



Optimal control of an SIR epidemic through finite-time non-pharmaceutical intervention

David I. Ketcheson¹

Received: 3 May 2020 / Revised: 27 May 2021 / Accepted: 13 June 2021 /
Published online: 26 June 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

We consider the problem of controlling an SIR-model epidemic by temporarily reducing the rate of contact within a population. The control takes the form of a multiplicative reduction in the contact rate of infectious individuals. The control is allowed to be applied only over a finite time interval, while the objective is to minimize the total number of individuals infected in the long-time limit, subject to some cost function for the control. We first consider the no-cost scenario and analytically determine the optimal control and solution. We then study solutions when a cost of intervention is included, as well as a cost associated with overwhelming the available medical resources. Examples are studied through the numerical solution of the associated Hamilton-Jacobi-Bellman equation. Finally, we provide some examples related directly to the current pandemic.

Mathematics Subject Classification: 92D30 · 34H05 · 49N90 · 92-10 · 49L12

1 Problem description and assumptions

The classical SIR model of Kermack and McKendrick (1927) is

$$x'(t) = -\gamma\sigma_0 y(t)x(t) \quad (1a)$$

$$y'(t) = \gamma\sigma_0 y(t)x(t) - \gamma y(t) \quad (1b)$$

$$(x(0), y(0)) \in \mathcal{D} := \{(x_0, y_0) : x_0 > 0, y_0 > 0, x_0 + y_0 \leq 1\}, \quad (1c)$$

where $x(t)$, $y(t)$ represent the susceptible and infected populations respectively, while the recovered population is $z(t) = 1 - x(t) - y(t)$. The region \mathcal{D} is forward-invariant and a unique solution exists for all time (Hethcote 2000). While the temporal dynamics

✉ David I. Ketcheson
david.ketcheson@kaust.edu.sa

¹ Computer, Electrical, and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, 4700 KAUST, Thuwal 23955, Saudi Arabia

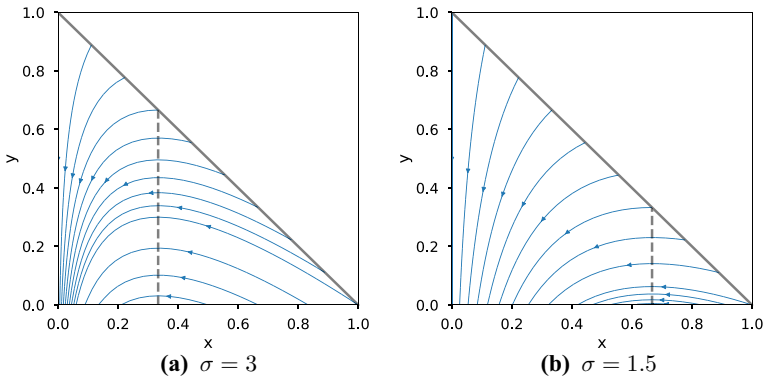


Fig. 1 Dynamics of the SIR model (1) for two values of the basic reproduction number. The critical value $x = 1/\sigma$ is shown with a dashed line

of (1) depend on both σ_0 and γ , the set of trajectories depends only on the basic reproduction number σ_0 . Dynamics for two values of σ_0 are shown in Fig. 1.

The system (1) is at equilibrium if $y(t) = 0$. This equilibrium is stable if and only if $x(t) \leq 1/\sigma_0$, a condition referred to as *herd immunity*. If this condition is not satisfied at the initial time, then $y(t)$ will first increase until it is, and then decrease, approaching zero asymptotically. The SIR model assumes that recovery confers permanent immunity.

For many diseases affecting humans, herd immunity is achieved through vaccination of a sufficient portion of the population. Herein we assume a vaccine is unavailable, so that herd immunity can only be achieved through infection and recovery. Our goal is to minimize $z_\infty := \lim_{t \rightarrow \infty} z(t)$, or equivalently (since $y_\infty = 0$) to maximize the long-time limit of the susceptible fraction: $x_\infty = \lim_{t \rightarrow \infty} x(t)$. This has the effect of minimizing the number of eventual deaths, which would be proportional to z_∞ .

