Applied Mathematical Modelling 151 (2026) 116527

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

Optimizing HIV antiretroviral therapy strategies: A multi-scale, multi-objective framework for reconciling individual–population trade-offs

Yuyi Xue [a], Mingwang Shen [b,c,d,e], Stacey R. Smith? [f,*], Yanni Xiao [g,*]

[a] School of Science, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210023, China
[b] China-Australia Joint Research Center for Infectious Diseases, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi, 710061, China
[c] Key Laboratory for Disease Prevention and Control and Health Promotion of Shaanxi Province, Xi'an, Shaanxi, 710061, China
[d] The Interdisciplinary Center for Mathematics and Life Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China
[e] Ministry of Education, Key Laboratory of Environment and Genes Related to Diseases (Xi'an Jiaotong University), Xi'an, Shaanxi, 710061, China
[f] School of Mathematics and Statistics, University of Ottawa, Ottawa, K1N 6N5, Canada
[g] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China

ARTICLE INFO

Keywords:
Multi-scale model
Multi-objective optimization
Antiretroviral therapy schemes
Treatment initiation.

ABSTRACT

Existing single-scale HIV models often overlook conflicts between within-host and population-level treatment strategies. To resolve this, we develop a multi-scale model with infection-age structure to optimize antiretroviral therapy (ART) initiation timing and efficacy through multi-objective optimization. By reformulating initiation time as a control variable in the system, we establish the existence of Pareto-optimal solutions through variational analysis and provide a theoretical characterization of the optimal treatment schedule. Numerical analyses reveal that optimal ART strategies must dynamically adapt to evolving priorities between individual outcomes and population-level benefits: early high-efficacy ART maximizes individual benefits, while delayed initiation optimizes population outcomes under persistent post-treatment risks. Post-treatment behavioural shifts influence this balance: reduced sexual activity supports earlier initiation, while high-risk populations require precisely timed delayed administration to balance transmission control and treatment costs. This framework provides quantitative principles for reconciling scale-specific treatment priorities.

1. Introduction

Human immunodeficiency virus (HIV) causes acquired immunodeficiency syndrome (AIDS) by infecting healthy $CD4^+$ T cells, which weakens the immune system and increases susceptibility to opportunistic infections such as tuberculosis, fungal infections, severe bacterial infections and certain cancers. By the end of 2023, there were an estimated 39.9 million people worldwide living with HIV. In the same year, approximately 630,000 people died from HIV-related causes, while 1.3 million new infections were reported [1]. Despite extensive research efforts, there is still no effective vaccine to prevent HIV infection, nor a definitive cure for AIDS. Antiretroviral therapy (ART) is the major tool used to control disease progression and prevent transmission. Therefore, it is crucial to explore and optimize ART implementation.

* Corresponding authors.
E-mail addresses: stacey.smith@uottawa.ca (S.R. Smith?), yxiao@mail.xjtu.edu.cn (Y. Xiao).

https://doi.org/10.1016/j.apm.2025.116527
Received 27 May 2025; Received in revised form 13 October 2025; Accepted 21 October 2025
Available online 31 October 2025

ART effectively suppresses viral replication within an infected individual, which alleviates symptoms, reduces HIV-related mortality and significantly extends life expectancy [2–4]. However, following ART initiation, patients transition to chronic disease management, incurring lifelong medication burdens, frequent clinical monitoring and strict adherence requirements [5–7]. Moreover, some of these individuals are vulnerable to drug interactions and adverse side effects, including metabolic, hematologic, cardiovascular and gastrointestinal complications, which pose challenges for clinical management. Thus, it is essential to identify an optimal ART regimen for an individual living with HIV that balances viral suppression, treatment-related side effects and costs.

As well as individual benefits, ART also serves as a key public-health intervention by reducing an individual's infectiousness and consequently limiting HIV transmission at the population level [8–11]. Current strategies emphasize early initiation of ART and expansion of treatment coverage [9,12]. However, whether the effectiveness of ART in reducing within-host viral load can successfully translate into a reduction in HIV transmission at the population level remains unclear. Some factors complicate this relationship. First, residual infectivity persists even among treated individuals, particularly in cases of virological failure or incomplete viral suppression [13]. Moreover, earlier treatment initiation prolongs lifespan [14], which inadvertently increases the duration of potential exposure and transmission opportunities. It has been shown that if ART coverage is scaled up to only 30–50 % of the infected population, earlier treatment leads to extended life expectancy without significantly reducing incidence rates [15,16]. It has been shown that ART could increase the overall the number of infections due to lower disease-induced mortality rate and longer survival of HIV-infected individuals [17]. Additionally, some research has suggested that earlier and scaled-up ART programs have been associated with increased risky behaviours, potentially offsetting the preventive benefits of treatment [18–20]. These findings demonstrate that the combined effects of prolonged survival and behavioural changes may undermine the overall effectiveness of early ART. Economically, ART imposes costs on both public-health systems and individuals, including treatment cost, clinical management, private expenditures and potential productivity losses [21,22]. Therefore, optimizing ART implementation — particularly determining the optimal timing for ART — is crucial for ensuring transmission control while also minimizing societal costs.

Given these trade-offs, ART policies must be carefully designed to balance individual and societal objectives. Optimal-control theory has been widely applied to HIV management, addressing different goals at the within-host and population levels. At the individual level, the objectives mainly include maximizing $CD4^+$ T cell counts [23,24], minimizing the viral load [25–27], reducing drug-resistant strains [28,29] and limiting the treatment duration [25]. At the population level, strategies aim to minimize the intervention costs while reducing the disease incidence and prevalence [30–34].

Multi-scale models have been applied to both HIV and other diseases, including COVID-19 [35–38]. Ford and Cuipe developed a multi-scale model that links the virus profile of infected individuals with population-level transmission and testing [36]. They also used a multi-scale model to predict the role of testing in a vaccinated population with amplified variants of the disease [37]. Xue et al. [38] used a data-driven model linking viral dynamics with population transmission dynamics to show that school closures are critical as the virus mutates.

Several studies have considered both within-host viral HIV dynamics and between-host transmission simultaneously, due to their complex interplay [17,39]. Notably, previous research has demonstrated conflicts between individual- and population-level optimal strategies. For example, in high-income countries, HIV incidence among men who have sex with men (MSM) continues to rise, despite reductions in community viral load [40]. Shen et al. linked individual and host scales in order to derive the reproduction number and show that antiretroviral therapy might not be able to reduce new infections [41]. Park et al. combined HIV virulence tradeoffs with a range of contact models, explicitly modelling partnership formation and dissolution, demonstrating that the simplest random-mixing structure best approximates the most realistic model [42]. Doekes et al. nested a within-host quasispecies model that incorporates a long-lived reservoir inside an epidemiological model of HIV dynamics [43]. They showed that the latent reservoir plays a role in both within- and between-host evolution. Gilchrist and Coombs revealed a conflict between the optimal selection pressures of HIV at the within-host and between-host levels [44]. Shen et al. found that the optimal drug efficacy for minimizing the population's basic reproduction number conflicts with within-host strategies for viral suppression when drug efficacy is suboptimal [17]. However, specific strategies for resolving such conflicts in treatment implementation remain unclear. Moreover, these studies have not explicitly considered the timing of ART initiation as a control variable, perhaps due to the mathematical complexity involved, particularly when the objective functions include integral terms whose limits of integration depend on the control variable.

In this study, we extend the single-objective optimization problem to a multi-objective framework in order to identify the optimal compromise between individual and population objectives, with particular emphasis on the initiation time of ART. We construct a multi-scale model with an infection-age structure and formulate a multi-objective optimization problem to determine the optimal ART initiation time and drug efficacy that minimize within-host viral load, the cumulative number of infections at the population level and treatment costs. To address the mathematical challenges arising from the dependence of integral bounds on the initiation time, we reformulate ART initiation as a control variable by employing the time-dependent switched-system method. We then apply Ekeland's Variational Principle to prove the existence of Pareto-optimal solutions and characterize the corresponding optimal-control strategies via the adjoint system. Numerical simulations complement the theoretical results and illustrate how optimal-therapy protocols can be coordinated under different priorities, followed by a comprehensive discussion of the results.

The main contributions of this work are twofold. First, unlike previous single-objective studies [17,39], we develop a rigorous multi-scale, multi-objective optimization framework that explicitly reconciles conflicts between within-host and population-level treatment objectives, thereby enabling the identification of Pareto-optimal strategies. Second, we introduce a methodological advancement by treating ART initiation time as a control variable. This reformulation resolves the technical difficulty arising from variable-dependent integration bounds and provides a theoretical characterization of the optimal initiation schedule — an aspect largely overlooked in previous multi-scale HIV models.

2. Coupling individual and epidemiological models

To simultaneously investigate the optimal ART regimen from both individual and population perspectives, we present a viral dynamics model and link the within-host model to the between-host transmission model depending on the relationship between viral load and infectiousness.

2.1. Within-host viral dynamic model

Let τ denote the infection-age, which is the time elapsed since an individual was infected. The variables $T(\tau), T^*(\tau)$ and $V(\tau)$ denote the concentrations of healthy CD4$^+$T cells, active infected CD4$^+$T cells and free virus at infection-age τ, respectively. If an individual is infected, the within-host viral dynamics are described by the following model [45]:

$$\begin{aligned}
\frac{dT(\tau)}{d\tau} &= \lambda_T - d_T T - (1-\epsilon(\tau))k_s TV,\\
\frac{dT^*(\tau)}{d\tau} &= (1-\epsilon(\tau))k_s TV - d_s T^*,\\
\frac{dV(\tau)}{d\tau} &= p_s T^* - d_{vs} V,
\end{aligned} \tag{1}$$

with $T(0)=T_0$, $T^*(0)=T_0^*$ and $V(0)=V_0$. The parameters λ_T and d_T represent the recruitment and death rates of healthy CD4$^+$T cells, d_s denotes the death rate of infected CD4$^+$T cells, while p_s and d_{vs} represent the production and clearance rates of virus. The function $\epsilon(\tau)$ is defined as a piecewise function

$$\epsilon(\tau) = \begin{cases} 0, & \text{for } \tau \in [0,\tau_s),\\ \epsilon_s(\tau), & \text{for } \tau \in [\tau_s, T_s], \end{cases} \tag{2}$$

where τ_s is the initiation time of ART, $\epsilon_s(\tau)$ denotes the time-dependent drug efficacy and T_s represents the time at which the infected individual progresses to the AIDS stage. According to Dodd et al [16], the relationship between τ_s and T_s is given by

$$T_s - \tau_s = (\hat{L}_1 - d_A) - l_1\tau_s \equiv L_1 - l_1\tau_s, \tag{3}$$

where $\hat{L}_1$ represents the average maximum life expectancies for those who are treated immediately after infection and d_A is the duration from AIDS stage to death. The parameter l_1 quantifies the reduction in lifespan per unit delay in ART initiation timing.

2.2. Between-host transmission model

The following coupled model describes the transmission of HIV in a homogeneous population. Let $S(t)$ be the number of susceptible individuals at time t and $I(\tau,t)$ denote the number of infected individuals (prior to the AIDS stage) with infection-age τ at time t. Uninfected individuals enter the susceptible population at a constant rate Λ. Susceptible individuals become infected through contact with infected individuals of infection-age τ at a transmission rate $\beta(V(\tau))$. Infected individuals exit this class either through natural death at rate μ, progression to the AIDS stage at rate σ or disease-related death at rate $\delta(V)$. The transmission and disease-related death rates are assumed to satisfy $\beta'(V)\geq 0, \beta''(V)\leq 0, \delta'(V)\geq 0, \delta''(V)\leq 0$ [46,47].

$$\begin{aligned}
\frac{dS}{dt} &= \Lambda - \mu S - \frac{S(t)}{N(t)}\int_0^{\tau_s}\beta(V)I(\tau,t)d\tau - \frac{S(t)}{N(t)}\int_{\tau_s}^{T_s(\tau_s)}\beta(V)I(\tau,t)d\tau,\\
\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial \tau}\right)I(\tau,t) &= -(\mu+\delta(V)+\sigma)I(\tau,t),\\
I(0,t) &= \frac{S(t)}{N(t)}\left(\int_0^{\tau_s}\beta(V)I(\tau,t)d\tau + \int_{\tau_s}^{T_s(\tau_s)}\beta(V)I(\tau,t)d\tau\right),\\
S(0) &= S_0, I(\tau,0) = I_0(\tau),
\end{aligned} \tag{4}$$

where the progression rate σ is given by

$$\sigma = \begin{cases} 0, & \tau \in [0,\tau_s),\\ \sigma_1 = \dfrac{1}{T_s-\tau_s}, & \tau\in[\tau_s,T_s]. \end{cases}$$

It is important to note that we focus solely on the optimal treatment scheme for first-line drugs, excluding the potential transition to second-line treatments due to drug resistance. Individuals are assumed to progress to the AIDS stage following treatment failure, and no reversion from the AIDS stage back to the chronic stage is considered. Furthermore, we assume that AIDS patients who experience treatment failure are assumed to be too ill to engage in sexual activity, thereby preventing further transmission of HIV. Consequently, they are not included in the transmission dynamics described above.

3. Existence, uniqueness and boundedness

For the within-host viral dynamics model (1), the existence and boundedness of solutions have been derived by Numfor et al. [39]. In this section, we prove the existence, uniqueness and boundedness of solutions for the coupled between-host model (4).

For convenience, we denote

$$I(0,t) = B_s(t)$$

and define the probability that an infected individual remains in the original class at infection age τ as

$$\kappa_s(\tau) = e^{-\int_0^\tau (\mu+\delta(V)+\sigma)da}.$$

Introducing a constant $\alpha \geq \max \beta(V)$ and denoting $B(s) = \int_0^{\tau_s} \beta(V)I(\tau,s)d\tau + \int_{\tau_s}^{T_s} \beta(V)I(\tau,s)d\tau$, the differential equation for $S(t)$ can be rewritten as

$$\frac{dS}{dt} + \alpha S = \Lambda + \alpha S - \mu S - \frac{S(t)}{N(t)}B(t).$$

Assuming that a solution to model (4) exists, it can be expressed using an integrating factor and the characteristics method as follows:

$$S(t) = S_0 e^{-(\mu+\alpha)t} + \frac{\Lambda}{\mu+\alpha}\left[1 - e^{-(\mu+\alpha)t}\right] + \int_0^t e^{-(\mu+\alpha)(t-u)}\left(\alpha - \frac{B(u)}{N(u)}\right)S(u)du,$$

$$I(\tau,t) = \begin{cases} B_s(t-\tau)\kappa_s, & \text{for } t > \tau, \\ I_0(\tau - t)\dfrac{\kappa_s(\tau)}{\kappa_s(\tau - t)}, & \text{for } t \leq \tau. \end{cases}$$

Denoting $I(t) = \int_0^{T_s} I(\tau,t)d\tau$ and $N(t) = S(t) + I(t)$, we have

$$\begin{aligned} \frac{dI(t)}{dt} &= \int_0^{T_s} \frac{\partial I(\tau,t)}{\partial t}d\tau = \int_0^{T_s}\left[-\frac{\partial I(\tau,t)}{\partial \tau} - (\mu+\delta(V)+\sigma)I\right]d\tau \\ &= I(0,t) - I(T_s,t) - \int_0^{T_s}(\mu+\delta(V)+\sigma)I(\tau,t)d\tau. \end{aligned}$$

Then

$$\begin{aligned} \frac{dN(t)}{dt} &= \frac{dS(t)}{d} + \frac{dI(t)}{dt} \\ &= \Lambda - \mu S - \frac{S(t)}{N(t)}\int_0^{T_s}\beta(V)I(\tau,t)d\tau + I(0,t) - I(T_s,t) - \int_0^{T_s}(\mu+\delta(V)+\sigma)I(\tau,t)d\tau. \end{aligned}$$

Since $I(0,t) = \frac{S(t)}{N(t)}\int_0^{T_s}\beta(V)I(\tau,t)d\tau$, we obtain

$$\begin{aligned} \frac{dN(t)}{dt} &= \Lambda - \mu S - I(T_s,t) - \int_0^{T_s}(\mu+\delta(V)+\sigma)I(\tau,t)d\tau \\ &\leq \Lambda - \mu S(t) - \int_0^{T_s}\mu I(\tau,t)d\tau = \Lambda - \mu S(t) - \mu I(t) = \Lambda - \mu N(t). \end{aligned}$$

It follows that $\lim_{t\to\infty} N(t) \leq \Lambda/\mu$. Thus, for sufficiently large t and any small $\varepsilon > 0$, we have $S(t) \leq \Lambda/\mu + \varepsilon, I(t) \leq \Lambda/\mu + \varepsilon$. Moreover, the expressions for $S(t)$ and $I(\tau,t)$ imply that $S(t) \geq \varsigma$ and $I(\tau,t) \geq 0$, where $\varsigma = \min\left\{S_0, \frac{\Lambda}{\mu+\alpha}\right\}$. Therefore, any solution to system (4) is nonnegative and bounded.

To prove the existence and uniqueness of the solution, we define the solution space as

$$\begin{aligned} X = \Big\{ &(S,I) \in \left(L^\infty(0,T_m) \times L^\infty\left(0,T_m; L^1(0,T_s)\right)\right) \Big| S(t) \geq \varsigma, I(\tau,t) \geq 0, \sup_t S(t) < \infty, \\ &\sup_t \int_0^{T_s} I(\tau,t)d\tau < \infty \ a.e. \text{ with respect to } t \Big\}, \end{aligned}$$

where $L^\infty(0,T_m)$ is the space of all essentially bounded functions on $(0,T_m)$. We define a mapping on the set X as

$$\tilde{\mathcal{L}} : X \to X, \quad \tilde{\mathcal{L}}(S,I) = \left(\tilde{L}_1(S,I), \tilde{L}_2(S,I)\right),$$

where

$$\tilde{L}_1(S,I)(t) = S_0 e^{-(\mu+\alpha)t} + \frac{\Lambda}{\mu+\alpha}\left(1 - e^{-(\mu+\alpha)t}\right) + \int_0^t e^{-(\mu+\alpha)(t-s)}\left(\alpha - \frac{B(s)}{N(s)}\right)S(s)ds,$$

$$\tilde{L}_2(S,I)(\tau,t) = \begin{cases} B_s(t-\tau)\kappa_s, & \text{for } t > \tau, \\ I_0(\tau - t)\dfrac{\kappa_s(\tau)}{\kappa_s(\tau - t)}, & \text{for } t \leq \tau. \end{cases}$$

Using Contraction Mapping Principle [48], we obtain the following theorem.

Theorem 1. *For $T_m < \infty$, there exists a unique bounded solution for system (4).*

Proof. See Appendix A. □

4. Multi-objective optimization

We formulate a multi-objective optimization problem (MOP) based on the multi-scale model. The goal is to investigate the optimal drug efficacy and initiation time of ART in order to minimize the viral load within a host, while also minimizing the number of infected individuals at the population level. The MOP is defined as

$$\text{(MOP)} \qquad \min_{(\epsilon(\tau),\tau_s)\in\mathcal{U}} \hat{J} = \left\{J_1\left(\epsilon(\tau),\tau_s\right), J_2\left(\epsilon(\tau),\tau_s\right)\right\},\ \tau_s \in [\bar{a},\bar{b}],\ \epsilon(\tau) \in [\bar{c},\bar{d}],$$

where the two objective functionals are:

$$J_1\left(\epsilon(\tau),\tau_s\right) = \int_0^{\tau_s} V(\tau)d\tau + \int_{\tau_s}^{T_s(\tau_s)} V(\tau)d\tau + \int_{\tau_s}^{T_s(\tau_s)} \frac{C_1}{2} D_0(\tau-\tau_s)d\tau + \int_{\tau_s}^{T_s(\tau_s)} \frac{C_2}{2}\epsilon_s^2(\tau)d\tau,$$

$$\begin{aligned} J_2\left(\epsilon(\tau),\tau_s\right) = &\int_0^{T_m}\int_0^{\tau_s} I(\tau,t)d\tau dt + \int_0^{T_m}\int_{\tau_s}^{T_s(\tau_s)} I(\tau,t)d\tau dt \\ &+ \int_0^{T_m}\int_{\tau_s}^{T_s(\tau_s)} \frac{C_1}{2} D_0(\tau-\tau_s) I(\tau,t)d\tau dt + \int_0^{T_m}\int_{\tau_s}^{T_s(\tau_s)} \frac{C_2}{2}\epsilon_s^2(\tau) I(\tau,t)d\tau dt, \end{aligned}$$

subject to control systems (1) and (4). The control set is

$$\mathcal{U} = \{(\epsilon,\tau_s) \in (L^\infty(0,T_s))^2 | \epsilon : (0,T_s) \to [\bar{c},\bar{d}], \tau_s \equiv C, \bar{a} \le C \le \bar{b}\}.$$

J_1 is an objective function at the within-host level, consisting of three components: (1) the cumulative viral load $\int_0^{T_s(\tau_s)} V(\tau)d\tau$ spanning both untreated $[0,\tau_s)$ and treated $[\tau_s,T_s]$ phases before AIDS onset; (2) the disability burden $\int_{\tau_s}^{T_s(\tau_s)} C_1 D_0(\tau-\tau_s)/2d\tau$, where D_0 represents disability weights for treatment-related side effects or complications; and (3) treatment costs $\int_{\tau_s}^{T_s(\tau_s)} C_2\epsilon_s^2(\tau)/2d\tau$ encompassing drug expenses and healthcare resource utilization. J_2 is an objective function at the population level, comprising: (1) the accumulated number of infected individuals $\int_0^{T_m}\int_0^{\tau_s} I(\tau,t)d\tau dt + \int_0^{T_m}\int_{\tau_s}^{T_s(\tau_s)} I(\tau,t)d\tau dt$ before the onset of AIDS over the time interval $[0,T_m]$; (2) the population disability burden $\int_{\tau_s}^{T_s(\tau_s)} C_1 D_0(\tau-\tau_s) I(\tau,t)/2d\tau$; and (3) public-health expenditure $\int_{\tau_s}^{T_s(\tau_s)} C_2\epsilon_s^2(\tau) I(\tau,t)/2d\tau$. The coefficients C_1, C_2 serve as balancing weights between disability impacts and economic costs in both objectives.

