorem 3 (Approximate fitted-Q error propagation)**.** *Let $Q^{(r)} \in \mathcal{B}(\mathcal{Z} \times \mathcal{A})$ be the fitted-Q approximation after iteration r. Suppose that for all $r \geq 0$,*

$$\|Q^{(r+1)} - \mathcal{T}Q^{(r)}\|_\infty \leq \varepsilon. \tag{5.2}$$

Then, for every integer $R \geq 1$,

$$\|Q^{(R)} - Q^*\|_\infty \leq \gamma^R\|Q^{(0)} - Q^*\|_\infty + \frac{1-\gamma^R}{1-\gamma}\varepsilon. \tag{5.3}$$

Proof. Using (5.2), the triangle inequality, Theorem 1, and $\mathcal{T}Q^* = Q^*$,

$$\begin{aligned}\|Q^{(r+1)} - Q^*\|_\infty &\leq \|Q^{(r+1)} - \mathcal{T}Q^{(r)}\|_\infty + \|\mathcal{T}Q^{(r)} - \mathcal{T}Q^*\|_\infty \\ &\leq \varepsilon + \gamma\|Q^{(r)} - Q^*\|_\infty.\end{aligned}$$

Iterating this scalar recursion from $r = 0$ to $R - 1$ gives (5.3). This is the standard perturbation argument behind approximate dynamic programming and fitted value iteration; finite-sample versions are developed in [18, 27, 30]. □

The next results provide the missing theory for the Lyapunov-penalized policy update. The idea is close in spirit to control-Lyapunov-function stabilization: one first specifies a bounded reference action that satisfies a one-step drift inequality, and then penalizes learned actions that move too far from this stabilizing action. Classical continuous-time control-Lyapunov constructions are due to Artstein and Sontag; see also the monograph treatment of Freeman and Kokotović [31–33]. In the present paper the action set is finite and the guarantee is a sampled-data practical tracking bound.

Assumption 3 (Bounded fitted-Q approximation)**.** *For the trained fitted-Q approximation and the compact projected learning domain, there is a constant $M_Q < \infty$ such that*

$$|Q_i(X, h, u; \theta_i)| \le M_Q$$

for all modes $i \in \mathcal{S}$, sampled states (X, h), and actions $u \in \mathcal{A}$.

Proposition 1 (Lyapunov-regularized policy comparison)**.** *Fix a sampled state $z = (X, h, i)$ and assume Assumption 3. Let*

$$\mu_i^Q(X, h) \in \arg\min_{u \in \mathcal{A}} Q_i(X, h, u; \theta_i)$$

and let $\mu_i^{\mathrm{LR}}(X, h)$ be defined by (4.4)*. Since $\mathcal{A} \subset [0, 1]^2$ and $\bar{u}(z) \in [0, 1]^2$, the following value-comparison bound holds:*

$$Q_i(X, h, \mu_i^{\mathrm{LR}}; \theta_i) \le Q_i(X, h, \mu_i^Q; \theta_i) + 2\lambda_L + 2\rho. \tag{5.4}$$

Moreover, let

$$\delta_{\mathcal{A}}(z) = \min_{u \in \mathcal{A}} \|u - \bar{u}(z)\|.$$

Then, for every $\lambda_L > 0$,

$$\|\mu_i^{\mathrm{LR}}(X, h) - \bar{u}(z)\|^2 \le \delta_{\mathcal{A}}(z)^2 + \frac{2M_Q + 2\rho}{\lambda_L}. \tag{5.5}$$

Proof. The optimality of μ_i^{LR} for the regularized objective gives

$$Q_i(z, \mu_i^{\mathrm{LR}}) + \lambda_L\|\mu_i^{\mathrm{LR}} - \bar{u}\|^2 + \rho\|\mu_i^{\mathrm{LR}}\|^2 \le Q_i(z, \mu_i^Q) + \lambda_L\|\mu_i^Q - \bar{u}\|^2 + \rho\|\mu_i^Q\|^2.$$

Dropping the two nonnegative terms on the left and using $\|\mu_i^Q - \bar{u}\|^2 \le 2$ and $\|\mu_i^Q\|^2 \le 2$ yields (5.4). For (5.5), choose $\hat{u}(z) \in \arg\min_{u \in \mathcal{A}} \|u - \bar{u}(z)\|$. Comparing the regularized objective at μ_i^{LR} with its value at $\hat{u}(z)$ gives

$$Q_i(z, \mu_i^{\mathrm{LR}}) + \lambda_L\|\mu_i^{\mathrm{LR}} - \bar{u}\|^2 + \rho\|\mu_i^{\mathrm{LR}}\|^2 \le Q_i(z, \hat{u}) + \lambda_L\delta_{\mathcal{A}}(z)^2 + \rho\|\hat{u}\|^2.$$

Since $|Q_i| \le M_Q$ and $\|\hat{u}\|^2 \le 2$, removing the nonnegative term $\rho\|\mu_i^{\mathrm{LR}}\|^2$ gives

$$\lambda_L\|\mu_i^{\mathrm{LR}} - \bar{u}\|^2 \le 2M_Q + \lambda_L\delta_{\mathcal{A}}(z)^2 + 2\rho,$$

which is (5.5). □

Assumption 4 (Reference-action drift condition)**.** *Let $V(z) = \|x - x_r\|^2$ be the squared tracking-error Lyapunov function. There exist constants $\alpha \in (0, 1)$, $b \ge 0$, and $L_u > 0$ such that the bounded reference action $\bar{u}$ in* (4.3) *satisfies*

$$\mathbb{E}[V(z^+(\bar{u})) \mid z] \le (1 - \alpha)V(z) + b, \tag{5.6}$$

and the expected one-step Lyapunov change is Lipschitz in the action:

$$\left|\mathbb{E}[V(z^+(u)) - V(z^+(v)) \mid z]\right| \le L_u\|u - v\| \tag{5.7}$$

for all $u, v \in \mathcal{A}$ and all sampled states z.