This is equivalent to minimizing the number of deaths, if we assume that some fixed fraction of the recovered population $z(t)$ dies from the disease. From the foregoing it is clear that $x_\infty \leq 1/\sigma_0$. The difference $1/\sigma_0 - x_\infty$ is referred to as *epidemiological overshoot*. For COVID-19, a review of early estimates of σ_0 can be found in Ying et al. (2020), Table 1 with mean 3.28 and median 2.79. In accordance with these estimates, we use a value $\sigma_0 = 3$ in most of the examples in this work. With this value, the SIR model implies that eventually at least two-thirds of the world population will eventually have COVID-19 antibodies; this number is likely to be significantly higher in reality due to epidemiological overshoot. For instance, it can be seen from Fig. 1a that, starting from a fully susceptible population and a small number of infected individuals, in the absence of control the SIR model predicts that over 90% of the population would be infected.

This overshoot can be reduced through non-pharmaceutical intervention (NPI), which is simply a means to reduce contact between infected and susceptible individuals; reductions of this kind occurred for instance as a result of NPIs imposed during the 1918 flu pandemic (Bootsma and Ferguson 2007). We model a NPI control via a time-dependent reproduction number $\sigma(t) \in [0, \sigma_0]$ with the system

$$x'(t) = -\gamma\sigma(t)yx \tag{2a}$$

$$y'(t) = \gamma\sigma(t)yx - \gamma y \tag{2b}$$

$$(x(0), y(0)) \in \mathcal{D} := \{(x_0, y_0) : x_0 > 0, y_0 > 0, x_0 + y_0 \leq 1\}. \tag{2c}$$

A temporary reduction in σ can account for both population-wide interventions and interventions specific to identified infectious (or possibly infectious) individuals. The SIR model with a time-dependent reproduction number (or equivalently, a time-dependent contact rate) has been considered before, for instance in Bootsma and Ferguson (2007); Sun et al. (2020).

Typically, an epidemic does not result in substantial *permanent* change in the contact rate of a population. We therefore assume

$$\sigma(t) = \sigma_0 \text{ for } t > T, \tag{3}$$

i.e., that intervention can only be applied over a finite interval $t \in [0, T]$. Since $x_\infty = 1/\sigma_0$ only at the single point $(x = 1/\sigma_0, y = 0)$, and since the $y = 0$ axis cannot be reached in a finite time, (3) implies that any solution must have $x_\infty < 1/\sigma_0$.

We state the control problem as follows:

Given $(x_0, y_0) \in \mathcal{D}, \sigma_0 > 0, T > 0,$

choose an admissible control $\sigma(t) : [0, T] \rightarrow [0, \sigma_0]$ to minimize

$$J(x(t), y(t), \sigma(t)) = -\lim_{t \rightarrow \infty} x(t) + \int_0^T L(x(t), y(t), \sigma(t))dt \tag{4}$$

subject to (2).

Here J is the objective function that accounts for the desire to minimize infections as well as a running cost of imposing control. We assume throughout that L is convex with respect to $q(t) = 1 - \sigma/\sigma_0$ and bounded uniformly by a constant for all $(x, y) \in \mathcal{D}, \sigma \in [0, \sigma_0]$.

There is a large body of work on compartmental epidemiological models and control for such models; see e.g. Hethcote (2000); Lenhart and Workman (2007) and references therein. A number of works focus on optimal control through vaccination; see e.g. Kar and Batabyal (2011). Other works, such as Yan and Zou (2008); Safi and Gumel (2013); Agosto (2013) focus on explicit modeling of and/or control through quarantined and isolated individuals. A review of work on optimal control in compartmental epidemiological models is presented in Sharomi and Malik (2017), along with the formulation of necessary conditions (based on Pontryagin’s maximum principle) for various extensions of the SIR model. For modeling and control based on even more detailed models incorporating spatial spread and human networks, see e.g. Ferguson et al. (2005).

1.1 Objectives and contributions

The modeling and assumptions in the present work are motivated by the current COVID-19 epidemic, which so far is being managed through broad NPIs and without

a vaccine. In order to understand the effects of NPIs imposed on an entire population, we stick to the simple model (2) rather than explicitly modeling quarantined individuals. Since such population-wide measures cannot be maintained indefinitely, we invoke the finite-time control assumption (3). This assumption is not new (see e.g. Greenhalgh (1988)), but unlike previous works our objective function is still based on the long-term outcome (rather than the outcome at time T). This drastically changes the nature of optimal solutions.

Although the broad motivation for this work comes from the current epidemic, our primary objective is to understand general properties of optimal controls for the variable- σ SIR system (2). To this end, we also investigate solutions in certain asymptotic regimes (such as when there is little or no cost associated with the control). Nevertheless, the values of the key parameters γ and σ_0 for all examples are chosen to fall in the range of current estimates for COVID-19.