There are two types of solutions for the MOP:

(1) There is a unique absolutely optimal solution that simultaneously optimizes both objective functions.

(2) There exist Pareto-optimal solutions where improvement in one objective necessitates worsening the other. That is, if (ϵ^*,τ_s^*) and $(\epsilon^{**},\tau_s^{**})$ are Pareto-optimal solutions, they satisfy $J_1(\epsilon^*,\tau_s^*) \le J_1(\epsilon^{**},\tau_s^{**})$ and $J_2(\epsilon^*,\tau_s^*) \ge J_2(\epsilon^{**},\tau_s^{**})$.

To solve the MOP, we apply the linear weighted sum method to convert it into a single-objective optimization problem (SOP):

$$\text{(SOP)} \qquad \min_{(\epsilon(\tau),\tau_s)\in\mathcal{U}} J = W_1 J_1 + W_2 J_2,$$

where the control system and control set remain consistent with those in the MOP. For any $W_1 \ge 0, W_2 \ge 0$, it can be shown that the following holds:

$$\arg\min_{(\epsilon(\tau),\tau_s)\in\mathcal{U}} J = \arg\min_{(\epsilon(\tau),\tau_s)\in\mathcal{U}} \left(\frac{W_1}{W_1+W_2}J_1 + \frac{W_2}{W_1+W_2}J_2\right).$$

Hence, we only need to analyze the case where $W_1 \ge 0, W_2 \ge 0$ and $W_1 + W_2 = 1$ for the SOP.

The relationship between the optimal solution of the SOP and the Pareto-optimal solutions of the MOP is established in the following.

Theorem 2. *For any $0 \le W_1 \le 1$ and $W_2 = 1 - W_1$, the optimal solution to the SOP is a Pareto-optimal solution to the MOP. In particular, if the optimal solutions to the SOP coincide for the cases $W_1 = 1, W_2 = 0$ and $W_1 = 0, W_2 = 1$, then this optimal solution is the absolutely optimal solution to the MOP.*

Proof. For any weight $W_1 \ge 0, W_2 \ge 0$ and $W_1 + W_2 = 1$, assume that there exists an optimal solution $u^* = (\epsilon^*(\tau),\tau_s^*)$ for the SOP such that

$$J(u^*) = \min_{u\in\mathcal{U}} J(u).$$

If u^* is not the Pareto-optimal solution for the MOP, there exists $\tilde{u}^* = (\tilde{\epsilon}^*(\tau),\tilde{\tau}_s)$ such that

$$J_1(\tilde{u}^*) \le J_1(u^*), J_2(\tilde{u}^*) \le J_2(u^*).$$

Therefore, the following holds:

$$W_1 J_1(\tilde{u}^*) + W_2 J_2(\tilde{u}^*) \le W_1 J_1(u^*) + W_2 J_2(u^*),$$

which implies

$$J(\tilde{u}^*) \le J(u^*).$$

This contradicts the assumption that u^* is the optimal solution for the SOP. Hence u^* is the Pareto-optimal solution for the MOP.

Assuming that for $W_1 = 1, W_2 = 0$ and $W_1 = 0, W_2 = 1$, the corresponding optimal solutions for the SOP are both (ϵ^*, τ_s^*). Then for any (ϵ, τ_s), the following holds:

$$J_1(\epsilon^*, \tau_s^*) \le J_1(\epsilon, \tau_s), J_2(\epsilon^*, \tau_s^*) \le J_2(\epsilon, \tau_s).$$

Hence (ϵ^*, τ_s^*) is the absolutely optimal solution for the MOP. □

Remark. We adopt the linear weighted-sum method because it is mathematically tractable and allows the multi-objective optimization problem to be reformulated into a single-objective problem, thereby making it amenable to variational analysis and optimal-control theory. By varying the weights, we can systematically trace out the Pareto-front, thereby revealing possible trade-offs between the micro-objective J_1 and the macro-objective J_2. Moreover, the weights themselves have a clear and intuitive interpretation: they directly reflect the decision-maker's relative emphasis on individual-level outcomes versus population-level outcomes. This not only ensures rigorous mathematical treatment but also facilitates policy discussion and scenario analysis in practical applications. While alternative methods (e.g., ε-constraint, evolutionary algorithms) are available, the weighted-sum method is particularly well-suited for analytical proofs and for providing interpretable results in this continuous optimization framework.

4.1. The optimal solution for the SOP

In this section, we prove the existence of the optimal solution for the SOP. The within-host viral dynamics switch over as antiretroviral therapy begins. Consequently, determining the optimal time to start ART involves identifying the optimal switching point within the switched system. Utilizing the time-dependent switched-system approach introduced by Xu et al. [49], we map the switching point into the control function of the switched system. This allows us to reformulate the optimal-control problem into an equivalent problem parameterized by the switching instants.

Introducing a new time variable a, we define a bijective mapping between τ and a as

$$\tau = \begin{cases} \tau_s a, & \text{for } a \in [0,1), \\ \tau_s + (L_1 - l_1\tau_s)(a-1), & \text{for } a \in [1,2], \end{cases} \tag{5}$$

where $a = 0$ corresponds to the infected time $\tau = 0$; $a = 1$ corresponds to the initiation time of ART $\tau = \tau_s$; and $a = 2$ corresponds to the starting time of the AIDS stage $\tau = T_s$.

Consequently, the SOP objective functionals J_1 and J_2 are converted to

$$\begin{aligned} J_1 &= \int_0^1 \tau_s V(a)da + \int_1^2 (L_1 - l_1\tau_s)V(a)da + \int_1^2 \frac{C_1}{2} D_0 (L_1 - l_1\tau_s)^2 (a-1)da \\ &\quad + \int_1^2 \frac{C_2}{2}(L_1 - l_1\tau_s)\epsilon_s^2(a)da, \\ J_2 &= \int_0^{T_m}\int_0^1 \tau_s I(a,t)dadt + \int_0^{T_m}\int_1^2 (L_1 - l_1\tau_s)I(a,t)dadt \\ &\quad + \int_0^{T_m}\int_1^2 \frac{C_1}{2} D_0 (L_1 - l_1\tau_s)^2 (a-1)I(a,t)dadt + \int_0^{T_m}\int_1^2 \frac{C_2}{2}(L_1 - l_1\tau_s)\epsilon_s^2(a)I(a,t)dadt, \end{aligned}$$

subject to the control system

$$\begin{aligned} \frac{dT}{da} &= g(\tau_s)\left[\lambda_T - d_T T - (1-\epsilon(a))k_s TV\right], \\ \frac{dT^*}{da} &= g(\tau_s)\left[(1-\epsilon(a))k_s TV - d_s T^*\right], \\ \frac{dV}{da} &= g(\tau_s)\left(p_s T^* - d_{vs} V\right), \\ \frac{dS}{dt} &= \Lambda - \mu S - \frac{S(t)}{N(t)}\int_0^1 \tau_s \beta(V) I(a,t)da - \frac{S(t)}{N(t)}\int_1^2 (L_1 - l_1\tau_s)\beta(V)I(a,t)da, \\ \left(g(\tau_s)\frac{\partial}{\partial t} + \frac{\partial}{\partial a}\right) I(a,t) &= -g(\tau_s)(\mu + \delta(V) + \sigma)I(a,t), \\ I(0,t) &= \frac{S(t)}{N(t)}\left[\int_0^1 \tau_s \beta(V) I(a,t)da + \int_1^2 (L_1 - l_1\tau_s)\beta(V)I(a,t)da\right], \end{aligned} \tag{6}$$

with initial conditions $T(0) = T_0, T^*(0) = T_0^*, V(0) = V_0, S(0) = S_0$ and $I(a,0) = I_0(a)$, where $g(\tau_s)$ is a piecewise function

$$g(\tau_s) = \begin{cases} \tau_s, & \text{for } a \in [0,1), \\ L_1 - l_1\tau_s, & \text{for } a \in [1,2]. \end{cases} \tag{7}$$

The control set is $\mathcal{U} = \left\{ (\epsilon, \tau_s) \in (L^\infty(0,2))^2 \middle| \epsilon : (0,2) \to [\bar{c}, \bar{d}], \tau_s \equiv C, \bar{a} \le C \le \bar{b} \right\}$.

In the following, we first show that the control system (6) is Lipschitz continuous with respect to the control variables. Building on this property, we derive the sensitivity equations and subsequently the adjoint system. Finally, we invoke Ekeland's Principle [50] to establish the existence of an optimal solution for the SOP.

4.1.1. Sensitivity equations and adjoint system

To prove the existence and uniqueness of the optimal-control pair, we begin by proving that the solution map of the control system (6) is Lipschitz continuous in terms of the control variables.

Theorem 3. *Define a map $S_1 : (\tau_s, \epsilon) \to (T, T^*, V, S, I)(\tau_s, \epsilon)$. The map S_1 is Lipschitz in the following ways:*

$$
\begin{aligned}
(1) \int_0^1 \left(|T-\bar{T}| + |T^*-\bar{T}^*| + |V-\bar{V}|\right) da + \int_1^2 \left(|T-\bar{T}| + |T^*-\bar{T}^*| + |V-\bar{V}|\right) da \\
+ \int_0^{T_m} |S-\bar{S}| dt + \int_0^{T_m} \int_0^1 |I-\bar{I}| da dt \\
+ \int_0^{T_m} \int_1^2 |I-\bar{I}| da dt \\
\le \tilde{K}_1 \int_0^2 \left(|\tau_s - \bar{\tau}_s| + |\epsilon - \bar{\epsilon}|\right) da,
\end{aligned}
$$

$$
\begin{aligned}
(2) \|T-\bar{T}\|_{L^\infty(0,1)} + \|T^*-\bar{T}^*\|_{L^\infty(0,1)} + \|V-\bar{V}\|_{L^\infty(0,1)} + \|T-\bar{T}\|_{L^\infty(1,2)} \\
+ \|T^*-\bar{T}^*\|_{L^\infty(1,2)} + \|V-\bar{V}\|_{L^\infty(1,2)} + \|S-\bar{S}\|_{L^\infty(0,T_m)} \\
+ \|I-\bar{I}\|_{L^\infty(0,T_m;L^1(0,1))} + \|I-\bar{I}\|_{L^\infty(0,T_m;L^1(1,2))} \\
\le \tilde{K}_2 \left(|\tau_s - \bar{\tau}_s| + \|\epsilon - \bar{\epsilon}\|_{L^\infty(0,2)}\right).
\end{aligned}
$$

Proof. See Appendix B. □

With the Lipschitz property established, we next analyze the differentiability of the solution map S_1 with respect to (τ_s, ϵ) to characterize first-order variations of the states. This leads to the sensitivity equations and, by duality, to the adjoint system, which transfers these variations from the state space to the control space and ensures that the derivative of the objective functional J with respect to the controls can be expressed entirely in terms of the adjoint variables and the original state trajectories, without explicit reference to the sensitivities.

Theorem 4. *The mapping $S_1 : (\tau_s, \epsilon) \to (T, T^*, V, S, I)(\tau_s, \epsilon)$ is differential in the following way: for any constants h_1 and $h_2 \in L^\infty(0,2)$, there is*

$$
\lim_{\varepsilon \to 0} \frac{(T, T^*, V, S, I)(\tau_s + \varepsilon h_1, \epsilon + \varepsilon h_2) - (T, T^*, V, S, I)(\tau_s, \epsilon)}{\varepsilon} = (\phi, \phi^*, v, s, i),
$$

with $\left(\tau_s + \varepsilon h_1, \epsilon + \varepsilon h_2\right) \in \mathcal{U}$ and $(\tau_s, \epsilon) \in \mathcal{U}$.

Proof. Since the solution map S_1 is Lipschitz with respect to control variables τ_s and ϵ, it follows from [39,51] that this mapping admits the Gâteaux derivatives ϕ, ϕ^*, v, s, i. □

Passing to the limit in the representation of the quotients, the derivatives (ϕ, ϕ^*, v, s, i) satisfy the following sensitivity equations:

$$
\begin{aligned}
\frac{d\phi}{da} &= g(\tau_s)\left[-d_T\phi - (1-\epsilon)k_s\phi V - (1-\epsilon)k_s T v\right] + g_1 h_1 \left(\lambda_T - d_T T - k_s T V\right) \\
&\quad + g_2 h_2 g(\tau_s) k T V, \\
\frac{d\phi^*}{da} &= g(\tau_s)\left[(1-\epsilon)k_s\phi V + (1-\epsilon)k_s T v - d_s\phi^*\right] + g_1 h_1 \left(k_s T V - d_s T^*\right) - g_2 h_2 g(\tau_s) k T V, \\
\frac{dv}{da} &= g(\tau_s)\left(p_s\phi^* - d_{vs} v\right) + g_1 h_1 \left(p_s T^* - d_{vs} V\right), \\
\frac{ds}{dt} &= -\mu s - \frac{1}{N(t)}\tilde{f}_1(t)s + \frac{S(t)}{N^2(t)}\tilde{f}_1(t)\left(s + \int_0^1 \tau_s i(a,t) da + \int_1^2 (L_1 - l_1\tau_s) i(a,t) da\right) \\
&\quad - \frac{S(t)}{N(t)}\tilde{f}_2(t) - \frac{S(t)}{N(t)}\tilde{f}_3(t) - h_1\frac{S(t)}{N(t)}\left[\int_0^1 \beta(V) I(a,t) da - \int_1^2 l_1 \beta(V) I(a,t) da\right] \\
&\quad - h_1 \frac{S(t)}{N^2(t)}\left(\int_0^1 I(a,t) da - \int_1^2 l_1 I(a,t) da\right), \\
g(\tau_s)\frac{\partial i}{\partial t} + \frac{\partial i}{\partial a} &= -g(\tau_s)\left[(\mu + \delta(V) + \sigma) i + \frac{\partial \delta(V)}{\partial V} v I\right] - g_1 h_1 \left[\frac{\partial I(a,t)}{\partial t} + (\mu + \delta(V)) I\right],
\end{aligned}
$$

$$
\begin{aligned}
i(0,t) = {} & \frac{1}{N(t)}\tilde{f}_1(t)s - \frac{S(t)}{N^2(t)}\tilde{f}_1(t)\left(s + \int_0^1 \tau_s i(a,t)da + \int_1^2 (L_1 - l_1\tau_s)i(a,t)da\right) \\
& + \frac{S(t)}{N(t)}\tilde{f}_2(t) + \frac{S(t)}{N(t)}\tilde{f}_3(t) + h_1\frac{S(t)}{N(t)}\left[\int_0^1 \beta(V)I(a,t)da - \int_1^2 l_1\beta(V)I(a,t)da\right] \\
& + h_1\frac{S(t)}{N^2(t)}\left(\int_0^1 I(a,t)da - \int_1^2 l_1 I(a,t)da\right), \\
\phi(0) = 0, \phi^*(0) = 0, & v(0) = 0, s(0) = 0, i(a,0) = 0,
\end{aligned} \tag{8}
$$

where $g(\tau_s)$ is given in (7) and g_1, g_2 are both piecewise functions:

$$
g_1 = \begin{cases} 1, & \text{for } a \in [0,1), \\ -l_1, & \text{for } a \in [1,2], \end{cases} \text{ and } g_2 = \begin{cases} 0, & \text{for } a \in [0,1), \\ 1, & \text{for } a \in [1,2]. \end{cases}
$$

The functions $\tilde{f}_1(t), \tilde{f}_2(t), \tilde{f}_3(t)$ are defined as

$$
\begin{aligned}
\tilde{f}_1(t) &= \int_0^1 \tau_s\beta(V)I(a,t)da + \int_1^2 (L_1 - l_1\tau_s)\beta(V)I(a,t)da, \\
\tilde{f}_2(t) &= \int_0^1 \tau_s\beta(V)i(a,t)da + \int_1^2 (L_1 - l_1\tau_s)\beta(V)i(a,t)da, \\
\tilde{f}_3(t) &= \int_0^1 \tau_s\frac{\partial\beta(V)}{\partial V}v(a)I(a,t)da + \int_1^2 (L_1 - l_1\tau_s)\frac{\partial\beta(V)}{\partial V}v(a)I(a,t)da.
\end{aligned}
$$

Next, we define the following sensitivity operators:

$$
\begin{aligned}
\mathcal{L}_1\begin{bmatrix}\phi\\ \phi^*\\ v\end{bmatrix} &= \frac{d}{da}\begin{bmatrix}\phi\\ \phi^*\\ v\end{bmatrix} + g(\tau_s)\begin{bmatrix} d_T + (1-\epsilon)k_sV & 0 & (1-\epsilon)k_sT \\ -(1-\epsilon)k_sV & d_s & -(1-\epsilon)k_sT \\ 0 & -p_s & d_{vs}\end{bmatrix}\begin{bmatrix}\phi\\ \phi^*\\ v\end{bmatrix}, \\
\mathcal{L}_2\begin{bmatrix}s\\ i\end{bmatrix} \equiv \begin{bmatrix}\mathcal{L}_{21}(s,i)\\ \mathcal{L}_{22}(s,i)\end{bmatrix} &= \begin{bmatrix} \frac{ds}{dt} + \tilde{M}_1(s,i) \\ g(\tau_s)\frac{\partial i}{\partial t} + \frac{\partial i}{\partial a} + \tilde{M}_2(s,i)\end{bmatrix},
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{M}_1(s,i) = {} & \mu s + \frac{1}{N(t)}\tilde{f}_1(t)s - \frac{S(t)}{N^2(t)}\tilde{f}_1(t)\left(s + \int_0^1 \tau_s i(a,t)da + \int_1^2 (L_1 - l_1\tau_s)i(a,t)da\right) \\
& + \frac{S(t)}{N(t)}\tilde{f}_2(t) + \frac{S(t)}{N(t)}\tilde{f}_3(t) \\
\tilde{M}_2(s,i) = {} & g(\tau_s)(\mu + \delta(V) + \sigma)i + g(\tau_s)\frac{\partial\delta(V)}{\partial V}vI.
\end{aligned}
$$

The sensitivity equations (8) can be rewritten as:

$$
\begin{aligned}
\mathcal{L}_1\begin{bmatrix}\phi\\ \phi^*\\ v\end{bmatrix} &= \begin{bmatrix} g_1h_1\left(\lambda_T - d_TT - (1-\epsilon)k_sTV\right) + g_2h_2g(\tau_s)k_sTV \\ g_1h_1\left((1-\epsilon)k_sTV - d_sT^*\right) - g_2h_2g(\tau_s)k_sTV \\ g_1h_1\left(p_sT^* - d_{vs}V\right)\end{bmatrix}, \\
\mathcal{L}_2\begin{bmatrix}s\\ i\end{bmatrix} &= \begin{bmatrix} -h_1\frac{S(t)}{N(t)}\left(\int_0^1 \beta(V)Ida - \int_1^2 l_1\beta(V)Ida\right) - h_1\frac{S(t)}{N^2(t)}\left(\int_0^1 Ida - \int_1^2 l_1Ida\right) \\ -g_1h_1\left(\frac{\partial I}{\partial t} + \mu I + \delta(V)I\right)\end{bmatrix}.
\end{aligned}
$$

Introducing the adjoint variables $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5, we establish the following duality relationship between the sensitivity and adjoint operators:

$$
\begin{aligned}
& \int_0^1 (\lambda_1,\lambda_2,\lambda_3)\mathcal{L}_1(\phi,\phi^*,v)da + \int_1^2 (\lambda_1,\lambda_2,\lambda_3)\mathcal{L}_1(\phi,\phi^*,v)da + \int_0^{T_m} \lambda_4\mathcal{L}_{21}(s,i)dt \\
& \qquad + \int_0^{T_m}\int_0^1 \lambda_5\mathcal{L}_{22}(s,i)dadt + \int_0^{T_m}\int_1^2 \lambda_5\mathcal{L}_{22}(s,i)dadt \\
= {} & \int_0^1 (\phi,\phi^*,v)\mathcal{L}_1^*(\lambda_1,\lambda_2,\lambda_3)da + \int_1^2 (\phi,\phi^*,v)\mathcal{L}_1^*(\lambda_1,\lambda_2,\lambda_3)da \\
& \qquad + \int_0^{T_m} s\mathcal{L}_{21}^*(\lambda_4,\lambda_5)dt + \int_0^{T_m}\int_0^1 i\mathcal{L}_{22}^*(\lambda_4,\lambda_5)dadt \\
& \qquad + \int_0^{T_m}\int_1^2 i\mathcal{L}_{22}^*(\lambda_4,\lambda_5)dadt.
\end{aligned} \tag{9}
$$

Next, we define the adjoint equations in the weak sense:

$$\begin{aligned}
\mathcal{L}_1^*\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{bmatrix} &= \begin{bmatrix}0\\ 0\\ W_1 g(\tau_s)\end{bmatrix},\\
\mathcal{L}_2^*\begin{bmatrix}\lambda_4\\ \lambda_5\end{bmatrix} \equiv \begin{bmatrix}\mathcal{L}_{21}^*(\lambda_4,\lambda_5)\\ \mathcal{L}_{22}^*(\lambda_4,\lambda_5)\end{bmatrix} &= \begin{bmatrix}0\\ W_2 g(\tau_s)\left(1+\frac{C_2\epsilon^2(a)}{2}+g(\tau_s)g_2\frac{C_1D_0(a-1)}{2}\right)\end{bmatrix},
\end{aligned} \tag{10}$$

in which the right-hand sides are obtained by differentiating the integrand of the objective functional with respect to each state variable.