Theorem 4 (Practical tracking guarantee for the Lyapunov-regularized policy)**.** *Suppose Assumptions 3 and 4 hold. Let μ^{LR} be the Lyapunov-penalized policy* (4.4)*. Define the uniform grid-resolution error*

$$\bar{\delta}_{\mathcal{A}} = \sup_{z\in\mathcal{Z}} \min_{u\in\mathcal{A}} \|u - \bar{u}(z)\|$$

and

$$\Delta_L = \left(\bar{\delta}_{\mathcal{A}}^2 + \frac{2M_Q + 2\rho}{\lambda_L}\right)^{1/2}. \tag{5.8}$$

Then the closed-loop sampled system satisfies the drift inequality

$$\mathbb{E}[V(z_{k+1}) \mid z_k] \le (1-\alpha)V(z_k) + b + L_u\Delta_L. \tag{5.9}$$

Consequently,

$$\mathbb{E}V(z_k) \le (1-\alpha)^k V(z_0) + \frac{b + L_u\Delta_L}{\alpha}. \tag{5.10}$$

Thus the expected tracking error is practically bounded, and the ultimate bound improves when the action grid is refined and λ_L is increased, provided the bounded fitted-Q approximation is kept fixed.

Proof. By Proposition 1, the Lyapunov-regularized policy satisfies

$$\|\mu^{\mathrm{LR}}(z) - \bar{u}(z)\| \le \Delta_L$$

uniformly over the sampled domain. Using (5.7) with $u = \mu^{\mathrm{LR}}(z)$ and $v = \bar{u}(z)$ gives

$$\mathbb{E}[V(z^+(\mu^{\mathrm{LR}}(z))) \mid z] \le \mathbb{E}[V(z^+(\bar{u}(z))) \mid z] + L_u\Delta_L.$$

Combining this inequality with the reference-action drift condition (5.6) proves (5.9). Taking total expectations and iterating the scalar recursion

$$a_{k+1} \le (1-\alpha)a_k + b + L_u\Delta_L, \qquad a_k = \mathbb{E}V(z_k),$$

yields (5.10). □

Remark 2. *The theorem above explains the role of the Lyapunov penalty. Pure fitted-Q greediness controls only the approximate value objective. The additional term in* (4.4) *provides a quantitative link to the stabilizing reference action through* (5.5)*, which then yields the practical tracking estimate* (5.10)*. The result is conditional because the drift condition* (5.6) *is a property of the chosen reference action and simulator, not a consequence of regression alone. Numerical verification of Assumption 4 for the specific biological simulator used in Section 6 is reported in Section 6.5.*

Proposition 2 (Positivity under projected inputs and states)**.** *Assume that the continuous biological vector field is locally Lipschitz and satisfies the inward-boundary condition*

$$f_j(x,u,i) \ge 0 \quad \textit{whenever } x_j = 0,$$

for every component j, every mode i, and every admissible input $u \in [0,1]^2$. Then the positive orthant is invariant for the continuous biological dynamics. In the implemented sampled simulator, the additional componentwise projection

$$x_{k+1} \leftarrow \min\{2, \max\{0, x_{k+1}\}\}$$

ensures $x_k \in [0,2]^4$ for all k whenever $x_0 \in [0,2]^4$.

Proof. The continuous-time claim is the standard positive-invariance criterion for ordinary differential equations: If the vector field points into, or is tangent to, the closed positive orthant at every boundary face, then trajectories starting in the orthant cannot leave it. This follows from Nagumo's tangent-cone theorem for invariant sets [34]; see also the positive-systems treatment in [35]. The sampled claim follows immediately from the explicit projection step, which maps every tentative update into the box $[0, 2]^4$ component by component. □

6. Data-calibrated numerical experiment

6.1. Data source and preprocessing

The numerical study uses the public data file from the ModelingMDSCs repository, available at [36]. The repository identifies the data as coming from the FIR atezolizumab non-small-cell lung cancer (NSCLC) study and as data used in subsequent modelling work [22, 37, 38]. The computational workflow downloads the Excel file directly from the public repository, parses repeated tumour blocks, and converts longest lesion diameter LD to tumour volume using

$$V = 0.5LD^3. \tag{6.1}$$

Each lesion trajectory is normalized by its initial volume. The parsed data contain 52 observations from six target-lesion trajectories.

To avoid reporting all calibration statistics as in-sample, the six trajectories are split at the trajectory level into a calibration set and a held-out set before any simulator parameters are chosen. Four trajectories (32 observations) form the calibration set; two trajectories (20 observations) form the held-out set. Simulator parameters are set using the calibration set only. Table 1 summarizes the split-level growth-rate statistics.

Table 1. Trajectory-level calibration and holdout split summary.

Split	Trajectories	Observations	Median rate	Positive median	IQR
Calibration	4	32	0.002401	0.004147	0.006604
Held-out	2	20	−0.003516	−0.002499	0.001017

Note: The simulator parameters r_σ are set using only the calibration-set statistics. Growth rates are log-linear rates per day fitted to normalized tumour-volume trajectories.

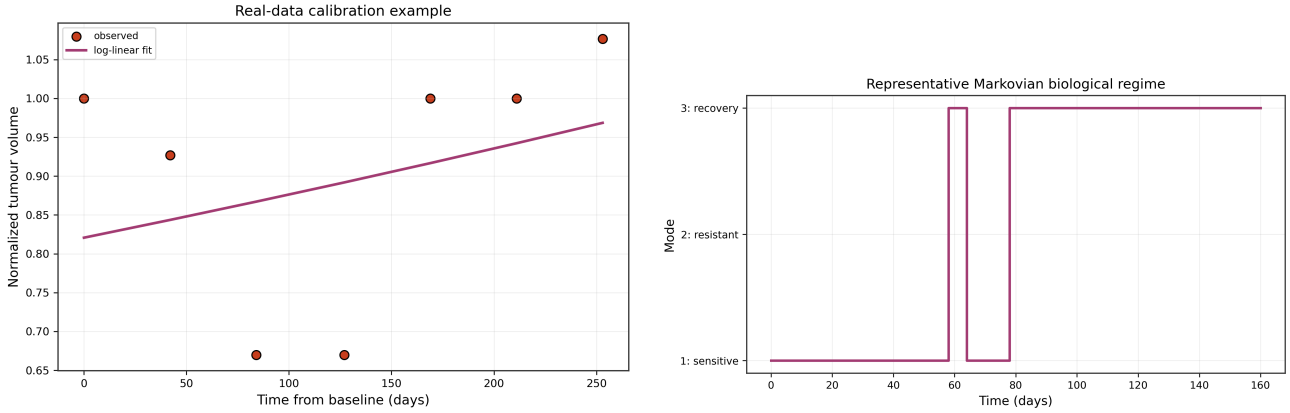
The fitted log-linear calibration on the calibration set produced a calibration-set median growth rate of 0.002401 per day, a positive-trajectory median of 0.004147 per day, and an interquartile range of 0.006604 per day. The held-out trajectories have a more negative median (−0.003516 per day), consistent with stronger treatment response, and lie within the range of rates covered by the simulator. Table 2 summarizes the four trajectories used for calibration.