One novel aspect of this work is that the problem is posed in terms of the infinite-time limit, but formulated in a way that only requires solution over a finite time interval. Indeed, without this reformulation we found that the problem was extremely ill-conditioned; this reformulation is also needed in order to compute approximate solutions via a Hamilton-Jacobi-Bellman equation. This reformulation is presented in Sect. 2. The main theoretical result is an exact characterization of the optimal control when $L = 0$, given as Theorem 3 in Sect. 3.

Typical results in the literature on control of compartmental epidemiological models are numerical and are based on Pontryagin's weak maximum principle, which gives only necessary conditions for optimality. At best, uniqueness is shown for small times; see e.g. Kirschner et al. (1997); Fister et al. (1998); Lenhart and Feng (2002); Yan and Zou (2008); Kar and Batabyal (2011); Sharomi and Malik (2017). In contrast, here the main result includes a proof of optimality for arbitrarily large times. In Sect. 4 we explore the behavior of optimal solutions for $L \neq 0$ under various interesting cost functions and parameter regimes. Here the results are based on solutions of the relevant Hamilton-Jacobi-Bellman equation, which is both necessary and sufficient for optimality. In Sect. 5 we consider direct application to the COVID-19 pandemic. Some conclusions are drawn in Sect. 6.

The code for all examples in this paper is publicly available (Ketcheson 2021).

2 Formulation over a finite time interval

In this section we reformulate the control problem (4) in terms of the solution over a finite time interval $[0, T]$. This reformulation is necessary both to facilitate the exact solution in Sect. 3 and to arrive at a numerically-tractable problem for computing approximate solutions, as described in Sect. 4.

In general, the solution of (2) depends on the initial data (x_0, y_0) , the control $\sigma(t)$, and time t , so it is natural to write $x(t; \sigma(t), x_0, y_0)$. In what follows it will be convenient to make a slight abuse of notation and write $x(t; \sigma(t))$ or $x(t)$ when there is no chance of confusion.

For a fixed reproduction number, the asymptotic susceptible fraction x_∞ can be obtained from the solution $x(t), y(t)$ at any time t , since solutions of (1) move along contours of x_∞ . Thus we will write $x_\infty(x, y)$ or $x_\infty(x, y, \sigma_0)$.

2.1 A formula for x_∞

In this subsection we review the solution of the SIR model without control (1). It can be shown that $x(t)$ satisfies (see (Harko et al. 2014; Pakes 2015) and Kermack and McKendrick 1927, pp.707-708)

$$x(t)e^{\sigma_0 z(t)} = x_0 e^{\sigma_0 z_0}.$$

Since $z = 1 - x - y$ we define

$$\mu(x, y, \sigma_0) := x(t)e^{-\sigma_0(x(t)+y(t))},$$

which is constant in time for any solution of (1). The trajectories in Fig. 1 are thus also contours of μ . Since $y_\infty = 0$, we have

$$x_\infty = x_0 e^{\sigma_0(x_\infty - x_0 - y_0)} = \mu(x_0, y_0, \sigma_0) e^{\sigma_0 x_\infty}.$$

Setting $w = -x_\infty \sigma_0$ we have

$$w e^w = -x_0 \sigma_0 e^{-\sigma_0(x_0 + y_0)} = -\mu \sigma_0.$$

Thus $w = W_0(-\mu \sigma_0)$ where W_0 is the principal branch of Lambert’s W -function (Pakes 2015), and

$$x_\infty(x, y, \sigma_0) = -\frac{1}{\sigma_0} W_0(-\mu(x, y, \sigma_0) \sigma_0). \tag{5}$$

Formula (5) allows us to rewrite the problem (4) in terms of the state at time $T < \infty$:

Given $(x_0, y_0) \in \mathcal{D}, \sigma_0 > 0, T > 0$,
 choose an admissible control $\sigma(t) : [0, T] \rightarrow [0, \sigma_0]$ to minimize

$$J = -x_\infty(x(T), y(T), \sigma_0) + \int_0^T L(x(t), y(t), \sigma(t)) dt \tag{6}$$

subject to (2).