Building on the sensitivity system (8) and the duality relation (9) between the sensitivity and adjoint operators, we apply integration by parts to transfer the differential operators from the sensitivity variables (ϕ, ϕ^*, v, s, i) to the adjoint variables $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$. Then, by matching the coefficients of (ϕ, ϕ^*, v, s, i) on both sides of (9) and selecting boundary conditions that cancel the boundary terms produced during integration by parts, we obtain the explicit forms of the adjoint operators $\mathcal{L}_1^*$ and $\mathcal{L}_2^*$ and thereby determine the corresponding adjoint equations. The detailed derivations are given in Appendix C. Combining these results with (10) yields the following complete adjoint system:

$$\begin{aligned}
\frac{d\lambda_1}{da} &= g(\tau_s)\left[d_T\lambda_1+(1-\epsilon)k_sV(\lambda_1-\lambda_2)\right],\\
\frac{d\lambda_2}{da} &= g(\tau_s)(d_s\lambda_2-p_s\lambda_3),\\
\frac{d\lambda_3}{da} &= g(\tau_s)\int_0^{T_m}\left[\frac{S(t)}{N(t)}\frac{\partial\beta(V)}{\partial V}I(\lambda_4-\lambda_5(0,t))+\frac{\partial\delta(V)}{\partial V}I\lambda_5\right]dt\\
&\quad+g(\tau_s)\left[d_{vs}\lambda_3+(1-\epsilon)k_sT(\lambda_1-\lambda_2)-W_1\right],\\
\frac{d\lambda_4}{dt} &= \mu\lambda_4+\left(\frac{1}{N(t)}-\frac{S(t)}{N^2(t)}\right)\tilde{f}_1(t)\left(\lambda_4-\lambda_5(0,t)\right),\\
g(\tau_s)\frac{\partial\lambda_5}{\partial t}+\frac{\partial\lambda_5}{\partial a} &= g(\tau_s)\frac{S(t)}{N(t)}\left(\beta(V)-\frac{\tilde{f}_1(t)}{N(t)}\right)(\lambda_4-\lambda_5(0,t))+g(\tau_s)(\mu+\delta(V)+\sigma)\lambda_5\\
&\quad-W_2g(\tau_s)\left[1+\frac{C_2\epsilon^2(a)}{2}+g(\tau_s)g_2\frac{C_1D_0(a-1)}{2}\right],
\end{aligned} \tag{11}$$

together with the transversality conditions

$$\lambda_1(2)=0,\ \lambda_2(2)=0,\ \lambda_3(2)=0,\ \lambda_4(T_m)=0,\ \lambda_5(a,T_m)=0,\ \lambda_5(2,t)=0.$$

The solution to the adjoint system (11) is considered in the following weak sense:

$$\begin{aligned}
&\int_0^1\left(\lambda_1\psi_1+\lambda_2\psi_2+\lambda_3\psi_3\right)da+\int_1^2\left(\lambda_1\psi_1+\lambda_2\psi_2+\lambda_3\psi_3\right)da+\int_0^{T_m}\lambda_4\psi_4dt\\
&\qquad+\int_0^{T_m}\int_0^1\lambda_5\psi_5dadt+\int_0^{T_m}\int_1^2\lambda_5\psi_5dadt\\
&=W_1\tau_s\int_0^1y_3da+W_1(L_1-l_1\tau_s)\int_1^2y_3da\\
&\qquad+W_2\tau_s\int_0^{T_m}\int_0^1y_5dadt+W_2(L_1-l_1\tau_s)\int_0^{T_m}\int_1^2y_5dadt\\
&\qquad+W_2(L_1-l_1\tau_s)\int_0^{T_m}\int_0^2\left(\frac{C_2\epsilon_s^2}{2}+(L_1-l_1\tau_s)\frac{C_1D_0(a-1)}{2}\right)y_5dadt
\end{aligned}$$

for all $\psi_1,\psi_2,\psi_3\in L^\infty(0,2), \psi_4\in L^\infty(0,T_m), \psi_5\in L^\infty\left(0,T_m;L^1(0,2)\right)$ that satisfy the system

$$\begin{aligned}
\psi_1 &= \frac{dy_1}{da}+g(\tau_s)\left[d_Ty_1+(1-\epsilon)k_sVy_1+(1-\epsilon)k_sTy_3\right],\\
\psi_2 &= \frac{dy_2}{da}-g(\tau_s)\left[(1-\epsilon)k_sVy_1+k_sTy_3-d_sy_2\right],\\
\psi_3 &= \frac{dy_3}{da}-g(\tau_s)\left(p_sy_2-d_{vs}y_3\right),\\
\psi_4 &= \frac{dy_4}{dt}+\mu y_4+\left(\frac{1}{N(t)}+\frac{S(t)}{N^2(t)}\right)\tilde{f}_1(t)y_4+\frac{S(t)}{N(t)}\tilde{f}_{2y}(t)+\frac{S(t)}{N(t)}\tilde{f}_{3y}(t)-\frac{S(t)}{N^2(t)}\tilde{f}_1(t)\tilde{f}_{4y}(t),\\
\psi_5 &= g(\tau_s)\frac{\partial y_5}{\partial t}+\frac{\partial y_5}{\partial a}+g(\tau_s)\left[(\mu+\delta(V)+\sigma)y_5+\frac{\partial\delta(V)}{\partial V}Iy_3\right],
\end{aligned}$$

where

$$\tilde{f}_{2y}(t) = \int_0^1 \tau_s \beta(V) y_5(a,t)da + \int_1^2 (L_1 - l_1\tau_s)\beta(V) y_5(a,t)da,$$
$$\tilde{f}_{3y}(t) = \int_0^1 \tau_s \frac{\partial \beta(V)}{\partial V} I(a,t) y_3(a)da + \int_1^2 (L_1 - l_1\tau_s)\frac{\partial \beta(V)}{\partial V} I(a,t) y_3(a)da,$$
$$\tilde{f}_{4y}(t) = \int_0^1 \tau_s y_5(a,t)da + \int_1^2 (L_1 - l_1\tau_s) y_5(a,t)da.$$

Theorem 5. *For $(\tau_s, \epsilon) \in \mathcal{U}$, the adjoint system admits a weak solution $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ that satisfies the Lipschitz conditions:*

$$\begin{aligned}
&\|\lambda_1 - \bar{\lambda}_1\|_{L^\infty(0,1)} + \|\lambda_2 - \bar{\lambda}_2\|_{L^\infty(0,1)} + \|\lambda_3 - \bar{\lambda}_3\|_{L^\infty(0,1)} + \|\lambda_1 - \bar{\lambda}_1\|_{L^\infty(1,2)} + \|\lambda_2 - \bar{\lambda}_2\|_{L^\infty(1,2)} \\
&\quad + \|\lambda_3 - \bar{\lambda}_3\|_{L^\infty(1,2)} + \|\lambda_4 - \bar{\lambda}_4\|_{L^\infty(0,T)} + \|\lambda_5 - \bar{\lambda}_5\|_{L^\infty(0,T_m;L^1(0,1))} \\
&\quad + \|\lambda_5 - \bar{\lambda}_5\|_{L^\infty(0,T_m;L^1(1,2))} \\
&\quad \le \tilde{K}_3\big(|\tau_s - \bar{\tau}_s| + \|\epsilon - \bar{\epsilon}\|_{L^\infty(0,2)}\big).
\end{aligned}$$

Proof. Regarding the existence of the solution, similar methods to those used in the proof of Theorem 1 are applied. The proof that the solution satisfies the Lipschitz conditions follows in a similar way as shown in Appendix B. □

It is worth emphasizing that the adjoint system can be regarded as the "transpose" or formal dual of the sensitivity system, serving to encode the gradient of the objective functional with respect to the control variables [52,53]. This perspective clarifies the structural relationship between the sensitivity and adjoint equations: the former quantifies the differential of the state with respect to control variations, whereas the latter conveys these variations from the state space to the control space, allowing them to be systematically projected onto the objective functional.

4.1.2. Characterization of the optimal control

The following analysis leverages the sensitivity and adjoint systems to differentiate the objective functional with respect to the control variables. This process allows us to infer the first-order necessary optimality conditions for both the ART initiation time τ_s and the drug efficacy $\epsilon(\tau)$. Consequently, we will obtain an explicit characterization of the optimal control in terms of the state and adjoint variables, thereby bridging abstract optimization theory with interpretable and clinically practical treatment strategies.

To characterize the optimal control of first-order PDEs (6), we first extend the objective function J to the space $L^1(0,2) \times L^1(0,2)$ via an embedding operator:

$$\mathcal{J}(\epsilon, \tau_s) = \begin{cases} J(\epsilon, \tau_s), & \text{if } (\epsilon, \tau_s) \in \mathcal{U}, \\ +\infty, & \text{if } (\epsilon, \tau_s) \notin \mathcal{U}. \end{cases}$$

Then we differentiate the objective function $\mathcal{J}$ with respect to the control variables to derive the optimality conditions, thereby providing a mathematical characterization of the optimal-control pair.

Theorem 6. *If $(\bar{\epsilon}(a), \bar{\tau}_s)$ is the optimal-control pair and $(\bar{T}, \bar{T}^*, \bar{V}, \bar{S}, \bar{I})$ and $(\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3, \bar{\lambda}_4, \bar{\lambda}_5)$ are the corresponding solutions to the control system (6) and the adjoint equations (11), then*

$$\bar{\epsilon}(a) = \begin{cases} 0, \text{ for } a \in [0,1), \\ \bar{\epsilon}_s(a) = \min\left\{\bar{d}, \max\left\{\bar{c}, \dfrac{\big(\bar{\lambda}_2(a) - \bar{\lambda}_1(a)\big)k_s \bar{T}(a)\bar{V}(a)}{C_2\Big(W_1 + W_2 \int_0^{T_m} \bar{I}(a,t)dt\Big)}\right\}\right\}, \text{ for } a \in [1,2], \end{cases}$$
$$\bar{\tau}_s = \min\left\{\bar{b}, \max\left\{\bar{a}, \frac{L_1}{l_1} + \frac{W_1\bar{g}_1 + W_2\bar{g}_2 + \bar{g}_3 + \bar{g}_4}{l_1 C_1 D_0\Big(W_1 \int_1^2 (a-1)da + W_2 \int_0^{T_m}\int_1^2 (a-1)\bar{I}(a,t)dadt\Big)}\right\}\right\},$$

where $\bar{g}_1, \bar{g}_2, \bar{g}_3, \bar{g}_4$ are given in (12).

Proof. Let $\bar{u} = (\bar{\tau}_s, \bar{\epsilon}(a))$ be the optimal control, where

$$\bar{\epsilon}(a) = \begin{cases} 0, & \text{for } a \in [0,1), \\ \bar{\epsilon}_s(a), & \text{for } a \in [1,2]. \end{cases}$$

Denote h_1 and $h_2(a)$ as admissible variations such that the perturbed control $(\bar{\tau}_s + \varepsilon h_1, \bar{c}(a) + \varepsilon h_2(a))$ remains within the admissible set $[\bar{a}, \bar{b}] \times [\bar{c}, \bar{d}]$ for sufficiently small $\varepsilon > 0$. This yields the first-order variational inequality in an appropriate weak sense:

$$
\begin{aligned}
0 &\le \lim_{\varepsilon \to 0^+} \frac{\mathcal{J}(\bar{c}(a) + h_2(a)\varepsilon, \bar{\tau}_s + h_1\varepsilon) - \mathcal{J}(\bar{c}(a), \bar{\tau}_s)}{\varepsilon} \\
&= W_1 \left[\int_0^1 \bar{\tau}_s \bar{v} da + \int_1^2 \left(L_1 - l_1 \bar{\tau}_s\right) \bar{v} da + \bar{f}_0 \right] + W_2 \int_0^{T_m} \int_0^1 \bar{\tau}_s \bar{i} da dt + W_2 \bar{f}_{00} \\
&\quad + W_2 \left\{ \int_0^{T_m} \int_1^2 \left[(L_1 - l_1 \bar{\tau}_s)\bar{i} + \frac{C_2}{2}(L_1 - l_1 \bar{\tau}_s)\bar{c}_s^2 \bar{i} + \frac{C_1 D_0}{2}(L_1 - l_1 \bar{\tau}_s)^2 (a-1)\bar{i} \right] da dt \right\} \\
&= W_1 \left[\int_0^1 (\bar{\phi}, \bar{\phi}^*, \bar{v}) \mathcal{L}_1^*(\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3) da + \int_1^2 (\bar{\phi}, \bar{\phi}^*, \bar{v}) \mathcal{L}_1^*(\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3) da + \bar{f}_0 \right] \\
&\quad + W_2 \left(\int_0^{T_m} \int_0^1 \mathcal{L}_{22}^* \bar{\lambda}_5 \bar{i} da dt + \int_0^{T_m} \int_1^2 \mathcal{L}_{22}^* \bar{\lambda}_5 \bar{i} da dt + \bar{f}_{00} \right) \\
&= W_1 \left[\int_0^1 (\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3) \mathcal{L}_1(\bar{\phi}, \bar{\phi}^*, \bar{v}) da + \int_1^2 (\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3) \mathcal{L}_1(\bar{\phi}, \bar{\phi}^*, \bar{v}) da + \bar{f}_0 \right] \\
&\quad + W_2 \left(\int_0^{T_m} \int_0^1 \bar{\lambda}_5 \mathcal{L}_{22} \bar{i} da dt + \int_0^{T_m} \int_1^2 \bar{\lambda}_5 \mathcal{L}_{22} \bar{i} da dt + \bar{f}_{00} \right) \\
&= -h_1 l_1 \left\{ \bar{g}_3 + \bar{g}_4 + W_1 \bar{g}_1 + W_1 \int_1^2 C_1 D_0 (L_1 - l_1 \bar{\tau}_s)(a-1) da + W_2 \bar{g}_2 \right. \\
&\quad \left. + W_2 \int_0^{T_m} \int_1^2 C_1 D_0 (L_1 - l_1 \bar{\tau}_s)(a-1) \bar{I} da dt \right\} \\
&\quad + \int_1^2 h_2(a)(L_1 - l_1 \bar{\tau}_s) \left[W_1 C_2 \bar{c}_s + W_2 C_2 \bar{c}_s \int_0^{T_m} \bar{I} dt + (\bar{\lambda}_1 - \bar{\lambda}_2) k_s \bar{T} \bar{V} dt \right] da \\
&\triangleq -h_1 \mathcal{K}_1 + \int_1^2 h_2(a) \mathcal{K}_2(a) da,
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{f}_0 &= \int_0^1 h_1 \bar{V} da - \int_1^2 l_1 h_1 \left[\bar{V} + C_1 D_0 (L_1 - l_1 \bar{\tau}_s)(a-1) + \frac{C_2}{2} \bar{c}_s^2 \right] da \\
&\quad + \int_1^2 h_2(a) C_2 (L_1 - l_1 \bar{\tau}_s) \bar{c}_s da, \\
\bar{f}_{00} &= \int_0^{T_m} \int_0^1 h_1 \bar{I} da dt - l_1 h_1 \int_0^{T_m} \int_1^2 \left[1 + \frac{C_2}{2} \bar{c}_s^2 + (L_1 - l_1 \bar{\tau}_s) C_1 D_0 (a-1) \right] \bar{I} da dt \\
&\quad + \int_0^{T_m} \int_1^2 h_2(a)(L_1 - l_1 \bar{\tau}_s) C_2 \bar{c}_s \bar{I} da dt. \\
\bar{g}_1 &= \int_0^1 -\frac{\bar{V}}{l_1} da + \int_1^2 \left(\bar{V} + \frac{C_2 \bar{c}_s^2}{2} \right) da, \\
\bar{g}_2 &= \int_0^{T_m} \int_0^1 -\frac{\bar{I}}{l_1} da dt + \int_0^{T_m} \int_1^2 \left(1 + \frac{C_2 \bar{c}_s^2}{2} \right) \bar{I} da dt, \\
\bar{g}_3 &= \int_1^2 \left[\bar{\lambda}_1 \left(\lambda_T - d_T \bar{T} - k_s \bar{T} \bar{V} \right) + \bar{\lambda}_2 \left(k_s \bar{T} \bar{V} - d_s \bar{T}^* \right) + \bar{\lambda}_3 \left(p_s \bar{T}^* - d_{vs} \bar{V} \right) \right] da \\
&\quad - \frac{1}{l_1} \int_0^1 \left[\bar{\lambda}_1 \left(\lambda_T - d_T \bar{T} - k_s \bar{T} \bar{V} \right) + \bar{\lambda}_2 \left(k_s \bar{T} \bar{V} - d_s \bar{T}^* \right) + \bar{\lambda}_3 \left(p_s \bar{T}^* - d_{vs} \bar{V} \right) \right] da \\
&\quad + \int_1^2 \bar{c}_s k_s \bar{T} \bar{V} \left(\bar{\lambda}_1 - \bar{\lambda}_2 \right) da, \\
\bar{g}_4 &= \int_0^{T_m} \int_0^1 \frac{\bar{\lambda}_5}{l_1} \left[\frac{\partial \bar{I}}{\partial t} + \left(\mu + \delta(\bar{V}) \right) \bar{I} \right] da dt - \int_0^{T_m} \int_1^2 \bar{\lambda}_5 \left[\frac{\partial \bar{I}}{\partial t} + \left(\mu + \delta(\bar{V}) + \bar{\sigma}_1 \right) \bar{I} \right] da dt.
\end{aligned}
\tag{12}
$$

If $\bar{c}_s(a)$ lies in the interior ($\bar{c} < \bar{c}_s(a) < \bar{d}$) and $\bar{a} < \bar{\tau}_s < \bar{b}$, then both positive and negative variations are admissible. For the inequality

$$
\lim_{\varepsilon \to 0^+} \frac{\mathcal{J}(\bar{c}(a) + h_2(a)\varepsilon, \bar{\tau}_s + h_1\varepsilon) - \mathcal{J}(\bar{c}(a), \bar{\tau}_s)}{\varepsilon} \ge 0
$$

to hold for both, we must have:

$$\mathcal{K}_1 = 0, \qquad \mathcal{K}_2(a) = 0,$$

which yield the interior formulas for $\bar{\epsilon}_s(a)$ and $\bar{\tau}_s$:

$$\bar{\epsilon}_s(a) = \frac{\left(\bar{\lambda}_2(a) - \bar{\lambda}_1(a)\right) k_s \bar{T}(a) \bar{V}(a)}{C_2\left(W_1 + W_2 \int_0^{T_m} \bar{I}(a,t)dt\right)}, \text{ for } a \in [1,2],$$

$$\bar{\tau}_s = \frac{L_1}{l_1} + \frac{W_1\bar{g}_1 + W_2\bar{g}_2 + \bar{g}_3 + \bar{g}_4}{l_1 C_1 D_0\left(W_1 \int_1^2 (a-1)da + W_2 \int_0^{T_m}\int_1^2 (a-1)\bar{I}(a,t)dadt\right)}.$$

If $\bar{\epsilon}_s(a)$ or $\bar{\tau}_s$ attains a boundary value, then only one-sided variations are allowed, leading to the following inequalities:

$$\begin{cases} \mathcal{K}_1 \geq 0, \quad \mathcal{K}_2(a) \geq 0, \quad \text{if } \bar{\tau}_s = \bar{a},\ \bar{\epsilon}_s(a) = \bar{c}, \\ \mathcal{K}_1 \geq 0, \quad \mathcal{K}_2(a) \leq 0, \quad \text{if } \bar{\tau}_s = \bar{a},\ \bar{\epsilon}_s(a) = \bar{d}, \end{cases}$$

and analogous conditions hold when $\bar{\tau}_s$ reaches its upper bound. Combining the interior and boundary cases yields the projection form stated in Theorem 6.