Table 2. Calibration-set tumour-volume summary.

Trajectory	Points	Days	Growth rate	Normalized-volume fit RMSE	Final normalized volume
Tumour 1	9	334	−0.009143	0.2094	0.0698
Tumour 4	7	253	0.000655	0.1460	1.0769
Tumour 5	11	461	0.004147	0.4974	5.0379
Tumour 6	5	167	0.006794	0.3548	3.4541

Note: Growth rates are obtained from log-linear fits to normalized tumour-volume trajectories. Normalized-volume fit RMSE denotes the root mean squared residual between the observed and fitted values on the normalized tumour-volume scale. Two additional held-out trajectories (Tumour 2 and Tumour 3) are not used to set the simulator parameters.

Figure 1 summarizes the real-data calibration and the realized Markovian regime sequence. Panel (a) shows the corresponding normalized tumour-volume observations and the representative log-linear calibration used to set plausible growth ranges.



(a) Real-data calibration.

(b) Markov biological regimes.

Figure 1. Representative real-data calibration using normalized tumour-volume observations and a log-linear fit, together with the realized Markov biological regimes during the evaluation run.

6.2. Simulator and learning settings

The normalized tumour–immune simulator is

$$\dot{T} = r_\sigma T\left(1 - \frac{T}{K_\sigma}\right) - \alpha_\sigma \frac{ET}{g_\sigma + T} - \beta_\sigma u_c T, \tag{6.2}$$

$$\dot{E} = s_\sigma + \rho_\sigma \frac{ET}{g_\sigma + T} - d_\sigma E + \gamma_\sigma u_i E - \eta_\sigma u_c E, \tag{6.3}$$

$$\dot{C} = -\lambda_c C + u_c, \tag{6.4}$$

$$\dot{I} = -\lambda_i I + u_i. \tag{6.5}$$

Figure 2 shows the main feedback and interaction pathways represented by the nonlinear simulator. The diagram separates biological state interactions, dimensionless treatment-like inputs, and the Markov and fuzzy mechanisms that modulate the active dynamics.

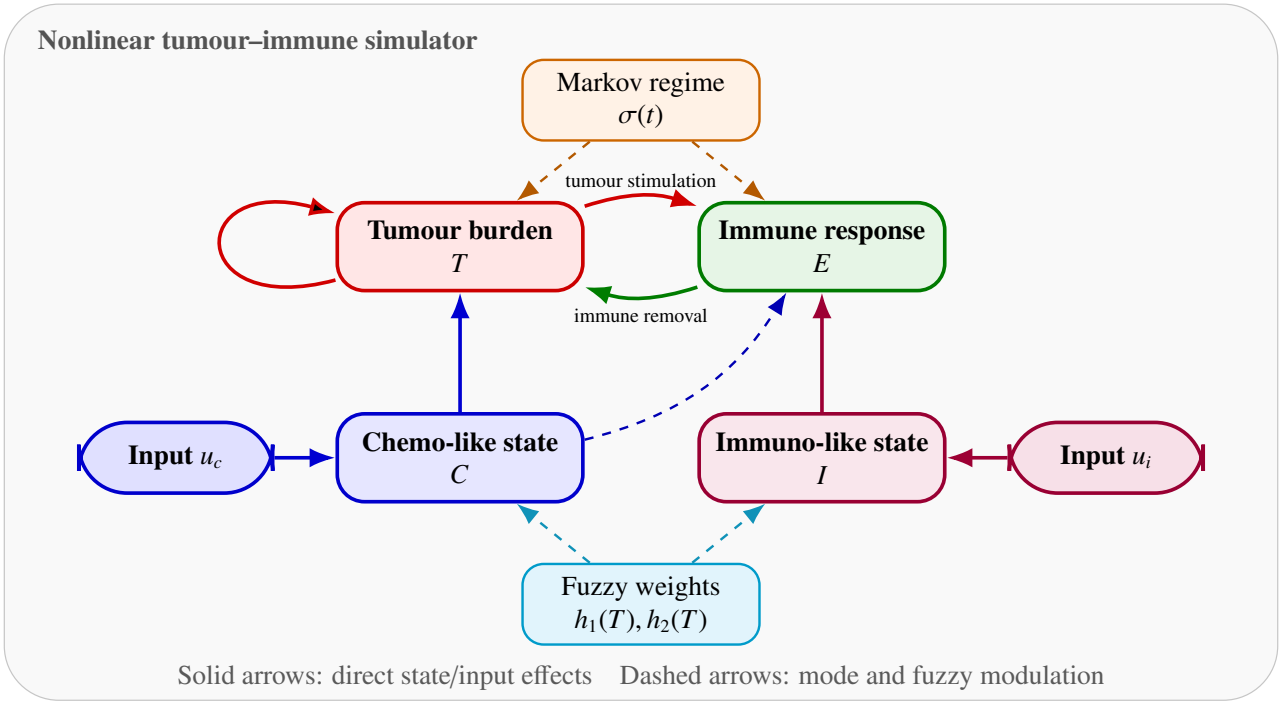


Figure 2. Flow diagram of the nonlinear tumour–immune simulator. Solid arrows denote direct biological or input effects; dashed arrows denote regime-dependent and fuzzy-weight modulation. The variables u_c and u_i are dimensionless computational inputs only.