In what follows we will also require the derivatives of x_∞ with respect to x, y , and μ . Direct computation gives

$$\frac{\partial x_\infty}{\partial y(t)} = -\frac{\sigma_0 x_\infty}{1 - \sigma_0 x_\infty} \tag{7a}$$

$$\frac{\partial x_\infty}{\partial x(t)} = \left(1 - \frac{1}{x(t)\sigma_0}\right) \frac{\partial x_\infty}{\partial y(t)} = \frac{1 - \sigma_0 x(t)}{1 - \sigma_0 x_\infty} \cdot \frac{x_\infty}{x(t)} \quad (7b)$$

$$\frac{\partial x_\infty}{\partial \mu} = \frac{e^{\sigma_0 x_\infty}}{1 - \sigma_0 x_\infty}. \quad (7c)$$

Using these expressions we can also compute the rate of change of x_∞ when some control $\sigma(t)$ is applied:

$$\frac{\partial x_\infty}{\partial t} = \frac{\gamma y x_\infty}{1 - \sigma_0 x_\infty} (\sigma_0 - \sigma(t)). \quad (8)$$

From this we see that the impact of an intervention on x_∞ is independent of $x(t)$ and directly proportional to $y(t)$. This indicates that intervention is more impactful when there is a larger infected population.

2.2 Bounds on x_∞

Now we turn our attention to the SIR system with control (2). Henceforth we assume that $\sigma(t) \in [0, \sigma_0]$ for almost every $t \in [0, T]$; we say that such a control is *admissible*.

It is straightforward to show that (2) has a unique solution for all time for any initial data in \mathcal{D} and any admissible control, by the same arguments used for (1). The proof of the next lemma shows that applying any control $\sigma(t) < \sigma_0$ over any length of time leads to an increase in x_∞ .

Lemma 1 *Let $\sigma_0 > 0$ and $(x_0, y_0) \in \mathcal{D}$ be given. Let $\sigma(t)$ be an admissible control. Then for $t \geq 0$ we have*

$$x_\infty(x(t; \sigma(t)), y(t; \sigma(t)), \sigma_0) \geq x_\infty(x_0, y_0, \sigma_0).$$

Proof Dividing (1b) by (1a) gives

$$\frac{dy}{dx} = -1 + \frac{1}{\sigma(t)x}. \quad (9)$$

Thus reducing $\sigma(t)$ has the effect of increasing dy/dx . Since all trajectories flow to the left (x is a decreasing function of t), this means that the solution trajectory obtained with $\sigma(t)$ lies below that obtained with σ_0 , for all $t > 0$. Since x_∞ is a decreasing function of y , this completes the proof. \square

Thus for any admissible control and any initial data we have

$$x_\infty(x_0, y_0, \sigma_0) \leq x_\infty(x(T), y(T), \sigma_0) \leq 1/\sigma_0.$$

2.3 Existence and necessary conditions for an optimal control

Let us define the Hamiltonian

$$H(x(t), y(t), \sigma(t), \lambda_{1,2}(t), t) = -\lambda_1(t)\gamma\sigma(t)y(t)x(t) + \lambda_2(t)\gamma y(t)(\sigma(t)x(t) - 1) + L(x(t), y(t), \sigma(t)), \tag{10}$$

and the adjoint variables $\lambda_1(t), \lambda_2(t)$, which are required to satisfy

$$\lambda'_1(t) = -\frac{\partial H}{\partial x} = (\lambda_1 - \lambda_2)\gamma\sigma(t)y(t) - \frac{\partial L}{\partial x} \tag{11a}$$

$$\lambda'_2(t) = -\frac{\partial H}{\partial y} = (\lambda_1 - \lambda_2)\gamma\sigma(t)x(t) + \lambda_2\gamma - \frac{\partial L}{\partial y} \tag{11b}$$

$$\lambda_1(T) = -\frac{\partial x_\infty(T)}{\partial x} = \frac{\partial}{\partial x(T)}(-x_\infty(x(T), y(T), \sigma_0)) = \left(1 - \frac{1}{x(T)\sigma_0}\right)\lambda_2(T) \tag{11c}$$

$$\lambda_2(T) = -\frac{\partial x_\infty(T)}{\partial y} = \frac{\partial}{\partial y(T)}(-x_\infty(x(T), y(T), \sigma_0)), \tag{11d}$$

where $x(t), y(t)$ satisfy (2). Note that the final conditions for $\lambda_{1,2}$ can be computed from (7). We have the following result.