This final characterization is crucial, because it transforms the abstract optimal-control problem into a concrete system of equations that can be solved numerically. It provides explicit, interpretable formulas that dictate how the optimal control should be adjusted based on the evolving biological and epidemiological dynamics, directly linking the mathematical solution to practical treatment insights. □

4.1.3. Existence of the optimal control

Since solutions of first-order partial differential equations are often non-smooth, the existence of an optimal control cannot be guaranteed in general. To address this, we apply Ekeland's Variational Principle, which ensures the existence of an approximating minimum point. By iteratively applying the principle, we can construct a minimizing sequence that converges to the optimal solution.

Theorem 7. *The function $\mathcal{J} : L^1(0,2) \times L^1(0,2) \to (-\infty, +\infty)$ is lower semi-continuous.*

Proof. Let $(\tau_{sn}, \epsilon_n) \to (\tau_s, \epsilon)$ as $n \to \infty$ in $L^1(0,2) \times L^1(0,2)$, where $\epsilon_n = \epsilon = 0$ for $a \in [0,1)$ and $\epsilon_n = \epsilon_{sn}$ for $a \in [1,2]$. Assume that $x = (T, T^*, V, S, I)$ and $x_n = (T_n, T_n^*, V_n, S_n, I_n)$ are the solutions for system (6) corresponding to $u = (\tau_s, \epsilon)$ and $u_n = (\tau_{sn}, \epsilon_n)$, respectively. For $a \in [0,1)$, denote $x = (T_1, T_1^*, V_1, S, I_1)$, $x_n = (T_{1n}, T_{1n}^*, V_{1n}, S_n, I_{1n})$; for $a \in [1,2]$, denote $x = (T_2, T_2^*, V_2, S, I_2)$, $x_n = (T_{2n}, T_{2n}^*, V_{2n}, S_n, I_{2n})$. Since the solution of system (6) with respect to the control variables is Lipschitz, we conclude that as $n \to \infty$, there exists

$$\begin{aligned} &\left(T_{1n}, T_{1n}^*, V_{1n}\right) \to \left(T_1, T_1^*, V_1\right) \text{ in } L^\infty(0,1),\ \left(T_{2n}, T_{2n}^*, V_{2n}\right) \to \left(T_2, T_2^*, V_2\right) \text{ in } L^\infty(1,2), \\ &S_n \to S \text{ in } L^\infty(0,T_m),\ I_{1n} \to I_1 \text{ in } L^\infty(0,T_m, L^1(0,1)), \\ &I_{2n} \to I_2 \text{ in } L^\infty(0,T_m, L^1(1,2)). \end{aligned}$$

Based on Theorem 5 in [54], there exists a subsequence, denoted by the same notation (τ_{sn}, ϵ_n), such that as $n \to \infty$, we have

$$\begin{aligned} \tau_{sn}V_{1n} &\to \tau_s V_1,\ a.e. \text{ in } (0,1),\ (L_1 - l_1\tau_{sn})V_{2n} \to (L_1 - l_1\tau_s)V_2,\ a.e. \text{ in } (1,2), \\ (L_1 - l_1\tau_{sn})\epsilon_{sn}^2 &\to (L_1 - l_1\tau_s)\epsilon_s^2,\ (L_1 - l_1\tau_{sn})^2 \to (L_1 - l_1\tau_s)^2,\ a.e. \text{ in } (1,2), \\ \tau_{sn}I_{1n} &\to \tau_s I_1,\ a.e. \text{ in } (0,1)\times(0,T_m), \\ (L_1 - l_1\tau_{sn})I_{2n} &\to (L_1 - l_1\tau_s)I_2,\ a.e. \text{ in } (1,2)\times(0,T_m), \\ \frac{C_2\epsilon_{sn}}{2}(L_1 - l_1\tau_{sn})I_{2n} &\to \frac{C_2\epsilon_s}{2}(L_1 - l_1\tau_s)I_2,\ a.e. \text{ in } (1,2)\times(0,T_m). \end{aligned}$$

Using Fatou's Lemma [54], we have

$$\begin{aligned} J_1(\epsilon,\tau_s) &= \int_0^1 \tau_s V_1(a)da + \int_1^2 (L_1 - l_1\tau_s)V_2(a)da + \int_1^2 \frac{C_1D_0}{2}(L_1 - l_1\tau_s)^2(a-1)da \\ &\quad + \int_1^2 \frac{C_2\epsilon_s^2}{2}(L_1 - l_1\tau_s)da \\ &= \int_0^1 \liminf_{n\to\infty} \tau_{sn}V_{1n}(a)da + \int_1^2 \liminf_{n\to\infty}(L_1 - l_1\tau_{sn})V_{2n}(a)da \\ &\quad + \int_1^2 \liminf_{n\to\infty} \frac{C_1D_0}{2}(L_1 - l_1\tau_{sn})^2(a-1)da + \int_1^2 \liminf_{n\to\infty} \frac{C_2\epsilon_{sn}^2}{2}(L_1 - l_1\tau_{sn})da \\ &\leq \liminf_{n\to\infty} \left[\int_0^1 \tau_{sn}V_{1n}da + \int_1^2 (L_1 - l_1\tau_{sn})V_{2n}da + \int_1^2 \frac{C_1D_0}{2}(L_1 - l_1\tau_{sn})^2(a-1)da \right. \\ &\quad \left. + \int_1^2 \frac{C_2\epsilon_{sn}^2}{2}(L_1 - l_1\tau_{sn})da\right]. \end{aligned}$$

$$
\begin{aligned}
J_2(\epsilon,\tau_s) &= \int_0^{T_m}\int_0^1 \tau_s I_1(a,t)dadt + \int_0^{T_m}\int_1^2 (L_1 - l_1\tau_s)I_2(a,t)dadt \\
&\quad + \int_0^{T_m}\int_1^2 \frac{C_1D_0}{2}(L_1 - l_1\tau_s)^2(a-1)I_2(a,t)dadt \\
&\quad + \int_0^{T_m}\int_1^2 \frac{C_2\epsilon_s^2}{2}(L_1 - l_1\tau_s)I_2(a,t)dadt \\
&= \int_0^{T_m}\int_0^1 \liminf_{n\to\infty} \tau_{sn} I_{1n}(a,t)dadt + \int_0^{T_m}\int_1^2 \liminf_{n\to\infty}(L_1 - l_1\tau_{sn})I_{2n}(a,t)dadt \\
&\quad + \int_0^{T_m}\int_1^2 \liminf_{n\to\infty} \frac{C_1D_0}{2}(L_1 - l_1\tau_{sn})^2(a-1)I_{2n}(a,t)dadt \\
&\quad + \int_0^{T_m}\int_1^2 \liminf_{n\to\infty} \frac{C_2\epsilon_{sn}^2}{2}(L_1 - l_1\tau_{sn})I_{2n}(a,t)dadt \\
&\le \liminf_{n\to\infty}\left[\int_0^{T_m}\int_0^1 \tau_{sn} I_{1n}(a,t)dadt + \int_0^{T_m}\int_1^2 (L_1 - l_1\tau_{sn})I_{2n}(a,t)dadt\right. \\
&\qquad + \int_0^{T_m}\int_1^2 \frac{C_1D_0}{2}(L_1 - l_1\tau_{sn})^2(a-1)I_{2n}(a,t)dadt \\
&\qquad \left. + \int_0^{T_m}\int_1^2 \frac{C_2\epsilon_{sn}^2}{2}(L_1 - l_1\tau_{sn})I_{2n}(a,t)dadt\right].
\end{aligned}
$$

Combining the above two inequalities, we derive:

$$\mathcal{J}(\epsilon,\tau_s) \le \liminf_{n\to\infty} \mathcal{J}(\epsilon_n,\tau_{sn}).$$

It follows that $\mathcal{J}$ is lower semi-continuous with respect to L^1 convergence. □

Based on Ekeland's Variational Principle, we can construct a minimizing sequence as follows: for $\varepsilon > 0$, there exists $(\epsilon^\varepsilon,\tau_s^\varepsilon) \in L^1(0,2)\times L^1(0,2)$ such that

(1) $\mathcal{J}(\epsilon^\varepsilon,\tau_s^\varepsilon) \le \inf_{(\epsilon(a),\tau_s)\in\mathcal{U}} \mathcal{J}(\epsilon,\tau_s) + \varepsilon$,

(2) $\mathcal{J}(\epsilon^\varepsilon,\tau_s^\varepsilon) = \min_{(\epsilon(a),\tau_s)\in\mathcal{U}} \mathcal{J}_\varepsilon$,

where $\mathcal{J}_\varepsilon = \mathcal{J}(\epsilon,\tau_s) + \sqrt{\varepsilon}\left(\|\tau_s^\varepsilon - \tau_s\|_{L^1(0,2)} + \|\epsilon^\varepsilon - \epsilon\|_{L^1(0,2)}\right)$.

Denote $(T^\varepsilon,(T^*)^\varepsilon,V^\varepsilon,S^\varepsilon,I^\varepsilon)$ and $\left(\lambda_1^\varepsilon,\lambda_2^\varepsilon,\lambda_3^\varepsilon,\lambda_4^\varepsilon,\lambda_5^\varepsilon\right)$ as the solutions of control equations (6) and adjoint system (11), respectively, corresponding to $(\epsilon^\varepsilon,\tau_s^\varepsilon)$. We further characterize the minimizing sequence $\{(\epsilon^\varepsilon,\tau_s^\varepsilon)\}$ in the following.

Theorem 8. *If $(\epsilon^\varepsilon,\tau_s^\varepsilon)$ is an optimal-control pair minimizing the function $\mathcal{J}_\varepsilon$, then*

$$
\epsilon^\varepsilon = \begin{cases} 0, \text{ for } a\in[0,1), \\ \epsilon_s^\varepsilon = \min\left\{\bar{d}, \max\left\{\bar{c}, \dfrac{\left(\lambda_2^\varepsilon - \lambda_1^\varepsilon\right)k_s T^\varepsilon V^\varepsilon - \frac{\sqrt{\varepsilon}\|h_2\|_{L^\infty(1,2)}}{h_2}}{C_2\left(W_1 + W_2\int_0^{T_m} I^\varepsilon dt\right)}\right\}\right\}, \text{ for } a\in[1,2], \end{cases}
$$

$$
\tau_s^\varepsilon = \min\left\{\bar{b}, \max\left\{\bar{a}, \frac{W_1 g_1^\varepsilon + W_2 g_2^\varepsilon + g_3^\varepsilon + g_4^\varepsilon + \frac{\sqrt{\varepsilon}|h_1|}{h_1 l_1}}{l_1 C_1 D_0\left(W_1\int_1^2 (a-1)da + W_2\int_0^{T_m}\int_1^2 (a-1)I^\varepsilon(a,t)dadt\right)}\right\}\right\},
$$

where $g_i^\varepsilon = \bar{g}_i\left(T^\varepsilon,(T^*)^\varepsilon,V^\varepsilon,S^\varepsilon,I^\varepsilon,\lambda_1^\varepsilon,\lambda_2^\varepsilon,\lambda_3^\varepsilon,\lambda_4^\varepsilon,\lambda_5^\varepsilon\right)$ for $i=1,2,3,4$ are defined by (12).

Proof. This result can be obtained by a similar method to the proof used in Theorem 6. □

4.1.4. Uniqueness of the optimal control

Next, we address the question of uniqueness.

Theorem 9. *If C_1, C_2 are sufficiently large, then there exists a unique optimal control $(\bar{\epsilon}(a),\bar{\tau}_s)$ to minimize the objective function $\mathcal{J}$.*

Proof. Define a mapping $\mathcal{G}: \mathcal{U}\to\mathcal{U}$ by

$$\mathcal{G}(\tau_s,\epsilon) = \left(\mathcal{G}_1(\tau_s,\epsilon),\mathcal{G}_2(\tau_s,\epsilon)\right),$$

where

$$\mathcal{G}_1(\tau_s,\epsilon)=\begin{cases}0, \text{ for } a\in[0,1),\\ \min\left\{\bar{d},\max\left\{\bar{c},\dfrac{\left(\lambda_2(a)-\lambda_1(a)\right)k_sT(a)V(a)}{C_2\left(W_1+W_2\int_0^{T_m}I(a,t)dt\right)}\right\}\right\}, \text{ for } a\in[1,2],\end{cases}$$

$$\mathcal{G}_2(\tau_s,\epsilon)=\min\left\{\bar{b},\max\left\{\bar{a},\frac{L_1}{l_1}+\frac{W_1g_1+W_2g_2+g_3+g_4}{l_1C_1D_0\left(W_1\int_1^2(a-1)da+W_2\int_0^{T_m}\int_1^2(a-1)I(a,t)dadt\right)}\right\}\right\}.$$

(T,T^*,V,S,I) and $(\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5)$ represent the solutions of control equations (6) and adjoint system (11), respectively. The functions $g_i=\bar{g}_i(T,T^*,V,S,I,\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5)$ for $i=1,2,3,4$ are given in (12). Using Lipschitz properties of the control and adjoint systems, there exist positive constants $\tilde{K}_4$ and $\tilde{K}_5$ such that

$$\begin{aligned}\left\|\mathcal{G}(\tau_s,\epsilon)-\mathcal{G}(\hat{\tau}_s,\hat{\epsilon})\right\|&=\left\|\mathcal{G}_1(\tau_s,\epsilon)-\mathcal{G}_1(\hat{\tau}_s,\hat{\epsilon})\right\|_{L^\infty(1,2)}+\left|\mathcal{G}_2(\tau_s,\epsilon)-\mathcal{G}_2(\hat{\tau}_s,\hat{\epsilon})\right|\\&\le\left\|\frac{(\lambda_2-\lambda_1)k_sTV}{C_2\left(W_1+W_2\int_0^{T_m}Idt\right)}-\frac{\left(\hat{\lambda}_2-\hat{\lambda}_1\right)k_s\hat{T}\hat{V}}{C_2\left(W_1+W_2\int_0^{T_m}\hat{I}dt\right)}\right\|_{L^\infty(1,2)}\\&\quad+W_1\left|\frac{\int_0^1-\frac{V}{l_1}da+\int_1^2\left(\frac{C_2}{2}\epsilon_s^2+V\right)da}{l_1C_1D_0g_5}-\frac{\int_0^1-\frac{\hat{V}}{l_1}da+\int_1^2\left(\frac{C_2}{2}\hat{\epsilon}_s^2+\hat{V}\right)da}{l_1C_1D_0\hat{g}_5}\right|\\&\quad+\left|\frac{g_3+g_4}{l_1C_1D_0g_5}-\frac{\hat{g}_3+\hat{g}_4}{l_1C_1D_0\hat{g}_5}\right|+W_2\left|\frac{\int_0^{T_m}\int_0^1Idadt}{l_1^2C_1D_0g_5}-\frac{\int_0^{T_m}\int_0^1\hat{I}dadt}{l_1^2C_1D_0\hat{g}_5}\right|\\&\quad+W_2\left|\frac{\int_0^{T_m}\int_1^2\left(1+\frac{C_2\epsilon_s^2}{2}\right)Idadt}{l_1C_1D_0g_5}-\frac{\int_0^{T_m}\int_1^2\left(1+\frac{C_2\hat{\epsilon}_s^2}{2}\right)\hat{I}dadt}{l_1C_1D_0\hat{g}_5}\right|\\&\le\left(\frac{\tilde{K}_4}{C_2}+\frac{\tilde{K}_5}{C_1}\right)\left(|\tau_s-\hat{\tau}_s|+\|\epsilon_s-\hat{\epsilon}_s\|_{L^\infty(1,2)}\right),\end{aligned}$$

where $g_5=W_1\int_1^2(a-1)da+W_2\int_0^{T_m}\int_1^2(a-1)Idadt$.

Obviously, $\mathcal{G}$ maps $\mathcal{U}$ into $\mathcal{U}$. Furthermore, it follows from the Banach Contraction Theorem that the map $\mathcal{G}$ admits a unique fixed point $(\bar{\epsilon},\bar{\tau}_s)$ for $\frac{\tilde{K}_4}{C_2}+\frac{\tilde{K}_5}{C_1}<1$. Subsequently, we identify that this fixed point is an optimal solution pair by proving that the minimizing sequence $\{(\epsilon^\epsilon,\tau_s^\epsilon)\}$ converges to $(\bar{\epsilon},\bar{\tau}_s)$.

$$\begin{aligned}\left\|(\bar{\tau}_s,\bar{\epsilon})-\left(\tau_s^\epsilon,\epsilon^\epsilon\right)\right\|&=|\bar{\tau}_s-\tau_s^\epsilon|+\|\bar{\epsilon}-\epsilon^\epsilon\|_{L^\infty(0,2)}\\&\le\left|\frac{W_1\bar{g}_1+W_2\bar{g}_2+\bar{g}_3+\bar{g}_4}{l_1C_1D_0\bar{g}_5}-\frac{W_1g_1^\epsilon+W_2g_2^\epsilon+g_3^\epsilon+g_4^\epsilon+\frac{\sqrt{\epsilon}|h_1|}{h_1l_1}}{l_1C_1D_0g_5^\epsilon}\right|\\&\quad+\left\|\frac{(\lambda_2-\lambda_1)k_sTV}{C_2\left(W_1+W_2\int_0^{T_m}Idt\right)}-\frac{\left(\lambda_2^\epsilon-\lambda_1^\epsilon\right)k_sT^\epsilon V^\epsilon-\frac{\sqrt{\epsilon}\|h_2\|_{L^\infty(1,2)}}{h_2}}{C_2\left(W_1+W_2\int_0^{T_m}I^\epsilon dt\right)}\right\|_{L^\infty(1,2)}\\&\le\left(\frac{\tilde{K}_4}{C_2}+\frac{\tilde{K}_5}{C_1}\right)\left(|\bar{\tau}_s-\tau_s^\epsilon|+\|\bar{\epsilon}_s-\epsilon_s^\epsilon\|_{L^\infty(1,2)}\right)\\&\quad+\frac{\sqrt{\epsilon}}{C_2(W_1+W_2)}+\frac{\sqrt{\epsilon}}{\frac{1}{2}l_1^2C_1D_0(W_1+W_2)}.\end{aligned}$$

Therefore

$$|\bar{\tau}_s-\tau_s^\epsilon|+\|\bar{\epsilon}_s-\epsilon_s^\epsilon\|_{L^\infty(1,2)}\le\frac{\sqrt{\epsilon}/C_2+2\sqrt{\epsilon}/(l_1^2C_1D_0)}{\left[1-\left(\frac{\tilde{K}_4}{C_2}+\frac{\tilde{K}_5}{C_1}\right)\right](W_1+W_2)}.$$

Choosing sufficiently large values of C_1 and C_2, we derive

$$\lim_{\epsilon\to0}\left(|\bar{\tau}_s-\tau_s^\epsilon|+\|\bar{\epsilon}_s-\epsilon_s^\epsilon\|_{L^\infty(1,2)}\right)=0.$$

Since $\mathcal{J}$ is lower semi-continuous, according to Ekeland's Variational Principle, we obtain

$$\mathcal{J}(\epsilon^\epsilon,\tau_s^\epsilon)\le\inf_{(\epsilon,\tau_s)\in\mathcal{U}}\mathcal{J}(\epsilon,\tau_s)+\varepsilon.$$

Thus

$$\mathcal{J}(\bar{\epsilon}, \bar{\tau}_s) \leq \inf_{(\epsilon,\tau_s)\in\mathcal{U}} \mathcal{J}(\epsilon, \tau_s).$$

This means $(\bar{\epsilon}, \bar{\tau}_s)$ is an optimal-control pair to minimize the objective function $\mathcal{J}$, so we obtain the result. □

4.2. Pareto-optimal solutions for the MOP

We have proved the existence of an optimal control for the SOP and given its characterization. Based on the relationship between the optimal solution of the SOP and Pareto-optimal solutions of the MOP, as presented in Theorem 2, we further analyze the Pareto-optimal solutions of the MOP. We investigate how to balance the micro-objective J_1 and the macro-objective J_2, and we derive the optimal coordination of therapy protocols for $\hat{J}$. The following are several special cases. Since drug efficacy satisfies $\epsilon = 0$ for $a \in [0, 1)$ before the initiation of ART, we only present the control $\epsilon = \epsilon_s$ for $a \in [1, 2]$.