In these equations, r_σ is the tumour growth rate, K_σ is the carrying capacity, α_σ is the immune-mediated tumour-removal coefficient, g_σ is a half-saturation constant, β_σ is the chemotherapy-like tumour-effect coefficient, s_σ is immune recruitment, ρ_σ is tumour-stimulated immune growth, d_σ is immune decay, γ_σ is immunotherapy-like immune stimulation, η_σ is chemotherapy-like immune suppression, and λ_c, λ_i are decay rates for C and I. The subscript σ means that the parameter value depends on the active Markov mode. The T–S fuzzy membership function is a two-rule logistic partition based on tumour burden:

$$h_2(T) = \frac{1}{1 + \exp[-8(T - 0.55)]}, \qquad h_1(T) = 1 - h_2(T). \tag{6.6}$$

The Markov transition matrix used by the simulator, but not supplied to the controller, is

$$P = \begin{bmatrix} 0.965 & 0.025 & 0.010 \\ 0.035 & 0.940 & 0.025 \\ 0.025 & 0.020 & 0.955 \end{bmatrix}. \tag{6.7}$$

Panel (b) of Figure 1 shows the realized Markovian regime sequence during evaluation, illustrating that the controller is tested under switching biological conditions rather than a fixed deterministic mode. The experiment used $\Delta t = 2$ days, training horizon 60 days, evaluation horizon 160 days, 18 previously collected training episodes, 5 fitted Q iterations, $\gamma = 0.98$, ridge parameter 10^{-4}, and a 7×7 action grid on $[0, 1]^2$. Thus 540 previously collected transitions were generated for learning. The active-treatment

regime was formed from the positive-growth subset of the real data because the public observations are treatment-response data rather than untreated natural-history data.

The learning residuals in Figure 3 decrease over fitted Q iterations and provide a numerical check that the fitted regressions stabilize before policy evaluation.

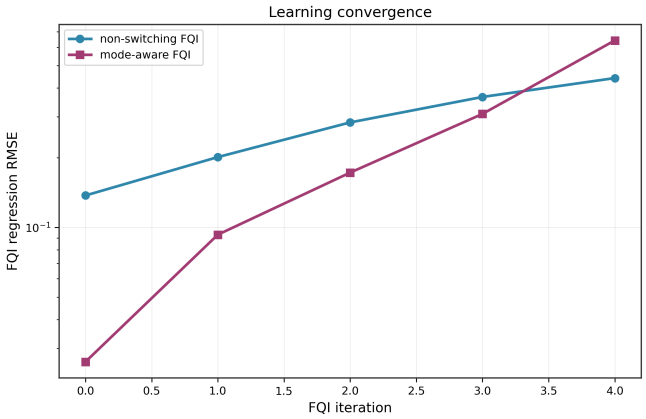
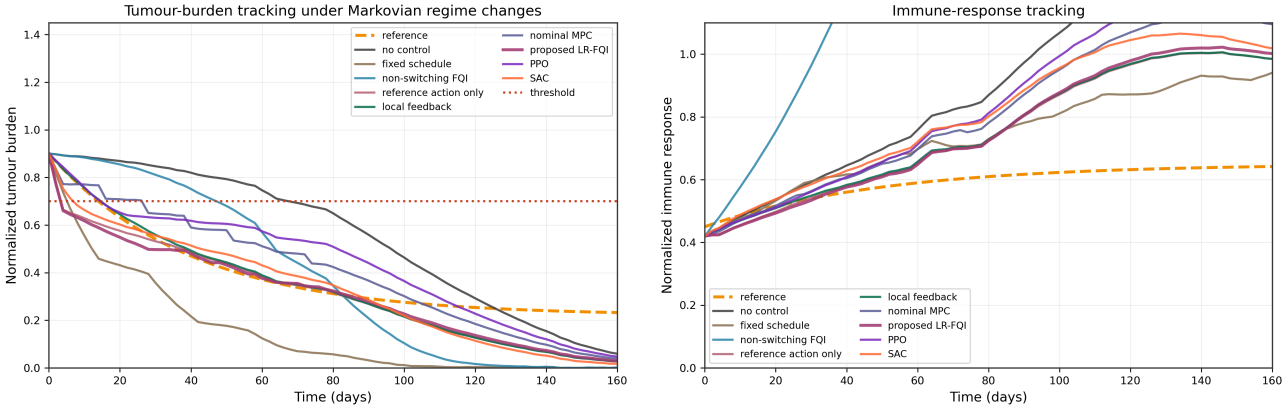


Figure 3. Fitted-Q regression RMSE over successive iterations for the mode-dependent and non-switching fitted-Q learners. At each iteration, the RMSE is computed between the fitted Q-function predictions and the corresponding Bellman targets over the training transitions.

6.3. Quantitative results

Figure 4 compares tumour-burden tracking and the corresponding immune-response trajectories across all compared strategies. The proposed Lyapunov-penalized mode-dependent fitted Q controller reaches the target region rapidly while avoiding the prolonged threshold exceedance observed under no control and the less efficient behaviour observed under the alternative active controllers. The immune-response comparison is important because a low tumour burden alone is not sufficient for acceptable tracking if the immune state is driven far from the reference trajectory.

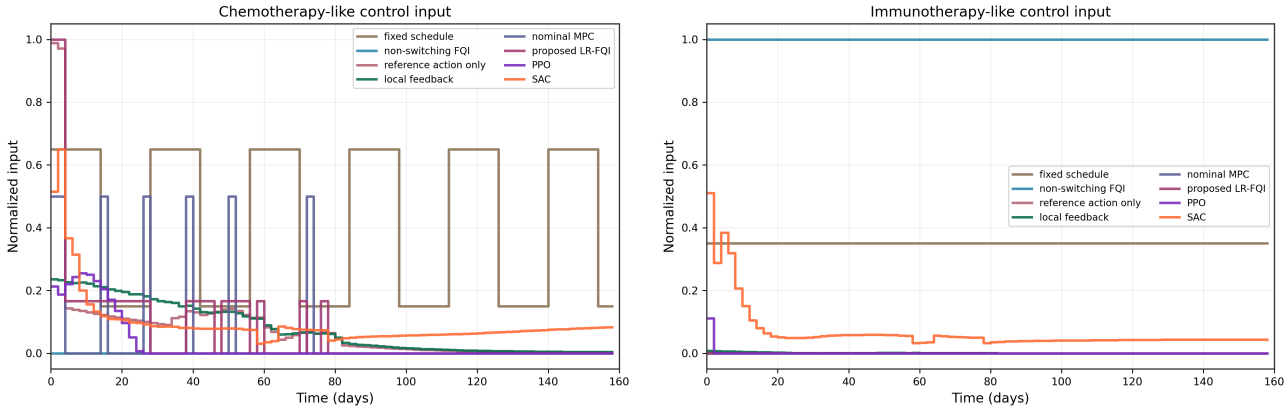


(a) Tumour-burden tracking.

(b) Immune-response trajectories.

Figure 4. Tumour-burden and immune-response trajectories under no control, fixed scheduling, non-switching fitted Q iteration, reference action, local feedback, nominal MPC, PPO, SAC, and proposed Lyapunov-penalized mode-dependent fitted Q iteration.

The dimensionless input profiles are shown in Figure 5. The proposed controller uses sparse chemotherapy-like input and avoids the saturated immunotherapy-like behaviour selected by the non-switching fitted Q controller.



(a) Chemotherapy-like input. **(b)** Immunotherapy-like input.

Figure 5. Dimensionless chemotherapy-like and immunotherapy-like inputs for all active-control strategies.

The aggregate performance metrics are shown in Table 3. Here, ISE (integral squared error) is approximated by summing the squared tracking errors over the sampled time steps and multiplying by Δt, that is,

$$\mathrm{ISE} = \sum_{k} (x_k - x_{r,k})^2 \Delta t.$$

The representative single-seed comparison should be interpreted as a balanced-control result rather than as uniform dominance on every individual metric. The reference-action baseline gives a slightly smaller weighted cost in this single realization, while local feedback gives a smaller tumour ISE. However, the proposed LR-FQI remains among the best-performing active controllers and achieves a short threshold-exceedance time of 4 days with substantially lower total input than fixed scheduling and non-switching FQI. Compared with PPO, SAC, nominal MPC, no control, and non-switching FQI, the proposed controller gives lower weighted cost and shorter threshold-exceedance time in the representative run. This supports the role of Lyapunov-regularized fitted Q learning as a balanced data-driven controller, rather than a controller that optimizes every metric separately. PPO and SAC achieved lower total input (4.75 and 25.89 respectively) but higher weighted cost (76.32 and 72.30) and higher tumour ISE (2.82 and 1.92). Nominal MPC produced the highest threshold-exceedance time (28 days) among active controllers. Relative to non-switching fitted Q iteration, the proposed controller reduced the weighted cost from 504.326 to 51.170, reduced tumour ISE from 6.825 to 1.699, reduced time above the tumour threshold from 48 to 4 days, and reduced total dimensionless input from 160.0 to 11.667.

Table 3. Performance metrics for the active-treatment simulation regime.

Controller	Cost	Tumour ISE	Immune ISE	Input	Final T	Time $T > 0.70$
Proposed LR-FQI	51.170	1.699	8.498	11.667	0.030	4.0
Reference action	50.785	1.733	7.887	12.772	0.028	4.0
Local feedback	53.930	1.487	7.985	12.844	0.028	16.0
Nominal MPC	59.632	2.098	13.890	7.000	0.039	28.0
SAC	72.301	1.917	12.091	25.885	0.017	8.0
PPO	76.322	2.824	18.111	4.754	0.048	16.0
No control	201.099	9.759	26.280	0.000	0.060	68.0
Fixed schedule	255.894	9.037	5.052	122.000	0.002	6.0
Non-switching FQI	504.326	6.825	211.895	160.000	0.001	48.0

Note: ISE denotes integral squared error, $\sum_k e_k^2 \Delta t$. The reported results correspond to the representative single-seed run. Input values are normalized computational quantities and do not represent clinical doses. Threshold-exceedance time is reported in days.

The no-control baseline eventually produced a small final tumour value because the simulator still contains immune-mediated tumour suppression. However, it remained above the tumour threshold for 68 days and had a higher tumour ISE than the proposed controller. Fixed scheduling reduced the threshold time to 6 days but required 122 units of dimensionless input and produced a larger weighted cost. Non-switching fitted Q iteration selected saturated immunotherapy-like input, which resulted in a large immune tracking error. In contrast, the proposed Lyapunov-penalized controller used sparse chemotherapy-like input, avoided immunotherapy saturation, and achieved one of the strongest balanced profiles across weighted cost, tumour tracking, threshold-exceedance time, immune-state behaviour, and input use.

6.4. Monte Carlo evaluation

To assess robustness under varying Markov transition sequences and process-noise realizations, all nine controllers are evaluated over 50 independent random seeds, each producing an independent realization of the Markov chain and the process noise over a 160-day evaluation horizon. The initial state and mode are fixed across seeds. Table 4 reports the mean and standard deviation of the weighted cost, tumour ISE, and threshold-exceedance time over the 50 seeds.

Table 5 shows the Monte Carlo comparison in tabular form. The Monte Carlo results provide the main robustness evidence. The proposed LR-FQI achieves the lowest mean weighted cost (47.29 ± 19.91), while the reference-action baseline is very close (47.88 ± 20.63). Both controllers achieve the shortest and most consistent threshold-exceedance time (4.00 ± 0.00 days). Local feedback gives the smallest mean tumour ISE, but it keeps the tumour above the threshold for substantially longer (16.32 ± 1.30 days). SAC and nominal MPC have higher mean costs and longer threshold-exceedance times, and PPO remains clearly weaker under the present 15000-timestep training budget. Thus, the expanded comparison supports the conclusion that LR-FQI provides the best overall balance across the primary weighted objective, threshold control, tumour tracking, and input moderation.

Table 4. Monte Carlo evaluation over 50 independent Markov and noise trajectories.