Case 1: For $W_1 = 1, W_2 = 0$, we obtain a Pareto-optimal treatment scheme $(\hat{\epsilon}, \hat{\tau}_s)$ for the multi-objective function $\hat{J}$. This scheme corresponds to the optimal control at the within-host level for the single-objective function J_1:

$$\begin{aligned}
\hat{\epsilon}_s(a) &= \min\left\{\bar{d}, \max\left\{\bar{c}, \frac{\left(\hat{\lambda}_2(a) - \hat{\lambda}_1(a)\right)k_s\hat{T}(a)\hat{V}(a)}{C_2}\right\}\right\}, \text{ for } a \in [1, 2], \\
\hat{\tau}_s &= \min\left\{\bar{b}, \max\left\{\bar{a}, \frac{L_1}{l_1} + \frac{2\hat{g}_1 + 2\hat{g}_3}{l_1 C_1 D_0}\right\}\right\},
\end{aligned} \tag{13}$$

where $(\hat{T}, \hat{T}^*, \hat{V}, \hat{S}, \hat{I})$ and $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4, \hat{\lambda}_5)$ are the corresponding solutions for the control system and the adjoint system. The forms of functions $\hat{g}_1$ and $\hat{g}_3$ are given in (12).

Case 2: For $W_1 = 0, W_2 = 1$, we derive a Pareto-optimal treatment scheme $(\tilde{\epsilon}, \tilde{\tau}_s)$ for the multi-objective function $\hat{J}$, which also serves as the optimal control at the between-host level for the single-objective function J_2:

$$\begin{aligned}
\tilde{\epsilon}_s(a) &= \min\left\{\bar{d}, \max\left\{\bar{c}, \frac{\left(\tilde{\lambda}_2(a) - \tilde{\lambda}_1(a)\right)k_s\tilde{T}(a)\tilde{V}(a)}{C_2\int_0^{T_m}\tilde{I}(a,t)dt}\right\}\right\}, \text{ for } a \in [1, 2], \\
\tilde{\tau}_s &= \min\left\{\bar{b}, \max\left\{\bar{a}, \frac{L_1}{l_1} + \frac{\tilde{g}_2 + \tilde{g}_3 + \tilde{g}_4}{l_1 C_1 D_0 \int_0^{T_m}\int_1^2 (a-1)\tilde{I}_2(a,t)dadt}\right\}\right\},
\end{aligned} \tag{14}$$

where $(\tilde{T}, \tilde{T}^*, \tilde{V}, \tilde{S}, \tilde{I})$ and $(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3, \tilde{\lambda}_4, \tilde{\lambda}_5)$ are the corresponding solutions for the control system and adjoint system. The forms of $\tilde{g}_2, \tilde{g}_3, \tilde{g}_4$ are given in (12).

By synthesizing the results from Cases 1 and 2, we derive the necessary conditions for the control pair $(\bar{\epsilon}, \bar{\tau}_s)$ to be the absolutely optimal solutions for the MOP. Specifically, they must satisfy one of the following conditions (A1)–(A3) for any a fixed time point $a \in [1, 2]$ while simultaneously satisfying one of the conditions (B1)–(B3):

$$\text{(A1)} \int_0^{T_m} \bar{I}(a,t)dt = 1,$$

$$\text{(A2)} \min\left\{\frac{\left(\bar{\lambda}_2(a) - \bar{\lambda}_1(a)\right)k_s\bar{T}(a)\bar{V}(a)}{C_2}, \frac{\left(\bar{\lambda}_2(a) - \bar{\lambda}_1(a)\right)k_s\bar{T}(a)\bar{V}(a)}{C_2\int_0^{T_m}\bar{I}(a,t)dt}\right\} > \bar{d},$$

$$\text{(A3)} \max\left\{\frac{\left(\bar{\lambda}_2(a) - \bar{\lambda}_1(a)\right)k_s\bar{T}(a)\bar{V}(a)}{C_2}, \frac{\left(\bar{\lambda}_2(a) - \bar{\lambda}_1(a)\right)k_s\bar{T}(a)\bar{V}(a)}{C_2\int_0^{T_m}\bar{I}(a,t)dt}\right\} < \bar{c},$$

$$\text{(B1)}\ 2\bar{g}_1 + 2\bar{g}_3 = \frac{\bar{g}_2 + \bar{g}_3 + \bar{g}_4}{\int_0^{T_m}\int_1^2 (a-1)\bar{I}(a,t)dadt},$$

$$\text{(B2)} \min\left\{\frac{L_1}{l_1} + \frac{2\bar{g}_1 + 2\bar{g}_3}{l_1 C_1 D_0}, \frac{L_1}{l_1} + \frac{\bar{g}_2 + \bar{g}_3 + \bar{g}_4}{l_1 C_1 D_0 \int_0^{T_m}\int_1^2 (a-1)\bar{I}dadt}\right\} > \bar{b},$$

$$\text{(B3)} \max\left\{\frac{L_1}{l_1} + \frac{2\bar{g}_1 + 2\bar{g}_3}{l_1 C_1 D_0}, \frac{L_1}{l_1} + \frac{\bar{g}_2 + \bar{g}_3 + \bar{g}_4}{l_1 C_1 D_0 \int_0^{T_m}\int_1^2 (a-1)\bar{I}dadt}\right\} < \bar{a},$$

where $(\bar{T}, \bar{T}^*, \bar{V}, \bar{S}, \bar{I})$ and $(\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3, \bar{\lambda}_4, \bar{\lambda}_5)$ are the corresponding solutions for the control system and adjoint system. The functions $\bar{g}_i (i = 1, 2, 3, 4)$ are given in (12).

Table 1
Parameters and initial values of the model.

Parameters	Description	Default values	source
λ_T (/μl/day)	Generation rate of uninfected CD4$^+$T cells	10	[56]
d_T (/day)	Death rate of uninfected CD4$^+$T cells	0.01	[57]
k_s (μl /day)	Infection rate of CD4$^+$T cells by virus	2.4×10^{-6}	[57]
d_s (/day)	Death rate of infected CD4$^+$T cells	0.35	[57]
p_s (/day)	The rate of production of virus by infected cells	1500	[57]
d_{vs} (/day)	Clearance rate of free virus before AIDS	3	[57]
Λ (/year)	Susceptible population admission rate for MSM	4161	[57]
μ(/year)	Removal rate	0.0143	[17]
$\beta(V)$ (/year)	Transmission rate	$c\times\beta_0 V$	[17]
$\beta_0 V$	Transmission rate per contact	$\beta_0=4.56\times10^{-8}$	[17]
c (/year)	Sexual contacts number	varied	see text
$\delta(V)$ (/year)	Disease-related death rate	$\delta_i\times V$	[17]
δ_1(/year)	Virion-associated mortality coefficient before treatment	3.871×10^{-6}	[17]
δ_2(/year)	Virion-associated mortality coefficient after treatment	1.584×10^{-6}	[58]
τ_s (year)	The initiation time of treatment	control variable	see text
T_s (year)	The time when the AIDS stage begins	$\tau_s+E(\tau_s)-d_A$	[57]
$E(\tau_s)$(year)	The extended life expectancies	$27.5-2.5\times\tau_s$	[17]
d_A (year)	Duration from AIDS to death	0.7	[17]
σ_1(/year)	Rate of progression to AIDS	$1/(T_s-\tau_s)$	[57]

Case 3: If an absolutely optimal solution — one that is optimal with respect to all objectives — does not exist, we can derive the following Pareto-optimal solutions for the multi-objective function $\hat{J}$ with $0\le W_1\le 1, W_2=1-W_1$:

$$\begin{aligned}\bar{\epsilon}_s(a)&=\min\left\{\bar{d},\max\left\{\bar{c},\frac{(\bar{\lambda}_2(a)-\bar{\lambda}_1(a))k_s\bar{T}(a)\bar{V}(a)}{C_2\left(W_1+(1-W_1)\int_0^{T_m}\bar{I}(a,t)dt\right)}\right\}\right\},\ \text{for } a\in[1,2],\\ \bar{\tau}_s&=\min\left\{\bar{b},\max\left\{\bar{a},\frac{L_1}{l_1}+\frac{W_1\bar{g}_1+(1-W_1)\bar{g}_2+\bar{g}_3+\bar{g}_4}{l_1C_1D_0\left(W_1\int_1^2(a-1)da+(1-W_1)\int_0^{T_m}\int_1^2(a-1)\bar{I}(a,t)dadt\right)}\right\}\right\}.\end{aligned}\tag{15}$$

Note that higher values of W_i indicate that decision-makers place greater emphasis on the objective J_i.

5. Numerical simulations

Building on our theoretical analysis, we apply a finite-difference method for ordinary and partial differential equations to perform numerical simulations of the optimal control using the sample set of parameters given in Table 1. Starting with baseline weights $C_1=200, C_2=10$, we analyze how weights adjustments influence the optimal treatment schedules. The initial drug efficacy is bounded within $[0,0.8]$ (i.e., $\bar{c}=0,\bar{d}=0.8$), and we subsequently modify the upper efficacy limit to assess its impact on determining the optimal ART initiation timing. Setting the baseline contact rate as $c=25$, we explore how variations in post-treatment sexual-contact behaviour affect the initiation timing of ART.

5.1. Optimal coordination of therapy protocols

The trade-off between the within-host objective J_1 and the population-level objective J_2 is mediated through the weighting parameters W_1 and W_2 in the objective function J. Increasing W_1 places greater emphasis on individual health outcomes, including viral-load suppression and minimizing treatment-related disability burden. In contrast, increasing W_2 prioritizes population-level benefits, such as reducing disease incidence and public-health costs.

Guided by these theoretical results, we vary W_1 from 0 to 1 to generate a family of Pareto-optimal strategies. This traces out the Pareto front (Fig. 1(a)), where each point corresponds to a pair (J_1,J_2) achievable under some Pareto-optimal strategy. Every point on this front represents an optimal compromise: improving one objective necessarily worsens the other. To illustrate this trade-off, we select three representative cases to further analyze the corresponding Pareto-optimal initiation time of ART and the drug efficacy for the MOP, consequently providing the optimal coordination of therapy protocols.

When $W_1=1,W_2=0$, Fig. 1 illustrates the optimal therapy scheme for the single micro-level objective J_1. The results indicate that, at the within-host level, the optimal initiation time of ART is 1.6 years post-infection and that drug efficacy should be maintained at a high level for approximately 22.7 years before the onset of AIDS. Table 2 shows that under this treatment scheme, the cumulative viral load is 2.3941×10^5 and the cumulative number of infected individuals during half a year is 33.62. At the between-host level $W_1=0,W_2=1$, treatment should be initiated 10.3 years after infection. To control drug costs, the required drug efficacy is relatively low. Compared to the optimal control for the individual objective, the cumulative number of infections decreases to 18.9 under

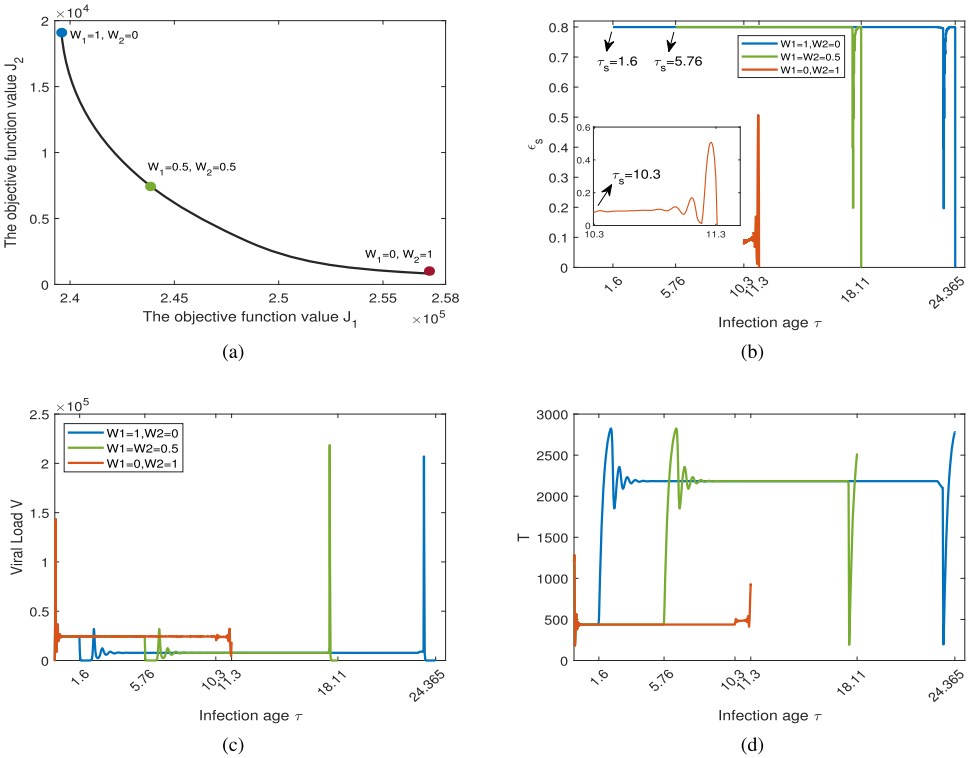

Fig. 1. (a) Pareto front. Each point on this curve corresponds to the objective function values J_1, J_2 under a Pareto-optimal solution. (b) The Pareto-optimal drug efficacy $\epsilon(\tau)$ and initiation time τ_s of ART for three cases $W_1 = 1, W_2 = 0$; $W_1 = W_2 = 0.5$; $W_1 = 0, W_2 = 1$. (c)–(d) The corresponding viral load and concentration of healthy CD4$^+$ T cells.

Table 2
The optimal initiation time of ART and the corresponding objective function values J_1, J_2 under three cases.

Weights	J_1	J_2	$\int_0^{T_s} V(\tau)d\tau$	$\int_0^{T_m}\int_0^{T_s} I(\tau,t)d\tau dt$	τ_s
$W_1 = 1, W_2 = 0$	2.3941×10^5	2.4448×10^4	2.1366×10^5	33.6197	1.5903
$W_1 = W_2 = 0.5$	2.4356×10^5	7.5443×10^3	2.3593×10^5	26.5614	5.7604
$W_1 = 0, W_2 = 1$	2.7582×10^5	0.1027×10^3	2.7577×10^5	18.9442	10.2929

this scheme, but the total viral load increases to 2.7582×10^5. This result demonstrates the conflict between the optimal controls at different levels. The optimal treatment scheme at the within-host level is not optimal for minimizing the objective function J_2.

To address this conflict, we balance the dual-level objectives by adjusting the weights W_1 and W_2. Assuming that within-host and between-host objectives are of equal importance, we set $W_1 = 1/2, W_2 = 1/2$. The resulting Pareto-optimal solution for the MOP suggests that treatment should be initiated 5.76 years after infection. The optimal drug efficacy is higher than that for the case $W_1 = 0, W_2 = 1$ and should be maintained at a high level for approximately 12.35 years before entering the AIDS stage. As shown in Table 2, the total viral load and the number of infected individuals are 2.4356×10^5 and 26.5614, respectively, which lie between the values for the two extreme cases $W_1 = 1, W_2 = 0$ and $W_1 = 0, W_2 = 1$.

These results highlight that if decision-makers prioritize individual benefits — minimizing within-host viral loads while reducing disability burden associated with side effects of drugs and treatment costs — infected individuals should be treated relatively early after infection and drug efficacy should be maintained at high levels for an extended period. However, the survival advantage conferred by early ART may inadvertently amplify transmission risks in subpopulations characterized by persistent viremia and sustained post-treatment high-risk behaviours. Hence, from the population perspective, strategically delayed initiation of ART may offer greater benefits in controlling infected individuals while increasing societal cost-effectiveness. One way to resolve the conflict between optimal treatment schedules at these two scales requires solving for Pareto-optimal solutions, which provide a framework for determining when and how to implement antiviral treatment based on the relative importance of the objectives.

This balancing act is operationalized by systematically varying the weights to trace out the Pareto front. Each pair (W_1, W_2) yields a specific Pareto-optimal treatment strategy (ϵ^*, τ_s^*) via Theorem 6. Large W_1 values favour early initiation and high-efficacy ART, whereas large W_2 values emphasize population-level outcomes, such as transmission control and cost reduction, leading to delayed initiation schemes. The Pareto front thus provides decision-makers with a menu of trade-off solutions, from which context-specific

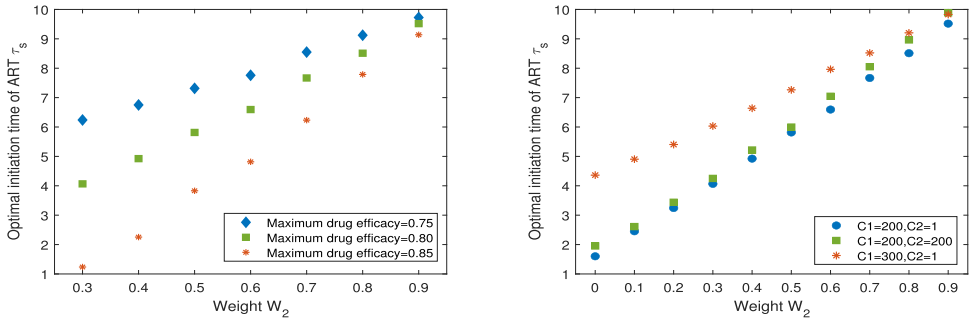

Fig. 2. (a) The effect of maximum drug efficacy on the optimal initiation time of ART. (b) The effect of cost weights C_1, C_2 on the optimal initiation time of ART.

strategies can be chosen according to epidemiological, economic and ethical priorities. Importantly, the "balance" is ultimately a policy decision: our framework does not prescribe a single optimal compromise but instead delivers the complete set of Pareto-optimal trade-offs.

5.2. *The effects of controllable parameters on the optimal initiation time of ART*

Given the heterogeneity in drug efficacy across different infected individuals, understanding the impact of drug efficacy on the optimal treatment strategy is crucial. We vary the maximum drug efficacy to investigate how it influences the optimal treatment time. Fig. 2(a) illustrates that as the maximum drug efficacy increases, the optimal initiation time of ART should be advanced. Specifically, we observe that when the maximum drug efficacy increases from 0.75 to 0.85, the optimal initiation time of ART decreases from 9.7248 years to 9.1386 years after infection for $W_2 = 0.9$. For $W_2 = 0.3$, the optimal initiation time decreases sharply from 6.24 years to 1.2364 years post-infection. These results highlight the critical role of maximum drug efficacy in determining the optimal initiation time of ART, particularly for smaller values of W_2. Notably, when individual benefits are prioritized, higher drug efficacy leads to a substantial advancement in the initiation time of ART.

Unlike Shen et al [17], our objective functions J_1 and J_2 incorporate a disability burden associated with side effects of drugs and treatment costs, weighted by parameters C_1 and C_2, respectively. To understand the influence of these factors on the optimal treatment strategy, we compare the optimal initiation time of ART for different values of C_1 and C_2 in Fig. 2(b). As C_1 or C_2 increases, the optimal initiation time of ART is delayed. This effect is particularly pronounced for increases in C_1, especially when W_1 is large (i.e., W_2 is small).

These findings emphasize that treatment timing is governed by a trade-off between suppressing viral loads, controlling the number of infections, mitigating the loss of healthy life quality due to drug side effects and minimizing treatment costs. If the focus is on viral suppression and infection control, ART should be initiated relatively early. Conversely, for individuals who experience severe side effects, if life quality and treatment costs are prioritized, the initiation of therapy should be delayed.

5.3. *The effects of post-treatment behavioural changes on the optimal initiation time of ART*

The sexual behaviour of HIV-infected individuals may change after initiating treatment. Some individuals voluntarily reduce their sexual activity, whereas some may engage in riskier behaviour after scaled-up ART due to increased confidence in the protective effects of ART [18–20]. Considering this variability, we investigate how changes in sexual-contact rates influence the optimal initiation time of ART.