Controller	Cost (mean±SD)	Tumour ISE (mean±SD)	Time $T > 0.70$ (mean±SD)
Proposed LR-FQI	47.29 ± 19.91	0.74 ± 0.65	4.00 ± 0.00
Reference action	47.88 ± 20.63	0.77 ± 0.76	4.00 ± 0.00
Local feedback	52.12 ± 23.58	0.63 ± 0.66	16.32 ± 1.30
SAC	59.92 ± 24.28	1.22 ± 0.81	8.28 ± 1.28
Nominal MPC	63.94 ± 32.13	2.96 ± 1.81	27.84 ± 5.49
PPO	89.45 ± 37.84	4.67 ± 2.60	15.92 ± 1.14
Fixed schedule	214.00 ± 41.68	7.02 ± 2.87	6.00 ± 0.00
No control	439.61 ± 386.67	24.88 ± 20.99	96.68 ± 41.93
Non-switching FQI	521.38 ± 169.00	12.99 ± 12.73	65.48 ± 33.42

Note: Values are reported as mean ± standard deviation over 50 independent realizations. ISE denotes integral squared error. Threshold-exceedance time, defined as the time for which $T > 0.70$, is reported in days.

Table 5. Monte Carlo controller comparison over 50 independent Markov and noise trajectories.

Controller	Cost mean	Cost SD	Tumour ISE mean	Tumour ISE SD	Time $T > 0.70$ mean	Input mean	Final T mean
Proposed LR-FQI	47.292	19.910	0.742	0.650	4.000	29.680	0.180
Reference action	47.882	20.628	0.772	0.758	4.000	32.590	0.166
Local feedback	52.124	23.579	0.632	0.657	16.320	30.676	0.178
SAC	59.922	24.283	1.215	0.815	8.280	35.683	0.206
Nominal MPC	63.935	32.133	2.964	1.812	27.840	14.240	0.291
PPO	89.448	37.837	4.670	2.601	15.920	13.294	0.360
Fixed schedule	214.001	41.682	7.019	2.870	6.000	122.000	0.022
No control	439.614	386.669	24.875	20.989	96.680	0.000	0.498
Non-switching FQI	521.385	169.001	12.991	12.728	65.480	160.240	0.117

Note: Results are based on 50 independent realizations. Means and standard deviations (SDs) are reported separately for cost and tumour ISE; the remaining columns report means only. ISE denotes integral squared error. Threshold-exceedance time, defined as the time for which $T > 0.70$, is reported in days. Input values are normalized computational quantities and do not represent clinical doses.

6.5. Numerical verification of the reference-action drift condition

Assumption 4 requires that the reference action $\bar{u}$ satisfies an expected one-step Lyapunov drift inequality $\mathbb{E}[V(z^+(\bar{u})) \mid z] \leq (1-\alpha)V(z) + b$ on the simulation domain, and that the one-step Lyapunov change is Lipschitz in the action with constant L_u. We verify both numerically on the compact domain $\mathcal{D} = \{T \in [0, 1.20], E \in [0.10, 1.00], C \in [0, 1.00], I \in [0, 1.00], i \in \{1, 2, 3\}\}$, which contains all trajectories observed during evaluation.

For the drift condition, we draw 5000 states uniformly from $\mathcal{D}$ and, for each state, estimate $\mathbb{E}[V(z^+(\bar{u})) \mid z]$ by averaging over 20 independent noise realizations. Table 6 reports, for each value of α, the smallest b such that $\mathbb{E}[V(z^+(\bar{u}))] \leq (1-\alpha)V(z) + b$ holds for all 5000 samples.

The drift condition holds for all $\alpha \in \{0.005, 0.01, 0.02, 0.05, 0.10\}$ with $b \approx 4.0$. The fraction of points satisfying the condition with $b = 0$ is approximately 19%, indicating that the reference action does not universally decrease V in one step but does satisfy the weaker practical stability form used in Theorem 4. For the Lipschitz condition, 2500 random action pairs are sampled from $\mathcal{A}$ and the

deterministic one-step ratio $|V(z^+(u)) - V(z^+(v))|/\|u - v\|$ is computed. The estimated maximum is $L_u^{\max} = 9.516$ and the 95th percentile is $L_u^{(95)} = 6.982$. Using $\alpha = 0.01$, $b = 4.024$, and $L_u = 9.516$ in (5.10) yields a finite practical bound, confirming that Assumption 4 is numerically supported on the stated compact domain. The constants are not claimed to hold globally outside this domain.

Table 6. Numerical verification of the reference-action drift condition.

α	Minimum required b	Fraction satisfying $b = 0$	Residual q_{95}
0.005	4.023	0.198	3.899
0.010	4.024	0.197	3.902
0.020	4.026	0.195	3.906
0.050	4.041	0.187	3.921
0.100	4.077	0.172	3.947

Note: Assumption 4 requires

$$\mathbb{E}[V(z^+(\bar{u})) \mid z] \le (1 - \alpha)V(z) + b.$$

For each value of α, the table reports the smallest b for which this inequality holds over all 5000 sampled domain points. The residual is defined as

$$R(z) = \mathbb{E}[V(z^+(\bar{u})) \mid z] - (1 - \alpha)V(z),$$

and q_{95} denotes its 95th percentile. The sampled domain is $T \in [0, 1.20]$, $E \in [0.10, 1.00]$, $C, I \in [0, 1.00]$, and $\sigma \in \{1, 2, 3\}$. Each conditional expectation is estimated using 20 noise realizations per sampled point.

7. Conclusions and discussion

The results support four conclusions. First, public tumour-volume data can calibrate a reproducible T–S fuzzy Markov jump simulator, but treatment-response data must be used carefully. The positive-growth subset is used to construct an active-treatment simulation regime because the median observed response is negative; a trajectory-level holdout split confirms that the calibration statistics are not unduly influenced by a particular subset. Second, purely greedy fitted Q iteration can produce undesirable policies, such as saturated immunotherapy-like input in the non-switching comparison. Third, Lyapunov regularization improves policy quality in this simulator by combining data-driven Q-estimation with a stabilizing reference action, leading to bounded and more interpretable dimensionless inputs. Fourth, the balanced performance of the proposed method is supported across 50 Monte Carlo seeds, and Assumption 4 is numerically supported over 5000 sampled domain points.