Assuming that the baseline contact before ART is c, Fig. 3(a) shows the optimal initiation time of ART when sexual contacts decrease after treatment. The results indicate that as sexual contacts decline, the optimal initiation time of ART should be earlier. This is likely because a reduction in sexual contacts after ART decreases the number of new infections, thus lowering associated costs and the objective function value J_1, which allows for an earlier initiation of ART. Fig. 3(c) further explores the relationship between the reduction in sexual-contact rates and the advancement of ART initiation, revealing a nonlinear correlation under different weights W_2. The results demonstrate that there exists a threshold range for sexual contact reduction, approximately between $7/15c$ and $9/15c$, under W_2 values ranging from 0.3 to 0.7. Below this threshold, further reductions in sexual contacts lead to earlier treatment initiation. However, above this threshold, decreasing contact rates no longer results in a continuous advancement of ART initiation. This suggests that aggressively advancing ART initiation may not always be optimal when post-treatment sexual-contact rates are reduced. While reducing contact rates can further reduce new infections, it does not sufficiently offset the increase in associated costs. Conversely, delaying ART initiation slightly may reduce overall costs, thereby minimizing the total objective function J, even if it does not minimize the number of infected individuals. Additionally, we found that if sexual contacts decrease by $2/15$, ART initiation occurs about 0.1 years earlier. When contacts decrease by $7/15$, ART should start 0.25 to 0.42 years earlier, with significant variations in the advancement of ART initiation time depending on the value of W_2. This indicates that as W_2 increases, the timing of ART initiation

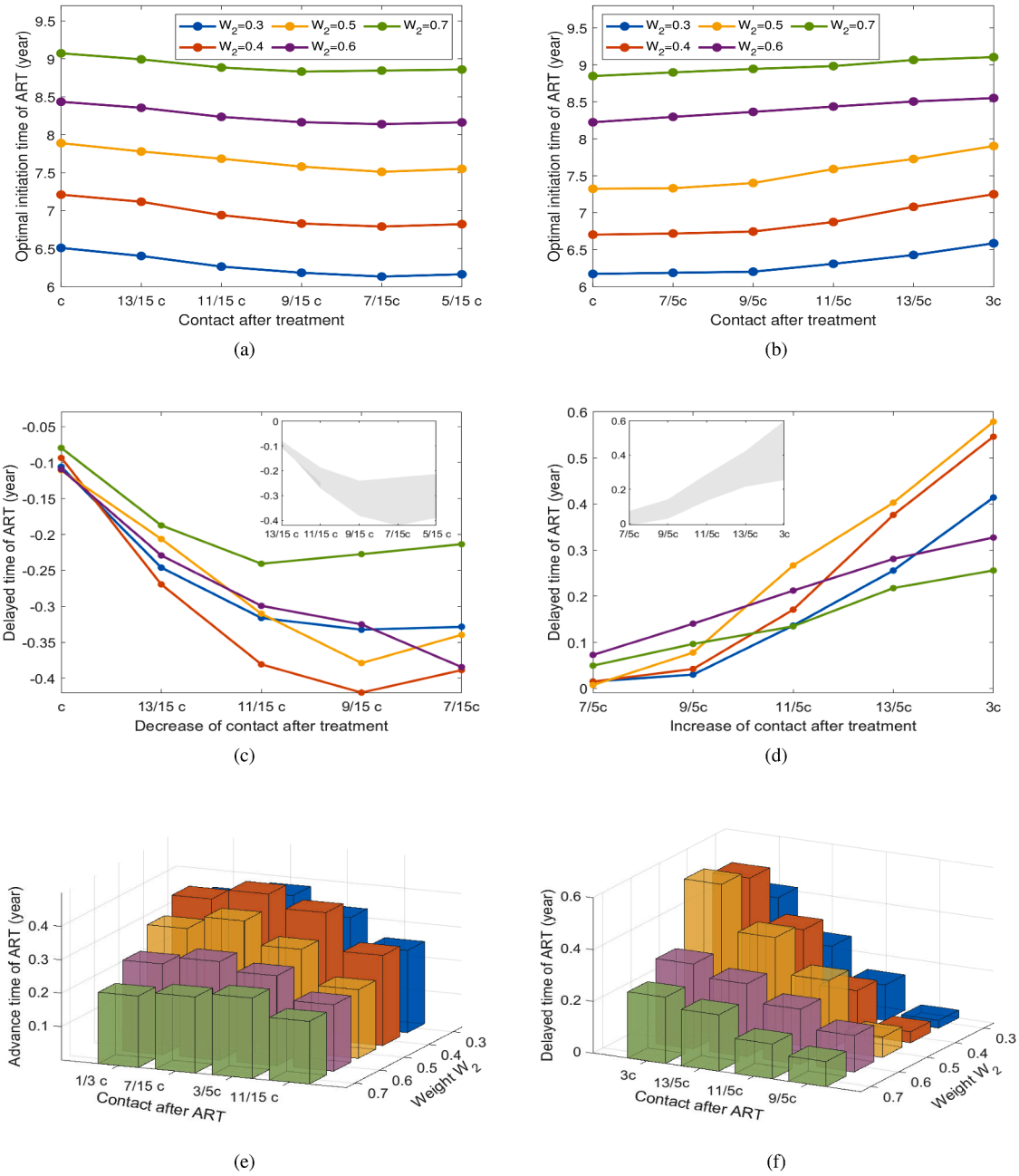

Fig. 3. (a)–(b) The optimal initiation time of ART for decreasing and increasing post-treatment sexual contacts under different weights W_2. To compare the results, we set baseline contacts at $c = 25$ per year in (a) and $c = 75$ per year in (b). (c)–(d) The relationship between the decrease/increase of contacts and the variation of initiation time of ART compared with not changing contact behaviour under different weights W_2. The shaded sub-figure represents the overall variation trends of ART initiation time as sexual contacts vary under weights W_2 from 0.3–0.7. (e)–(f) Relative to the baseline contact c, the variation in the initiation time of ART with changes in weight W_2 and sexual contacts after ART.

becomes more sensitive to reductions in sexual contacts, implying that prioritizing infection control costs (W_2) requires more precise calibration of initiation schedules in response to behavioural changes.

Next, we investigate how the weight W_2 governs the impact of reduced contact rates on the initiation time of ART in Fig. 3(e). Relative to the baseline contact c, our findings reveal that for a fixed reduction in contacts, the advancement of ART initiation first increases (peaking near $W_2 \approx 0.4$) and then decreases as W_2 rises. This occurs because earlier treatment can decrease J_1 while increasing the value of J_2. Therefore, if sexual activity decreases among HIV-infected individuals after treatment, the optimal initiation time of ART should be earlier. However, the extent of this advancement depends on the relative weights W_1 and W_2.

Fig. 3(b) shows the optimal initiation time of ART as post-treatment contacts increase. The results indicate that as sexual contacts increase after ART, the initiation of treatment should be delayed. This is due to the fact that increased contacts after ART may result in more new infections. Delaying treatment helps reduce the infection rate and associated costs, thus lowering J_2. Fig. 3(d) offers a detailed depiction of the relationship between increased sexual contacts and the delayed initiation time of ART. It shows that a higher frequency of sexual behaviour after ART results in a longer delay in starting treatment. Furthermore, as the contact rate rises, the delay in initiation time also grows. This suggests that, to some extent, delaying the start of ART is more beneficial as sexual contact increases. However, there are significant differences in the delay of ART initiation for a fixed contact rate when varying weights W_2 as shown in Fig. 3(f). The results show that there exists a threshold for weight W_2, below which as weight increases, the delayed time should also be longer, but above which the delayed time declines. Notably, this threshold varies with different increases in sexual contacts. When the contact level increases to $9/5c$, the threshold is 0.6. If the contact level exceeds $9/5c$ after treatment, the corresponding threshold is 0.5. These findings suggest that for HIV-infected individuals engaging in more risky post-treatment sexual behaviour, the optimal initiation time of ART should be delayed, and the delay duration is affected by the weights W_2. These findings emphasize the need for dynamic, context-specific protocols — particularly for high-risk cohorts — where treatment timelines adapt to both behavioural patterns and systemic cost-infection trade-offs.

To highlight the novelty and effectiveness of our formulation, we compared our numerical results with two representative multi-scale HIV control studies [17,39]. Shen et al. [17] minimized either the basic reproduction number R_0 or the viral load in two separate single-objective problems and showed that low maximal drug efficacy can lead to conflicting optimal strategies and even make early ART increase R_0. However, they did not determine the optimal initiation time of ART or propose a way to reconcile the conflicting objectives. Our simulations reproduce this phenomenon: under extreme weightings ($W_1 = 1, W_2 = 0$ and $W_1 = 0, W_2 = 1$), we likewise observe that the optimal strategies at the two scales are not aligned. But, in addition, we treat the ART initiation time τ_s as a control variable and construct the Pareto front, quantifying a continuum of optimal trade-offs and identifying compromise strategies unattainable in single-objective models. For example, a balanced weighting ($W_1 = W_2 = 0.5$) yields a strategy initiating ART at 5.76 years, which reduces infections compared to the within-host optimum while keeping the viral load lower than the population-level optimum. Numfor et al. [39] proved the existence of optimal controls for a coupled HIV model but limited their analysis to the post-ART stage and did not examine how initiation timing influenced outcomes. Our formulation incorporates these timing effects and further evaluates how maximal drug efficacy and behavioural factors shift the optimal initiation window.

6. Discussion

Many mathematical models have been developed to investigate the optimal drug schedules for HIV control. Most are implemented either at the within-host or the between-host scale, with a single objective. However, the optimal treatment strategy for an individual may not be optimal for the population and vice versa. In this study, on the basis of a multi-scale model with infection-age structure, we formulated a multi-objective optimization problem to investigate the optimal initiation time and drug efficacy of antiretroviral therapy.

We constructed a bijective mapping to reformulate the initiation timing of ART as a control variable in the control system. Subsequently, we converted a multi-objective optimization problem (MOP) into a single objective optimization problem (SOP) and demonstrated that the optimal solution for the SOP is also the Pareto-optimal solution for the MOP. Furthermore, we proved the existence of the optimal solution for the SOP using Ekeland's variational principle and provided a characterization for the Pareto-optimal schemes. We performed numerical simulations to answer two questions: when and how to implement antiretroviral treatment if there is conflict between optimal-control schemes at the within-host and between-host levels.

Our findings demonstrate that optimal ART strategies must dynamically adapt to evolving priorities between individual outcomes and population-level benefits, while accounting for heterogeneity in regional income and post-treatment behavioural response of an individual. When prioritizing individual benefits — such as viral level, personal quality of life and treatment costs — early ART initiation combined with high drug efficacy is crucial for viral suppression and delaying disease progression. Conversely, when focusing on population-level transmission control and financial sustainability, delayed ART initiation may be preferable if post-treatment behavioural risks persist, as this approach reduces short-term costs despite increasing transmission risks. This trade-off requires context-specific calibration: high-income settings may justify earlier, higher-efficacy ART regimens to align with global targets like UNAIDS 95-95-95, while resource-limited regions may adopt delayed ART strategies paired with targeted prevention campaigns (e.g., condom distribution, education) to balance fiscal constraints and epidemic containment. As drug efficacy improves, earlier initiation of ART becomes more beneficial, particularly for individuals who respond well to therapy or who have access to newer treatments. As drug efficacy improves, the initiation time of ART should be earlier, particularly for individuals who respond well to therapy or who have access to newer treatments. Behavioural variations after ART further influence this balance. Reducing post-treatment sexual activities supports earlier ART to capitalize on fewer transmission opportunities, while high-risk populations with elevated sexual contacts after ART may require careful delays contingent on behavioural thresholds. For such populations, delaying ART until counselling has reduced contacts below critical levels can help reduce untreated individuals with high viral loads resulting in more transmission.

A dual-strategy framework emerges: some subgroups should receive mandatory early ART with maximal efficacy to disrupt transmission chains, supported by behavioural interventions to sustain risk reduction. In contrast, adaptive delayed ART protocols for high-risk groups in resource-limited settings must prioritize cost containment without compromising surveillance, ensuring rigorous viral load monitoring to prevent undetected spread.

While earlier work [17,39] used multi-scale models to explore optimal HIV therapy, our study provides key methodological advances. First, unlike studies that optimized within-host and between-host objectives separately, we formulated a multi-objective framework that directly identifies Pareto-optimal compromises between these often-conflicting scales. Second, Shen et al. [17] only numerically examined how ART timing affects R_0 and found that early therapy may increase infections; however, they did not treat the initiation time as a control variable. We rigorously incorporated the ART initiation time into the control set and overcame the mathematical challenge of control-dependent integral bounds. Together, these contributions establish a quantitative principle for reconciling individual- and population-level treatment priorities that constitutes the main advance of our study: a rigorous, multi-objective, multi-scale optimization framework that not only reconciles conflicting goals across scales but also provides the theoretically optimal initiation time for ART.

Our study has several limitations, which should be acknowledged. First, the model does not account for the emergence of drug resistance, which may lead to an overestimation of the benefits associated with early treatment, as resistant viral strains could compromise long-term therapeutic efficacy [55]. Second, we omit the possibility that AIDS patients may revert to the chronic stage due to treatment effects, which may influence cost-effectiveness analyses, and we ignore the possibility of sexual activity in the AIDS cohort.

In summary, we theoretically obtained the optimal coordination of therapy protocols by solving a multi-objective optimization problem for the multi-scale model and answered the question of when and how to implement the optimal treatment scheme. This may help to develop a synergistic treatment protocol that is beneficial to both individuals and the whole population.

CRediT authorship contribution statement

Yuyi Xue: Writing – original draft, Investigation, Formal analysis; **Mingwang Shen:** Validation, Software, Resources, Methodology, Investigation; **Stacey R. Smith?:** Writing – review & editing, Writing – original draft, Supervision, Project administration; **Yanni Xiao:** Writing – review & editing, Validation, Funding acquisition, Conceptualization.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (No. 12401642 (YXue), 12171387 (MS), 12220101001 (YXiao)), by the Natural Science Foundation of Jiangsu Province (No. BK20240606 (YXue)), by the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (No. NY223198 (YXue)), the National Key R & D Program of China (No. 2022YFC2505100) and Young Elite Professional of the Shaanxi Three Qin Talents Special Support Program (No. 2024STZZK08). Stacey R. Smith? was supported by an NSERC Discovery Grant. Please note that the question mark in "Smith?" is part of her name.

Appendix A. Proof of Theorem 1

Denoting $\beta_1 = \max \beta(V)$, we have

$$\begin{aligned}
&\tilde{L}_1(S,I)(t) \geq S_0 e^{-(\mu+\alpha)t} + \frac{\Lambda}{\mu+\alpha}\left(1-e^{-(\mu+\alpha)t}\right) \geq \min\left\{S_0, \frac{\Lambda}{\mu+\alpha}\right\},\\
&\left|\tilde{L}_1(S,I)\right|(t) \leq \left|S_0 e^{-(\mu+\alpha)t} + \frac{\Lambda}{\mu+\alpha}\left(1-e^{-(\mu+\alpha)t}\right)\right| + \left|\alpha\int_0^t e^{-(\mu+\alpha)(t-s)} S(s)ds\right|\\
&\qquad + \left|\int_0^t e^{-(\mu+\alpha)(t-s)}\frac{S(s)}{N(s)}B(s)ds\right|\\
&\qquad \leq S_0 + \frac{\Lambda}{\mu+\alpha} + \frac{\alpha}{\mu+\alpha}\sup_t S(t) + \frac{\beta_1}{\mu+\alpha}\left(\sup_t\int_0^{\tau_s} I(\tau,t)d\tau + \sup_t\int_{\tau_s}^{T_s} I(\tau,t)d\tau\right)\\
&\qquad < \infty,\\
&0 \leq \int_0^{T_s}\left|\tilde{L}_2(S,I)\right|(\tau,t)d\tau = \int_0^t \left|B_s(t-\tau)\kappa_s(\tau)\right|d\tau + \int_t^{T_s}\left|I_0(\tau-t)\frac{\kappa_s(\tau)}{\kappa_s(\tau-t)}\right|d\tau\\
&\qquad \leq \beta_1 T_m \sup_t\int_0^{T_s} I_s(\tau,t)d\tau + T_s\sup_\tau I_0(\tau) < \infty.
\end{aligned}$$

Hence, $\tilde{\mathcal{L}}$ is a self-mapping on the space X.

Next, we show that $\tilde{\mathcal{L}}$ is a contraction mapping. Define an iterative sequence as

$$S^{(n+1)}(t) = S_0 e^{-(\mu+\alpha)t} + \frac{\Lambda}{\mu+\alpha}\left(1 - e^{-(\mu+\alpha)t}\right) + \int_0^t e^{-(\mu+\alpha)(t-s)} S^{(n)}(s)\left(\alpha - \frac{B^{(n)}(s)}{N^{(n)}(s)}\right) ds,$$

$$I^{(n+1)}(\tau,t) = \begin{cases} B_s{}^{(n)}(t-\tau)\kappa_s(\tau), & \text{for } t > \tau, \\ I_0(\tau - t)\dfrac{\kappa_s(\tau)}{\kappa_s(\tau - t)}, & \text{for } t \le \tau, \end{cases}$$

$$N^{(n+1)}(t) = S^{(n)}(t) + \int_0^{\tau_s} I^{(n)}(\tau,t)d\tau + \int_{\tau_s}^{T_s} I^{(n)}(\tau,t)d\tau,$$

with $B^{(n)}(s) = \int_0^{\tau_s} \beta(V) I^{(n)}(\tau,s)d\tau + \int_{\tau_s}^{T_s} \beta(V) I^{(n)}(\tau,s)d\tau$, which follows by the definition of

$$F_n(t) = \left|S^{(n+1)}(t) - S^{(n)}(t)\right|,$$

$$\check{I}_n(t) = \int_0^{T_s} \left|I^{(n+1)}(\tau,t) - I^{(n)}(\tau,t)\right| d\tau,$$

$$\check{N}_n(t) = F_n(t) + \check{I}_n(t).$$

Assuming $S^{(0)}(t) = 0, I^{(0)}(\tau,t) = 0$, we derive

$$I^{(1)}(\tau,t) = \begin{cases} 0, & \text{for } t > \tau,\ \tau \in [0,T_s], \\ I_0(\tau - t)\dfrac{\kappa_s(\tau)}{\kappa_s(\tau - t)}, & \text{for } t \le \tau,\ \tau \in [0,T_s]. \end{cases}$$

For $n = 0$, the following holds:

$$F_0(t) = \left|S^{(1)}(t) - S^{(0)}(t)\right| \le \left|S_0 e^{-(\mu+\alpha)t} + \frac{\Lambda}{\mu+\alpha}\left(1 - e^{-(\mu+\alpha)t}\right)\right| \le \max\left\{S_0, \frac{\Lambda}{\mu+\alpha}\right\},$$

$$\check{I}_0(t) = \int_0^{T_s} \left|I^{(1)}(\tau,t) - I^{(0)}(\tau,t)\right| d\tau = \int_t^{T_s} I_0(\tau - t)\frac{\kappa_s(\tau)}{\kappa_s(\tau - t)} d\tau \le \int_0^{T_s} I_0(\tau) d\tau,$$

$$\check{N}_0(t) = F_0(t) + \check{I}_0(t) \le \max\left\{S_0, \frac{\Lambda}{\mu+\alpha}\right\} + \int_0^{T_s} I_0(\tau) d\tau.$$

For $n = 1$, we obtain

$$\begin{aligned} F_1(t) &= \left|S^{(2)}(t) - S^{(1)}(t)\right| = \left|\int_0^t e^{-(\mu+\alpha)(t-s)} S^{(1)}(s)\left(\alpha - \frac{B^{(1)}(s)}{N^{(1)}(s)}\right) ds\right| \\ &\le \max\left\{S_0, \frac{\Lambda}{\mu+\alpha}\right\}\frac{\alpha+\beta_1}{\mu+\alpha}, \\ \check{I}_1(t) &= \int_0^{T_s} \left|I^{(2)}(\tau,t) - I^{(1)}(\tau,t)\right| d\tau = \int_0^t B_s^{(1)}(t-\tau)\kappa_s(\tau) d\tau \\ &\le \int_0^t \frac{S^{(1)}(t-\tau)}{N^{(1)}(t-\tau)} \int_0^{T_s} \beta(V) I^{(1)}(s, t-\tau) ds d\tau \\ &\le \beta_1 \int_0^t \int_t^{T_s} I_0(s - t + \tau)\frac{\kappa_s(\tau)}{\kappa_s(s - t + \tau)} ds d\tau \\ &\le \beta_1 T_m \int_0^{T_s} I_0(\tau) d\tau, \\ \check{N}_1(t) &= F_1(t) + \check{I}_1(t) \le \max\left\{S_0, \frac{\Lambda}{\mu+\alpha}\right\}\frac{\alpha+\beta_1}{\mu+\alpha} + \beta_1 T_m \int_0^{T_s} I_0(\tau) d\tau \\ &\le \max\left\{\frac{\alpha+\beta_1}{\mu+\alpha}, \beta_1 T_m\right\}\check{N}_0 \doteq K_1 \check{N}_0. \end{aligned}$$