This paper presented a Lyapunov-penalized data-driven adaptive tracking-control framework for data-calibrated T–S fuzzy Markov jump biological systems with unknown dynamics and unknown transition probabilities. The proposed controller learns mode-dependent fitted Q-functions from previously collected transition data and uses a Lyapunov-penalized policy update step to obtain bounded and interpretable dimensionless inputs. The theoretical analysis established Bellman contraction, uniqueness of the fitted update, an approximate fitted Q error bound, quantitative regularized-policy bounds, a practical tracking guarantee for the Lyapunov-regularized policy, and positivity preservation under projection. The numerical study used 32 public tumour-volume observations from four calibration-set target-lesion trajectories to calibrate plausible growth ranges in the simulator. In the active-treatment Markov regime, the expanded comparison shows that the proposed LR-FQI controller provides the best overall balanced performance among nine reported

controllers. In the 50-seed Monte Carlo evaluation, it achieves the lowest mean weighted cost (47.29 ± 19.91) and the shortest threshold-exceedance time jointly with the reference-action baseline (4.00 ± 0.00 days). Some individual metrics favour other controllers; for example, local feedback gives a slightly smaller mean tumour ISE and nominal MPC uses less input, but those gains come with worse threshold-control or weighted-cost performance. The results therefore support a balanced-performance claim rather than the stronger claim that LR-FQI dominates every baseline on every metric.

The main limitations are as follows. (i) The drift constant $b \approx 4.0$ in Assumption 4 is empirically estimated on a compact sampled domain and is not analytically derived; the practical tracking bound in Theorem 4 is therefore an in-simulation practical guarantee rather than a global stability certificate. (ii) PPO and SAC are trained for only 15000 timesteps, which produces suboptimal but transparent deep-RL baselines; longer training could improve their performance at the expense of greater computation. (iii) The biological parameters are fixed in-silico quantities and do not represent any individual patient. (iv) The framework does not address hidden or partially observable Markov modes, which are common in clinical settings.

The proposed framework is mathematical and computational. It does not provide clinical treatment recommendations, patient-specific dosing rules, or medical decision support. The dimensionless inputs u_c and u_i are simulation variables only. The real data are used to calibrate tumour-growth ranges, not to validate an RL dosing policy. The immune state, Markov mode, treatment response, and control constraints are simplified for reproducible analysis; they should not be interpreted as patient-specific pharmacokinetic/pharmacodynamic variables. Clinical translation would require pharmacokinetic/pharmacodynamic modelling, toxicity constraints, patient-specific uncertainty quantification, hidden-state estimation, and ethically approved prospective validation. Future work should address hidden Markov modes, partial observability of immune variables, safe exploration, patient-specific uncertainty, and pharmacokinetic/pharmacodynamic constraints.

Author contributions

Shirali Kadyrov: Conceptualization, methodology, software, writing–original draft. Ardak Kashkynbayev: Methodology, supervision, validation, writing–review and editing. Yershat Sapazhanov: Formal analysis, validation, writing–review and editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

An AI assistant was used to support LaTeX formatting, template conversion, and language-editing checks. The authors reviewed and approved all scientific content.

Acknowledgments

The authors acknowledge the public data resources cited in this manuscript.

Funding

This work is supported by a grant from the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of the project BR24993094.

Data availability statement

The tumour-volume data used for calibration are publicly available from the ModelingMDSCs repository cited in the references. The numerical results in this manuscript are generated from the stated preprocessing, simulator, action grid, fitted Q settings, and evaluation horizon.

Conflict of interest

The authors declare no conflicts of interest.

References

1. V. A. Kuznetsov, I. A. Makalkin, M. A. Taylor, A. S. Perelson, Nonlinear dynamics of immunogenic tumors: parameter estimation and global bifurcation analysis, *Bull. Math. Biol.*, **56** (1994), 295–321. https://doi.org/10.1007/BF02460644
2. D. Kirschner, J. C. Panetta, Modeling immunotherapy of the tumor–immune interaction, *J. Math. Biol.*, **37** (1998), 235–252. https://doi.org/10.1007/s002850050127
3. R. S. Sutton, A. G. Barto, *Reinforcement learning: an introduction*, MIT Press, 1998.
4. S. He, M. Zhang, H. Fang, F. Liu, X. Luan, Z. Ding, Reinforcement learning and adaptive optimization of a class of Markov jump systems with completely unknown dynamic information, *Neural Comput. Appl.*, **32** (2020), 14311–14320. https://doi.org/10.1007/s00521-019-04180-2
5. H. Fang, Y. Tu, H. Wang, S. He, F. Liu, Z. Ding, et al., Fuzzy-based adaptive optimization of unknown discrete-time nonlinear Markov jump systems with off-policy reinforcement learning, *IEEE Trans. Fuzzy Syst.*, **30** (2022), 5276–5290. https://doi.org/10.1109/TFUZZ.2022.3171844
6. X. Shi, Y. Li, C. Du, C. Chen, G. Zong, W. Gui, Reinforcement learning-based optimal control for Markov jump systems with completely unknown dynamics, *Automatica*, **171** (2025), 111886. https://doi.org/10.1016/j.automatica.2024.111886
7. Y. Lou, M. Luo, J. Cheng, X. Wang, K. Shi, Double-quantized-based H_∞ tracking control of T–S fuzzy semi-Markovian jump systems with adaptive event-triggered, *AIMS Math.*, **8** (2023), 6942–6969. https://doi.org/10.3934/math.2023351
8. G. Zong, H. Xie, D. Yang, X. Zhao, Y. Yi, Adaptive fuzzy tracking control for switched nonlinear systems under FDI attacks and input saturation: a flexible transient performance approach, *IEEE Trans. Cybern.*, **54** (2024), 7479–7488. https://doi.org/10.1109/TCYB.2024.3463689
9. Y. Wang, G. Zong, Dynamic event-triggered adaptive fixed-time practical tracking control for nonlinear systems through funnel function, *IEEE Trans. Autom. Sci. Eng.*, **22** (2025), 7008–7017. https://doi.org/10.1109/TASE.2024.3458176

10. Z. Gao, Y. Wang, Z. Ji, Finite-time H_∞ control of switched affine non-linear systems based on the Lie derivative and iteration technique, *IET Control Theory Appl.*, **13** (2019), 3148–3154. https://doi.org/10.1049/iet-cta.2018.5902

11. M. Sharifi, A. A. Jamshidi, N. N. Sarvestani, An adaptive robust control strategy in a cancer tumor–immune system under uncertainties, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **16** (2019), 865–873. https://doi.org/10.1109/TCBB.2018.2803175

12. H. Jiao, Q. Shen, Y. Shi, P. Shi, Adaptive tracking control for uncertain cancer–tumor–immune systems, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **18** (2021), 2753–2758. https://doi.org/10.1109/TCBB.2020.3036069

13. E. Ahmadi, J. Zarei, R. Razavi-Far, M. Saif, A dual approach for positive T–S fuzzy controller design and its application to cancer treatment under immunotherapy and chemotherapy, *Biomed. Signal Process. Control*, **58** (2020), 101822. https://doi.org/10.1016/j.bspc.2019.101822

14. A. Kashkynbayev, R. Rakkiyappan, Sampled-data output tracking control based on T–S fuzzy model for cancer–tumor–immune systems, *Commun. Nonlinear Sci. Numer. Simul.*, **128** (2024), 107642. https://doi.org/10.1016/j.cnsns.2023.107642

15. T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, *IEEE Trans. Syst., Man, Cybern.*, **SMC-15** (1985), 116–132. https://doi.org/10.1109/TSMC.1985.6313399

16. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, *arXiv Preprint*, 2017. https://doi.org/10.48550/arXiv.1707.06347

17. T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor–critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor, *Proceedings of the 35th International Conference on Machine Learning*, **80** (2018), 1861–1870.

18. D. Ernst, P. Geurts, L. Wehenkel, Tree-based batch mode reinforcement learning, *J. Mach. Learn. Res.*, **6** (2005), 503–556.

19. X. Liu, Q. Li, J. Pan, A deterministic and stochastic model for the system dynamics of tumor–immune responses to chemotherapy, *Phys. A: Stat. Mech. Appl.*, **500** (2018), 162–176. https://doi.org/10.1016/j.physa.2018.02.118

20. M. Dassow, S. Djouadi, K. Moussa, Optimal control of a tumor–immune system with a modified Stepanova cancer model, *IFAC-PapersOnLine* **54** (2021), 227–232. https://doi.org/10.1016/j.ifacol.2021.10.260

21. M. Farman, A. Ahmad, A. Akgül, M. U. Saleem, K. S. Nisar, V. Vijayakumar, Dynamical behavior of tumor–immune system with fractal-fractional operator, *AIMS Math.*, **7** (2022), 8751–8773. https://doi.org/10.3934/math.2022489

22. N. Ghaffari Laleh, C. M. L. Loeffler, J. Grajek, K. Staňková, A. T. Pearson, H. S. Muti, et al., Classical mathematical models for prediction of response to chemotherapy and immunotherapy, *PLOS Comput. Biol.*, **18** (2022), e1009822. https://doi.org/10.1371/journal.pcbi.1009822

23. L. A. Zadeh, Fuzzy sets, *Inf. Control*, **8** (1965), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

24. O. L. V. Costa, R. P. Marques, M. D. Fragoso, *Discrete-time Markov jump linear systems*, London: Springer, 2005. https://doi.org/10.1007/b138575
25. O. L. V. Costa, M. D. Fragoso, M. G. Todorov, *Continuous-time Markov jump linear systems*, Springer Science & Business Media, 2013. https://doi.org/10.1007/978-3-642-34100-7
26. R. Bellman, *Dynamic programming*, Princeton: Princeton University Press, 1957.
27. D. P. Bertsekas, Neuro-dynamic programming, In: P. M. Pardalos, O. A. Prokopyev, *Encyclopedia of optimization*, Springer, 2025. https://doi.org/10.1007/978-3-030-54621-2_440-1
28. P. J. Werbos, Approximate dynamic programming for real-time control and neural modeling, In: D. A. White, D. A. Sofge, *Handbook of intelligent control: neural, fuzzy, and adaptive approaches*, Van Nostrand Reinhold, 1992.
29. M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, New York: John Wiley & Sons, 2014.
30. R. Munos, C. Szepesvári, Finite-time bounds for fitted value iteration, *J. Mach. Learn. Res.*, **9** (2008), 815–857.
31. Z. Artstein, Stabilization with relaxed controls, *Nonlinear Anal.: Theory, Methods Appl.*, **7** (1983), 1163–1173. https://doi.org/10.1016/0362-546X(83)90049-4
32. E. D. Sontag, A universal construction of Artstein's theorem on nonlinear stabilization, *Syst. Control Lett.*, **13** (1989), 117–123. https://doi.org/10.1016/0167-6911(89)90028-5
33. R. A. Freeman, P. V. Kokotović, *Robust nonlinear control design: state-space and Lyapunov techniques*, Springer Science & Business Media, 2008.
34. M. Nagumo, Über die lage der integralkurven gewöhnlicher differentialgleichungen, *Proceedings of the Physico-Mathematical Society of Japan, 3rd Series*, **24** (1942), 551–559. https://doi.org/10.11429/ppmsj1919.24.0_551
35. W. M. Haddad, V. Chellaboina, Q. Hui, *Nonnegative and compartmental dynamical systems*, Princeton: Princeton University Press, 2010. https://doi.org/10.1515/9781400832248
36. A. L. MacLean, E. T. Roussos Torres, J. Kreger, ModelingMDSCs: tumour volume data from the FIR NSCLC atezolizumab study, GitHub repository, 2023. Available from: `https://github.com/maclean-lab/ModelingMDSCs`.
37. D. R. Spigel, J. E. Chaft, S. Gettinger, B. H. Chao, L. Dirix, P. Schmid, et al., FIR: efficacy, safety, and biomarker analysis of a phase II open-label study of atezolizumab in PD-L1-selected patients with NSCLC, *J. Thorac. Oncol.*, **13** (2018), 1733–1742. https://doi.org/10.1016/j.jtho.2018.05.004
38. J. Kreger, E. T. Roussos Torres, A. L. MacLean, Myeloid-derived suppressor-cell dynamics control outcomes in the metastatic niche, *Cancer Immunol. Res.*, **11** (2023), 614–628. https://doi.org/10.1158/2326-6066.CIR-22-0617