For $n > 1$, the following holds:

$$\begin{aligned} F_n(t) &= \left|S^{(n+1)}(t) - S^{(n)}(t)\right| \\ &\le \alpha \int_0^t e^{-(\mu+\alpha)(t-s)} \left|S^{(n)}(s) - S^{(n-1)}(s)\right| ds \\ &\quad + \int_0^t e^{-(\mu+\alpha)(t-s)} \int_0^{T_s} \beta(V) \left|\frac{I^{(n)}(\tau,s)S^{(n)}(s)}{N^{(n)}(s)} - \frac{I^{(n-1)}(\tau,s)S^{(n-1)}(s)}{N^{(n-1)}(s)}\right| d\tau ds \\ &\le \alpha \int_0^t \left|S^{(n)}(s) - S^{(n-1)}(s)\right| ds + \int_0^t \int_0^{T_s} \beta(V) \left|G_1(\tau,s)\right| d\tau ds \end{aligned}$$

$$\leq \left(\alpha + 2\beta_1 T_s\right) \int_0^t \left|S^{(n)}(s) - S^{(n-1)}(s)\right| ds + \beta_1(1+T_s) \int_0^t \int_0^{T_s} \left|I^{(n)}(\tau,s) - I^{(n-1)}(\tau,s)\right| d\tau ds$$

$$\leq \max\left\{\alpha + 2\beta_1 T_s, \beta_1(1+T_s)\right\} \int_0^t \left(F_{n-1}(s) + \check{I}_{n-1}(s)\right) ds,$$

$$\check{I}_n(t) = \int_0^{T_s} \left|I^{(n+1)}(\tau,t) - I^{(n)}(\tau,t)\right| d\tau \leq \int_0^t \left|B_s^{(n)}(t-\tau) - B_s^{(n-1)}(t-\tau)\right| d\tau$$

$$\leq \beta_1 \int_0^t \int_0^{T_s} \left| \frac{S^{(n)}(t-\tau)I^{(n)}(s,t-\tau)}{N^{(n)}(t-\tau)} - \frac{S^{(n-1)}(t-\tau)I^{(n-1)}(s,t-\tau)}{N^{(n-1)}(t-\tau)} \right| ds d\tau$$

$$\leq 2\beta_1 T_s \int_0^t \left|S^{(n)}(s) - S^{(n-1)}(s)\right| ds + \beta_1(1+T_s) \int_0^t \int_0^{T_s} \left|I^{(n)}(\tau,s) - I^{(n-1)}(\tau,s)\right| d\tau ds$$

$$\leq \max\left\{2\beta_1 T_s, \beta_1(1+T_s)\right\} \int_0^t \left(F_{n-1}(s) + \check{I}_{n-1}(s)\right) ds,$$

$$\check{N}_n(t) = F_n(t) + \check{I}_n(t) \leq \max\left\{\alpha + 4\beta_1 T_s, 2\beta_1(1+T_s)\right\} \int_0^t \check{N}_{n-1}(s) ds \doteq K \int_0^t \check{N}_{n-1}(s) ds,$$

where

$$G_1(\tau,s) = \frac{S^{(n)}}{N^{(n)}(s)} \left(I^{(n)}(\tau,s) - I^{(n-1)}(\tau,s)\right) + \frac{I^{(n-1)}(\tau,s)}{N^{(n-1)}(s)} \left(S^{(n)}(s) - S^{(n-1)}(s)\right)$$
$$+ \frac{I^{(n-1)}(\tau,s)S^{(n-1)}(s)}{N^{(n)}(s)} - \frac{I^{(n-1)}(\tau,s)S^{(n)}(s)}{N^{(n-1)}(s)}$$
$$= \frac{S^{(n)}(s)}{N^{(n)}(s)} \left(I^{(n)}(\tau,s) - I^{(n-1)}(\tau,s)\right) + \frac{I^{(n-1)}(\tau,s)}{N^{(n-1)}(s)} \left(S^{(n)}(s) - S^{(n-1)}(s)\right)$$
$$+ \frac{I^{(n-1)}(\tau,s)S^{(n-1)}(s)}{N^{(n)}(s)N^{(n-1)}(s)} \left[\left(S^{(n-1)}(s) - S^{(n)}(s)\right) + \int_0^{T_s} \left(I^{(n-1)}(\tau,s) - I^{(n)}(\tau,s)\right) d\tau\right].$$

Hence, we derive

$$\check{N}_2(t) \leq K \int_0^t \check{N}_1(s) ds \leq K \int_0^t K_1 \check{N}_0 ds = K K_1 \check{N}_0 t,$$

$$\check{N}_3(t) \leq K \int_0^t \check{N}_2(s) ds \leq \frac{K^2 K_1 \check{N}_0 t^2}{2},$$

$$\vdots$$

$$\check{N}_n(t) \leq K \int_0^t \check{N}_{n-1}(s) ds \leq \frac{K^{n-1} K_1 \check{N}_0 t^{n-1}}{(n-1)!}.$$

Further, the following holds:

$$\left|S^{(n+m)}(t) - S^n(t)\right| \leq \sum_{j=n+1}^{n+m} N_j \leq K_1 \check{N}_0 \sum_{j=n+1}^{n+m} \frac{K^{n-1} t^{n-1}}{(n-1)!} \to 0, n \to \infty,$$

$$\left|I^{(n+m)}(t) - I^{(n)}(t)\right| \leq \sum_{j=n+1}^{n+m} \int_0^{T_i} \left|I^{(j)}(t) - I^{(j-1)}(t)\right| dt \leq \sum_{j=n+1}^{n+m} \check{N}_j$$
$$\leq K_1 \check{N}_0 \sum_{j=n+1}^{n+m} \frac{K^{n-1} t^{n-1}}{(n-1)!} \to 0, n \to \infty.$$

It is evident that the generated sequence $\{S^{(n)}(t), I^{(n)}(\tau,t)\}$ is a Cauchy sequence in X. Since X is complete, the sequence is convergent. Consequently, there exists $\{S(t), I(\tau,t)\} \in X$, which is the limit of the generated sequence. From the definition of the operator $\tilde{\mathcal{L}}$, we derive

$$\tilde{\mathcal{L}}(S(t), I(\tau,t)) = (S(t), I(\tau,t)).$$

It follows that the limit $(S(t), I(\tau,t))$ is a fixed point of the operator $\tilde{\mathcal{L}}$. This establishes the existence of a solution to the coupled model (3).

Finally, we prove the uniqueness of the solution. Assuming there exist two solutions $(S(t), I(\tau,t))$ and $(\hat{S}(t), \hat{I}(\tau,t))$, set

$$F(t) = |S(t) - \hat{S}(t)|, \ \check{I}(t) = \int_0^{T_s} \left|I(\tau,t) - \hat{I}(\tau,t)\right| d\tau.$$

We can derive $\check{N}(t) \leq K \int_0^t \check{N}(s) ds$, which means $\check{N}(t) = 0$ for any $t > 0$. Hence $F(t) + \check{I}(t) = 0$ with $F(t) \geq 0, \check{I}(t) \geq 0$. Consequently, the equation $(S(t), I(\tau,t)) = (\hat{S}(t), \hat{I}(\tau,t))$ holds, so the solution is unique.

Appendix B. Proof of Theorem 3

Proof of inequality (1) in Theorem 3. For $a \in [0,1)$, it follows from (6) that

$$\frac{d(T-\bar{T})}{da} = \tau_s\left(\lambda_T - d_T T - k_s TV\right) - \bar{\tau}_s\left(\lambda_T - d_T\bar{T} - k_s\bar{T}\bar{V}\right).$$

Integrating, we have

$$|T-\bar{T}|(a) \le \tilde{D}_1\int_0^a |\tau_s-\bar{\tau}_s|da + \hat{D}_1\int_0^a \left(|T-\bar{T}| + |V-\bar{V}|\right)da, \tag{B.1}$$

since $T(a)$ and $V(a)$ is bounded. Similar estimates for $T^*-\bar{T}^*$ and $V-\bar{V}$ lead to

$$\begin{aligned} |T^*-\bar{T}^*|(a) &\le \tilde{D}_2\int_0^a |\tau_s-\bar{\tau}_s|da + \hat{D}_2\int_0^a \left(|T-\bar{T}| + |T^*-\bar{T}^*| + |V-\bar{V}|\right)da,\\ |V-\bar{V}|(a) &\le \tilde{D}_3\int_0^a |\tau_s-\bar{\tau}_s|da + \hat{D}_3\int_0^a \left(|T^*-\bar{T}^*| + |V-\bar{V}|\right)da. \end{aligned} \tag{B.2}$$

Collecting terms, we obtain

$$\begin{aligned} |T-\bar{T}|(a) + |T^*-\bar{T}^*|(a) + |V-\bar{V}|(a) \le{}& \left(\tilde{D}_1+\tilde{D}_2+\tilde{D}_3\right)\int_0^1 |\tau_s-\bar{\tau}_s|da\\ &+ \left(\hat{D}_1+\hat{D}_2+\hat{D}_3\right)\int_0^a \left(|T-\bar{T}| + |T^*-\bar{T}^*| + |V-\bar{V}|\right)da. \end{aligned}$$

Using Gronwall's inequality, the following holds:

$$|T-\bar{T}|(a) + |T^*-\bar{T}^*|(a) + |V-\bar{V}|(a) \le \breve{P}_1\left(1+\hat{P}_1 e^{\hat{P}_1}\right)\int_0^1 |\tau_s-\bar{\tau}_s|da, \tag{B.3}$$

where $\breve{P}_1 = \tilde{D}_1+\tilde{D}_2+\tilde{D}_3$ and $\hat{P}_1 = \hat{D}_1+\hat{D}_2+\hat{D}_3$.

For $a \in [1,2]$, we derive the following estimates:

$$\begin{aligned} |T-\bar{T}|(a) &\le \tilde{D}_4\int_1^a \left(|\tau_s-\bar{\tau}_s| + |\epsilon_s-\bar{\epsilon}_s|\right)da + \hat{D}_4\int_1^a \left(|T-\bar{T}| + |V-\bar{V}|\right)da,\\ |T^*-\bar{T}^*|(a) &\le \tilde{D}_5\int_1^a \left(|\tau_s-\bar{\tau}_s| + |\epsilon_s-\bar{\epsilon}_s|\right)da + \hat{D}_5\int_1^a \left(|T-\bar{T}| + |T^*-\bar{T}^*| + |V-\bar{V}|\right)da,\\ |V-\bar{V}|(a) &\le \tilde{D}_6\int_1^a \left(|\tau_s-\bar{\tau}_s| + |\epsilon_s-\bar{\epsilon}_s|\right)da + \hat{D}_6\int_1^a \left(|T^*-\bar{T}^*| + |V-\bar{V}|\right)da. \end{aligned} \tag{B.4}$$

Using the same method, we have

$$|T-\bar{T}|(a) + |T^*-\bar{T}^*|(a) + |V-\bar{V}|(a) \le \breve{P}_2\left(1+\hat{P}_2 e^{\hat{P}_2}\right)\int_1^2 \left(|\tau_s-\bar{\tau}_s| + |\epsilon_s-\bar{\epsilon}_s|\right)da, \tag{B.5}$$

Integrating (B.3) and (B.5) over the intervals $a=0$ to $a=1$ and $a=1$ to $a=2$ separately, we obtain

$$\begin{aligned} \int_0^1 \left(|T-\bar{T}| + |T^*-\bar{T}^*| + |V-\bar{V}|\right)da &\le \breve{P}_1\left(1+\hat{P}_1 e^{\hat{P}_1}\right)\int_0^1 |\tau_s-\bar{\tau}_s|da,\\ \int_1^2 \left(|T-\bar{T}| + |T^*-\bar{T}^*| + |V-\bar{V}|\right)da &\le \breve{P}_2\left(1+\hat{P}_2 e^{\hat{P}_2}\right)\int_1^2 (|\tau_s-\bar{\tau}_s| + |\epsilon_s-\bar{\epsilon}_s|)da. \end{aligned} \tag{B.6}$$

Based on the characterization of the solution for S, there exist positive constants $\tilde{D}_7, \hat{D}_7$ such that

$$\begin{aligned} |S-\bar{S}|(t) &\le \alpha\int_0^t |S-\bar{S}|(s)ds + \int_0^t \left|\frac{B(s)S(s)}{N(s)} - \frac{\bar{B}(s)\bar{S}(s)}{\bar{N}(s)}\right|ds\\ &\le \alpha\int_0^t |S-\bar{S}|(s)ds + \int_0^t\int_0^1 \left|\frac{\tau_s\beta(V)I(\xi,s)S(s)}{N(s)} - \frac{\bar{\tau}_s\beta(\bar{V})\bar{I}(\xi,s)\bar{S}(s)}{\bar{N}(s)}\right|d\xi ds\\ &\quad + \int_0^t\int_1^2 \left|\frac{(L_1-l_1\tau_s)\beta(V)I(\xi,s)S(s)}{N(s)} - \frac{(L_1-l_1\bar{\tau}_s)\beta(\bar{V})\bar{I}(\xi,s)\bar{S}(s)}{\bar{N}(s)}\right|d\xi ds\\ &\le \tilde{D}_7\left(\int_0^t |S-\bar{S}|(s)ds + \int_0^t\int_0^1 |I-\bar{I}|(\xi,s)d\xi ds + \int_0^t\int_1^2 |I-\bar{I}|(\xi,s)d\xi ds\right)\\ &\quad + \tilde{D}_7\left(\int_0^1 |V-\bar{V}|(\xi)d\xi + \int_1^2 |V-\bar{V}|(\xi)d\xi\right) + \hat{D}_7\int_0^2 |\tau_s-\bar{\tau}_s|da. \end{aligned} \tag{B.7}$$

Next, we consider the equation for $I(a,t)$. For $a \in [0,1)$, the solution of $I(a,t)$ can be described as

$$I(a,t) = \begin{cases} B_1(t-\tau_s a)\kappa_1(a), & \text{for } t > \tau_s a, \\ I_0\left(a - \frac{t}{\tau_s}\right)\frac{\kappa_1(a)}{\kappa_1\left(a-\frac{t}{\tau_s}\right)}, & \text{for } t < \tau_s a, \end{cases}$$

where

$$B_1(t) = I(0,t),\ \kappa_1(a) = e^{-\int_0^a \tau_s(\mu+\delta(V))da}.$$

Without loss of generality, assuming $\tau_s < \bar{\tau}_s$, there is a positive constant $\tilde{P}_1$ such that

$$\begin{aligned} \int_0^1 |I-\bar{I}|(a,t)da &= \int_0^{\frac{t}{\bar{\tau}_s}} |I-\bar{I}|da + \int_{\frac{t}{\bar{\tau}_s}}^{\frac{t}{\tau_s}} |I-\bar{I}|da + \int_{\frac{t}{\tau_s}}^1 |I-\bar{I}|da \\ &\le \tilde{P}_1\left[\frac{1}{\bar{a}}\int_0^t |S-\bar{S}|(s)ds + \int_0^t\int_0^1 |I-\bar{I}|(\xi,s)d\xi ds + \int_0^t\int_1^2 |I-\bar{I}|(\xi,s)d\xi ds\right. \\ &\quad \left. + \int_0^1 |V-\bar{V}|(\xi)d\xi + \int_1^2 |V-\bar{V}|(\xi)d\xi + \left(1+\frac{1}{\bar{a}^2}\right)\int_0^1 |\tau_s-\bar{\tau}_s|d\xi\right]. \end{aligned} \tag{B.8}$$

For $a \in [1,2]$, let $\hat{a} = a-1$. Then the solution of $I(a,t)$ is described as

$$I(\hat{a},t) = \begin{cases} B_2\left(t-(L_1-l_1\tau_s)\hat{a}\right)\kappa_2(\hat{a}), & \text{for } t > (L_1-l_1\tau_s)\hat{a}, \\ I_0\left(\hat{a} - \frac{t}{L_1-l_1\tau_s}\right)\frac{\kappa_2(\hat{a})}{\kappa_2\left(\hat{a}-\frac{t}{L_1-l_1\tau_s}\right)}, & \text{for } t < (L_1-l_1\tau_s)\hat{a}, \end{cases}$$

where

$$B_2(t) = I(1,t),\ k_2(\hat{a}) = e^{-\int_0^{\hat{a}}(L_1-l_1\tau_s)\left[\mu+\delta(V(\hat{a}+1))+\sigma_1\right]d\hat{a}}.$$

Assuming $\tau_s < \bar{\tau}_s$, the inequality $L_1 - l_1\tau_s > L_1 - l_1\bar{\tau}_s$ holds. There is a positive constant $\tilde{P}_2$ such that

$$\begin{aligned} \int_1^2 |I-\bar{I}|(a,t)da &= \int_0^1 |I-\bar{I}|(\hat{a},t)d\hat{a} \\ &= \int_0^{\frac{t}{L_1-l_1\tau_s}} |I-\bar{I}|(\hat{a},t)d\hat{a} + \int_{\frac{t}{L_1-l_1\bar{\tau}_s}}^{\frac{t}{L_1-l_1\tau_s}} |I-\bar{I}|(\hat{a},t)d\hat{a} + \int_{\frac{t}{L_1-l_1\bar{\tau}_s}}^1 |I-\bar{I}|(\hat{a},t)d\hat{a} \\ &\le \tilde{P}_2\left\{\frac{1}{L_1-l_1\bar{b}}\int_0^t\int_0^1 |I-\bar{I}|(\xi,s)d\xi ds + \int_0^1 |V-\bar{V}|(\xi)d\xi + \int_1^2 |V-\bar{V}|(\xi)d\xi\right. \\ &\quad \left. + \left[1+\frac{1}{(L_1-l_1\bar{b})^2}\right]\int_1^2 |\tau_s-\bar{\tau}_s|d\xi\right\}. \end{aligned} \tag{B.9}$$

By combining inequalities (B.7)–(B.9) and applying Gronwall's inequality in its integral form, there exist positive constants $\tilde{D}_8, \hat{D}_8$ such that

$$\begin{aligned} &|S-\bar{S}|(t) + \int_0^1 |I-\bar{I}|(\xi,t)d\xi + \int_1^2 |I-\bar{I}|(\xi,t)d\xi \\ &\le \tilde{D}_8\left(\int_0^t |S-\bar{S}|(s)ds + \int_0^t\int_0^1 |I-\bar{I}|(\xi,s)d\xi ds + \int_0^t\int_1^2 |I-\bar{I}|(\xi,s)d\xi ds\right) \\ &\quad + \hat{D}_8\left(\int_0^2 |\tau_s-\bar{\tau}_s|d\xi + \int_1^2 |\epsilon_s-\bar{\epsilon}_s|(\xi)d\xi\right) \\ &\le \hat{D}_8\left(1+\tilde{D}_8T_m e^{\tilde{D}_8T_m}\right)\left(\int_0^2 |\tau_s-\bar{\tau}_s|d\xi + \int_1^2 |\epsilon_s-\bar{\epsilon}_s|(\xi)d\xi\right). \end{aligned}$$

By integrating both sides of above inequality over the interval $t \in [0,T_m]$, we obtain

$$\begin{aligned} &\int_0^{T_m} |S-\bar{S}|(t)dt + \int_0^{T_m}\int_0^1 |I-\bar{I}|(\xi,t)d\xi dt + \int_0^{T_m}\int_1^2 |I-\bar{I}|(\xi,t)d\xi dt \\ &\le \hat{D}_8T_m\left(1+\tilde{D}_8T_m e^{\tilde{D}_8T_m}\right)\left(\int_0^2 |\tau_s-\bar{\tau}_s|d\xi + \int_1^2 |\epsilon_s-\bar{\epsilon}_s|(\xi)d\xi\right). \end{aligned} \tag{B.10}$$

Finally, by combining inequalities (B.6) and (B.10) and denoting $\tilde{K}_1 = \check{P}_1(1+\hat{P}_1e^{\hat{P}_1}) + \check{P}_2(1+\hat{P}_2e^{\hat{P}_2}) + \hat{D}_8T_m\left(1+\tilde{D}_8T_me^{\tilde{D}_8T_m}\right)$, we derive the results in Theorem 3.

Proof of inequality (1) in Theorem 3. By combining inequalities (B.1), (B.2) and (B.6), for $a \in [0,1)$, the following inequalities hold:

$$\begin{aligned}
|T-\bar{T}|(a) &\le \tilde{D}_1 \int_0^a |\tau_s-\bar{\tau}_s|da + \hat{D}_1\check{P}_1\left(1+\hat{P}_1 e^{\hat{P}_1}\right)\int_0^1 |\tau_s-\bar{\tau}_s|da,\\
|T^*-\bar{T}^*|(a) &\le \tilde{D}_2 \int_0^a |\tau_s-\bar{\tau}_s|da + \hat{D}_2\check{P}_1\left(1+\hat{P}_1 e^{\hat{P}_1}\right)\int_0^1 |\tau_s-\bar{\tau}_s|da\\
|V-\bar{V}|(a) &\le \tilde{D}_3 \int_0^a |\tau_s-\bar{\tau}_s|da + \hat{D}_3\check{P}_1\left(1+\hat{P}_1 e^{\hat{P}_1}\right)\int_0^1 |\tau_s-\bar{\tau}_s|da.
\end{aligned}$$

Over the interval $a \in [0,1)$, there exists a positive constant $\check{P}_3$ such that

$$\|T-\bar{T}\|_{L^\infty(0,1)} + \|T^*-\bar{T}^*\|_{L^\infty(0,1)} + \|V-\bar{V}\|_{L^\infty(0,1)} \le \check{P}_3|\tau_s-\bar{\tau}_s|.$$

Using the same method, we have

$$\begin{aligned}
\|T-\bar{T}\|_{L^\infty(1,2)} + \|T^*-\bar{T}^*\|_{L^\infty(1,2)} + \|V-\bar{V}\|_{L^\infty(1,2)} &\le \check{P}_4\big(|\tau_s-\bar{\tau}_s| + \|\epsilon_s-\bar{\epsilon}_s\|_{L^\infty(1,2)}\big),\\
\big\|S-\bar{S}\big\|_{L^\infty(0,T_m)} &\le \check{P}_5 T_m\big(|\tau_s-\bar{\tau}_s| + \|\epsilon_s-\bar{\epsilon}_s\|_{L^\infty(1,2)}\big),\\
\big\|I-\bar{I}\big\|_{L^\infty(0,T_m;L^1(0,1))} &\le \check{P}_6 T_m\big(|\tau_s-\bar{\tau}_s| + \|\epsilon_s-\bar{\epsilon}_s\|_{L^\infty(1,2)}\big),\\
\big\|I-\bar{I}\big\|_{L^\infty(0,T_m;L^1(1,2))} &\le \check{P}_7 T_m\big(|\tau_s-\bar{\tau}_s| + \|\epsilon_s-\bar{\epsilon}_s\|_{L^\infty(1,2)}\big).
\end{aligned}$$

Collecting above inequalities, we derive the results in Theorem 3.

Appendix C. Derivation of Adjoint System (11)

The adjoint operators $\mathcal{L}_1^*$ and $\mathcal{L}_2^*$ in (10) are derived by applying integration by parts to the left-hand side of (9) to transfer the differential operators from the sensitivity variables (ϕ,ϕ^*,v,s,i) onto the adjoint variables $(\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5)$. The detailed calculations are as follows:

$$\begin{aligned}
&\int_0^1(\lambda_1,\lambda_2,\lambda_3)\mathcal{L}_1(\phi,\phi^*,v)da + \int_1^2(\lambda_1,\lambda_2,\lambda_3)\mathcal{L}_1(\phi,\phi^*,v)da\\
&= \int_0^2 \lambda_1\left[\frac{d\phi}{da} + g(\tau_s)\big(d_T\phi + (1-\epsilon)k_s\phi V + (1-\epsilon)k_sTv\big)\right]da\\
&\quad + \int_0^2 \lambda_2\left[\frac{d\phi^*}{da} - g(\tau_s)\big((1-\epsilon)k_s\phi V + (1-\epsilon)k_sTv - d_s\phi^*\big)\right]da\\
&\quad + \int_0^2 \lambda_3\left[\frac{dv}{da} - g(\tau_s)(p_s\phi^* - d_{vs}v)\right]da\\
&= \int_0^2 \phi\left[-\frac{d\lambda_1}{da} + g(\tau_s)\big(d_T + (1-\epsilon)k_sV\big)\lambda_1\right]da + \int_0^2 v\big[g(\tau_s)(1-\epsilon)k_sT\lambda_1\big]da + \lambda_1\phi\Big|_0^2\\
&\quad - \int_0^2 \phi\big[g(\tau_s)(1-\epsilon)k_sV\lambda_2\big]da + \int_0^2 \phi^*\left[-\frac{d\lambda_2}{da} + g(\tau_s)d_s\lambda_2\right]da - \int_0^2 v\big[g(\tau_s)(1-\epsilon)k_sT\lambda_2\big]da + \lambda_2\phi^*\Big|_0^2\\
&\quad + \int_0^2 v\left[-\frac{d\lambda_3}{da} + g(\tau_s)d_{vs}\lambda_3\right]da - \int_0^2 \phi^* g(\tau_s)p_s\lambda_3 da + \lambda_3 v\Big|_0^2,
\end{aligned}$$

$$\begin{aligned}
&\int_0^{T_m}\lambda_4\mathcal{L}_{21}(s,i)dt = \int_0^{T_m}\lambda_4\left[\frac{ds}{dt} + \tilde{M}_1(s,i)\right]dt\\
&= \int_0^{T_m} s\left[-\frac{d\lambda_4}{dt} + \left(\mu + \left(\frac{1}{N(t)} - \frac{S(t)}{N^2(t)}\right)\tilde{f}_1(t)\right)\lambda_4\right]dt + \int_0^{T_m}\int_0^2 i\left[g(\tau_s)\frac{S(t)}{N(t)}\left(\beta(V) - \frac{\tilde{f}_1(t)}{N(t)}\right)\lambda_4\right]dadt\\
&\quad + \int_0^{T_m}\int_0^2 v\left[g(\tau_s)\frac{S(t)}{N(t)}\frac{\partial\beta(V)}{\partial V}I(a,t)\lambda_4\right]dadt + s\lambda_4\Big|_0^{T_m},
\end{aligned}$$

$$
\begin{aligned}
&\int_0^{T_m}\int_0^1 \lambda_5\mathcal{L}_{22}(s,i)dadt + \int_0^{T_m}\int_{1,2} \lambda_5\mathcal{L}_{22}(s,i)dadt \\
&= \int_0^{T_m}\int_0^2 \lambda_5\left[\frac{\partial i}{\partial a} + g(\tau_s)\frac{\partial i}{\partial t} + g(\tau_s)(\mu+\delta(V)+\sigma)i + g(\tau_s)\frac{\partial\delta(V)}{\partial V}Iv\right]dadt \\
&= \int_0^{T_m}\int_0^2 i\left[-\frac{\partial\lambda_5}{\partial a} - g(\tau_s)\frac{\partial\lambda_5}{\partial t} + g(\tau_s)(\mu+\delta(V)+\sigma)\lambda_5 - \frac{S(t)}{N(t)}g(\tau_s)\left(\beta(V) - \frac{\tilde{f}_1(t)}{N(t)}\right)\lambda_5(0,t)\right]dadt \\
&\quad - \int_0^{T_m} s\left[\left(\frac{1}{N(t)} - \frac{S(t)}{N^2(t)}\right)\tilde{f}_1(t)\lambda_5(0,t)\right]dt \\
&\quad + \int_0^{T_m}\int_0^2 v\left[g(\tau_s)\frac{\partial\delta(V)}{\partial V}I(a,t)\lambda_5 - g(\tau_s)\frac{\partial\beta(V)}{\partial V}\frac{S(t)}{N(t)}I(a,t)\lambda_5(0,t)\right]dadt \\
&\quad + \int_0^{T_m} \lambda_5(2,t)i(T_s,t)dt + g(\tau_s)\int_0^2 \lambda_5(s,T_m)i(s,T_m)dt.
\end{aligned}
$$

By matching the coefficients of (ϕ,ϕ^*,v,s,i) on both sides of (9) and imposing the boundary conditions

$$\lambda_1(2)=0,\lambda_2(2)=0,\lambda_3(2)=0,\lambda_4(T_m)=0,\lambda_5(2,\cdot)=0,\lambda_5(\cdot,T_m)=0,$$

which cancel the boundary terms produced during the integrations by parts, we obtain explicit forms of the adjoint operators $\mathcal{L}_1^*$ and $\mathcal{L}_2^*$ and thereby the adjoint equations with the transversality conditions given in (11).

References

[1] World Health Organization, HIV and AIDS, 2023, (https://www.who.int/news-room/fact-sheets/detail/hiv-aids).
[2] F.J. Palella, Jr, K.M. Delaney, A.C. Moorman, M.O. Loveless, et al., Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection, N Top N. Engl. J. Med. 338 (13) (1998) 853–860.
[3] P.L. Vernazza, L. Troiani, M.J. Flepp, R.W. Cone, J. Schock, Roth, et al., Potent antiretroviral treatment of HIV-infection results in suppression of the seminal shedding of HIV, AIDS 14 (2) (2000) 117–121.
[4] S.J. Reynolds, F. Makumbi, G. Nakigozi, J. Kagaayi, R.H. Gray, M. Wawer, T.C. Quinn, D. Serwadda, HIV-1 Transmission among HIV-1 discordant couples before and after the introduction of antiretroviral therapy, AIDS 25 (4) (2011) 473.
[5] T. Bärnighausen, J.A. Salomon, N. Sangrujee, HIV Treatment as prevention: issues in economic evaluation, PLoS Med. 9 (7) (2012) e1001263.
[6] K.A. Gebo, J.A. Fleishman, R. Conviser, J. Hellinger, F.J. Hellinger, Josephs, et al., Contemporary costs of HIV healthcare in the HAART era, AIDS 24 (17) (2010) 2705–2715.
[7] M. Al-arydah, R.J. Smith?, Adding education to “test and treat”: can we overcome drug resistance?, J. Appl. Math. 2015 (1) (2015) 781270.
[8] R.M. Anderson, S. Gupta, R.M. May, Potential of community-wide chemotherapy or immunotherapy to control the spread of HIV-1, Nature 350 (6316) (1991) 356–359.
[9] J.S.G. Montaner, R. Hogg, E. Wood, T. Kerr, M. Tyndall, A.R. Levy, P.R. Harrigan, The case for expanding access to highly active antiretroviral therapy to curb the growth of the HIV epidemic, The Lancet 368 (9534) (2006) 531–536.
[10] S. Attia, M. Egger, M. Müller, M. Zwahlen, N. Low, Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis, AIDS 23 (11) (2009) 1397–1404.
[11] J.P. Hughes, J.M. Baeten, J.R. Lingappa, A.S. Magaret, A. Wald, de Bruyn, et al., Determinants of per-coital-act HIV-1 infectivity among african HIV-1–serodiscordant couples, J. Infect. Dis. 205 (3) (2012) 358–365.
[12] W.T.S. Consortium, et al., Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies, The Lancet 373 (9672) (2009) 1352–1363.
[13] S.M. Blower, H.B. Gershengorn, R.M. Grant, A tale of two futures: HIV and antiretroviral therapy in san francisco, Science 287 (5453) (2000) 650–654.
[14] R.S. Hogg, K.V. Heath, B. Yip, K.J.P. Craib, M.V. O'Shaughnessy, M.T. Schechter, J.S.G. Montaner, Improved survival among HIV-infected individuals following initiation of antiretroviral therapy, JAMA 279 (6) (1998) 450–454.
[15] S. Baral, A. Rao, P. Sullivan, N. Phaswana-Mafuya, D. Diouf, Millett, et al., The disconnect between individual-level and population-level HIV prevention benefits of antiretroviral treatment, The Lancet HIV 6 (9) (2019) e632–e638.
[16] P.J. Dodd, G.P. Garnett, T.B. Hallett, Examining the promise of HIV elimination by ‘test and treat’ in hyper-endemic settings, AIDS 24 (5) (2010) 729.
[17] M. Shen, Y. Xiao, L. Rong, L.A. Meyers, Conflict and accord of optimal treatment strategies for HIV infection within and between hosts, Math. Biosci. 309 (2019) 107–117.
[18] I.G. Stolte, N.H. Dukers, R.B. Geskus, R.A. Coutinho, J.B. Wit, Homosexual men change to risky sex when perceiving less threat of HIV/AIDS since availability of highly active antiretroviral therapy: a longitudinal study, AIDS 18 (2) (2004) 303–309.
[19] K.K. Venkatesh, T.P. Flanigan, K.H. Mayer, Is expanded HIV treatment preventing new infections?: impact of antiretroviral therapy on sexual risk behaviors in the developing world, AIDS 25 (16) (2011) 1939.
[20] M.H. Katz, S.K. Schwarcz, T.A. Kellogg, J.D. Klausner, J.W. Dilley, S. Gibson, W. McFarland, Impact of highly active antiretroviral treatment on HIV seroincidence among men who have sex with men: san francisco, Am. J. Public Health 92 (3) (2002) 388–394.
[21] R. Hecht, L. Bollinger, J. Stover, W. McGreevey, F. Muhib, C.E. Madavo, D. de Ferranti, Critical choices in financing the response to the global HIV/AIDS pandemic, Health Aff. 28 (6) (2009) 1591–1605.
[22] S. Duwal, S. Winkelmann, C. Schütte, M. von Kleist, Optimal treatment strategies in the context of ‘treatment for prevention’ against HIV-1 in resource-poor settings, PLoS Comput. Biol. 11 (4) (2015) e1004200.
[23] Y. Zhou, Y. Liang, J. Wu, An optimal strategy for HIV multitherapy, J. Comput. Appl. Math. 263 (2014) 326–337.
[24] W. Garira, S.D. Musekwa, T. Shiri, Optimal control of combined therapy in a single strain HIV-1 model, Electron. J. Diff. Equs. 2005 (2005) 1–22.
[25] T. Jang, H.-D. Kwon, J. Lee, Free terminal time optimal control problem of an HIV model based on a conjugate gradient method, Bull. Math. Biol. 73 (2011) 2408–2429.
[26] H.-D. Kwon, J. Lee, S.-D. Yang, Optimal control of an age-structured model of HIV infection, Appl. Math. Comput. 219 (5) (2012) 2766–2779.
[27] A.-M. Croicu, Short-and long-Term optimal control of a mathematical model for HIV infection of CD4$^+$ t cells, Bull. Math. Biol. 77 (2015) 2035–2071.
[28] R.F. Stengel, Mutation and control of the human immunodeficiency virus, Math. Biosci. 213 (2) (2008) 93–102.
[29] P. Ngina, R.W. Mbogo, L.S. Luboobi, HIV Drug resistance: insights from mathematical modelling, Appl. Math. Model. 75 (2019) 141–161.
[30] T.T. Yusuf, F. Benyah, Optimal strategy for controlling the spread of HIV/AIDS disease: a case study of south africa, J. Biol. Dyn. 6 (2) (2012) 475–494.
[31] M. Mastahun, X. Abdurahman, Optimal control of an HIV/AIDS epidemic model with infective immigration and behavioral change, Appl. Math. (Irvine) 8 (1) (2017) 87–106.

[32] K.O. Okosun, O.D. Makinde, I. Takaidza, Impact of optimal control on the treatment of HIV/AIDS and screening of unaware infectives, Appl. Math. Model. 37 (6) (2013) 3802–3820.
[33] I. Takaidza, O.D. Makinde, K.O. Okosun, Computational modelling and optimal control of HIV/AIDS transmission in a community with substance abuse problem, in: A.R. Ansari (Ed.), Advances in Applied Mathematics, 87 of Springer Proceedings in Mathematics & Statistics, Springer, Cham, 2014, pp. 31–40.
[34] B. Seidu, O.D. Makinde, M.I. Daabo, Optimal control analysis of an HIV/AIDS model with linear incidence rate, J. Math. Comput. Sci. 6 (1) (2016) 58–75.
[35] J. Yang, L. Yang, L. Xue, Optimal control of a multi-scale immuno-influenza a transmission model with viral load-dependent infection, Appl. Math. Model. 132 (2024) 1–21.
[36] J.E. Forde, S.M. Ciupe, Quantification of the tradeoff between test sensitivity and test frequency in a COVID-19 epidemic—A multi-Scale modeling approach, Viruses 13 (3) (2021) 457.
[37] J.E. Forde, S.M. Ciupe, Modeling the influence of vaccine administration on COVID-19 testing strategies, Viruses 13 (12) (2021) 2546.
[38] Y. Xue, D. Chen, S.R. Smith?, X. Ruan, S. Tang, Coupling the within-host process and between-host transmission of COVID-19 suggests vaccination and school closures are critical, Bull. Math. Biol. 85 (1) (2023) 6.
[39] E. Numfor, S. Bhattacharya, S. Lenhart, M. Martcheva, Optimal control in coupled within-host and between-host models, Math. Model. Nat. Phenom. 9 (4) (2014) 171–203.
[40] M.G. Law, I. Woolley, D.J. Templeton, N. Roth, J. Chuah, B. Mulhall, P. Canavan, H. McManus, D.A. Cooper, K. Petoumenos, et al., Trends in detectable viral load by calendar year in the australian HIV observational database, J. Int. AIDS Soc. 14 (1) (2011) 10.
[41] M. Shen, Y. Xiao, L. Rong, Global stability of an infection-age structured HIV-1 model linking within-host and between-host dynamics, Math. Biosci. 263 (2015) 37–50.
[42] S.W. Park, B.M. Bolker, Effects of contact structure on the transient evolution of HIV virulence, PLoS Comput. Biol. 13 (3) (2017) e1005453.
[43] H.M. Doekes, C. Fraser, K.A. Lythgoe, Effect of the latent reservoir on the evolution of HIV at the within-and between-Host levels, PLoS Comput. Biol. 13 (1) (2017) e1005228.
[44] M.A. Gilchrist, D. Coombs, Evolution of virulence: interdependence, constraints, and selection using nested models, Theor. Popul. Biol. 69 (2) (2006) 145–153.
[45] A.S. Perelson, P.W. Nelson, Mathematical analysis of HIV-1 dynamics in vivo, SIAM Rev. 41 (1) (1999) 3–44.
[46] D. Coombs, M.A. Gilchrist, C.L. Ball, Evaluating the importance of within-and between-host selection pressures on the evolution of chronic pathogens, Theor. Popul. Biol. 72 (4) (2007) 576–591.
[47] Z. Feng, J. Velasco-Hernandez, B. Tapia-Santos, A mathematical model for coupling within-host and between-host dynamics in an environmentally-driven infectious disease, Math. Biosci. 241 (1) (2013) 49–55.
[48] V.I. Istratescu, Fixed point theory: an introduction, Springer, 1981.
[49] X. Xu, P.J. Antsaklis, Optimal control of switched systems based on parameterization of the switching instants, IEEE Trans. Automat. Contr. 49 (1) (2004) 2–16.
[50] I. Ekeland, On the variational principle, J. Math. Anal. Appl. 47 (2) (1974) 324–353.
[51] V. Barbu, Mathematical methods in optimization of differential systems, 310, Springer Science & Business Media, 2012.
[52] R. Glowinski, J.-L. Lions, Exact and approximate controllability for distributed parameter systems, Acta Numerica 4 (1995) 159–328.
[53] K.L. Teo, B. Li, C. Yu, V. Rehbock, et al., Applied and computational optimal control, Optim. Appl. (2021).
[54] L.C. Evans, R.F. Gariepy, Measure theory and fine properties of functions, CRC press, Boca Raton, Boca Raton, 2015.
[55] R.J. Smith?, J.T. Okano, J.S. Kahn, E.N. Bodine, S. Blower, Evolutionary dynamics of complex networks of HIV drug-resistant strains: the case of san francisco, Science 327 (5966) (2010) 697–701.
[56] R.A. Saenz, S. Bonhoeffer, Nested model reveals potential amplification of an HIV epidemic due to drug resistance, Epidemics 5 (1) (2013) 34–43.
[57] M. Shen, Y. Xiao, L. Rong, G. Zhuang, Global dynamics and cost-effectiveness analysis of HIV pre-exposure prophylaxis and structured treatment interruptions based on a multi-scale model, Appl. Math. Model. 75 (2019) 162–200.
[58] M. Shen, Y. Xiao, L. Rong, L.A. Meyers, S.E. Bellan, Early antiretroviral therapy and potent second-line drugs could decrease HIV incidence of drug resistance, Proc. R. Soc. B: Biol. Sci. 284 (1857) (2017) 20170525.