Theorem 1 *Let $(x_0, y_0) \in \mathcal{D}$ and $\sigma_0, \gamma, T \geq 0$ be given. Let the running cost L be given such that it is convex with respect to q , bounded uniformly by a constant for all $(x, y) \in \mathcal{D}, \sigma \in [0, \sigma_0]$, and continuously differentiable with respect to x and y . Then there exists an admissible control $\sigma^*(t)$ for (6) and corresponding response $(x^*(t), y^*(t))$ such that J is minimized over the set of admissible controls. Furthermore, there exist adjoint functions $\lambda_{1,2}(t)$ satisfying (11) for almost all $t \in [0, T]$ with $x(t) = x^*(t), y(t) = y^*(t)$, and such that the Hamiltonian is minimized pointwise with respect to σ :*

$$H(x^*(t), y^*(t), \sigma^*(t), \lambda_{1,2}(t), t) = \inf_{\sigma \in [0, \sigma_0]} H(x, y, \sigma, \lambda_{1,2}, t) \tag{12}$$

for almost all $t \in [0, T]$.

Proof The existence of an optimal control is guaranteed by Clarke (2013) Theorem 23.11

since L is convex with respect to $q(t) = 1 - \sigma(t)/\sigma_0$, the state solutions $(x(t), y(t))$ and their derivatives in time are bounded, the system (2) is Lipschitz with respect to x, y , and the control $\sigma(t) = \sigma_0$ is admissible and leads to a finite cost. The second part of the Theorem follows from applying Pontryagin’s weak maximum principle as stated e.g. in Clarke (2013), Theorem 22.2 which applies due to the assumptions on L and since x_∞ is continuously differentiable with respect to x and y . \square

Observe that condition (12) implies that the optimal control $\sigma^*(t)$ satisfies the optimality condition

$$\sigma^*(t) = \max(0, \min(\sigma_0, \hat{\sigma}(t))), \quad (13)$$

where

$$\left. \frac{\partial L}{\partial \sigma} \right|_{\sigma(t)=\hat{\sigma}(t)} = -(\lambda_2(t) - \lambda_1(t))\gamma yx. \quad (14)$$

2.4 Infinite-time control

In this section only, we consider controls that reach the optimal value $x_\infty = 1/\sigma_0$. This is achieved only at $(x, y) = (1/\sigma_0, 0)$, a state that cannot be reached from any other state without imposing some control, and which in any case can only be reached after an infinite time. Thus we momentarily set aside the restriction (3) and consider controls extending up to an arbitrarily large time T . We still require that the system approach a stable equilibrium point as $t \rightarrow \infty$. We assume that $x_0 \geq 1/\sigma_0$, since otherwise the maximum achievable value of x_∞ is x_0 , which would be achieved by taking simply $\sigma(t) = 0$ for all t . We also take $L = 0$ so that an optimal control is any control satisfying

$$\lim_{t \rightarrow \infty} x(t, \sigma(t)) = 1/\sigma_0.$$

There are infinitely many such controls. Two are particularly simple and are of interest.

The first is a constant control $\sigma(t) = \sigma_*(x_0, y_0, \sigma_0)$. By (5) we must have $x_\infty(x_0, y_0, \sigma_*) = 1/\sigma_0$, so σ_* is the solution of

$$W_0(-\mu(x_0, y_0, \sigma_*)\sigma_*) = -\frac{\sigma_*}{\sigma_0}.$$

The second is a bang-bang control in which

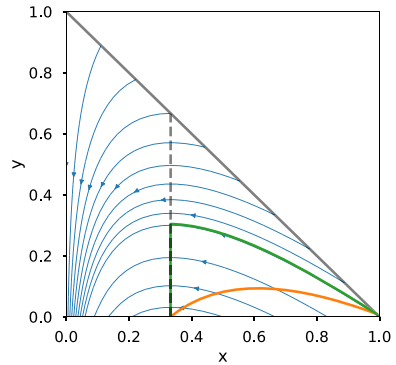
$$\sigma(t) = \begin{cases} \sigma_0 & x > 1/\sigma_0 \\ 0 & x = \sigma_0. \end{cases}$$

The response for each of these controls is shown for a specific example in Fig. 2.

3 Optimal control with $L = 0$

In this section we derive the exact solution of the control problem (6) with $L = 0$ (i.e., when the goal of increasing x_∞ completely trumps any associated costs or other concerns). Then (6) becomes

Fig. 2 Two infinite-time controls that give $x_\infty = 1/\sigma_0$. Here $\sigma_0 = 3$ and $(x_0, y_0) = (0.99, 0.01)$. For the constant control, $\sigma(t) = \sigma_* \approx (1 - 0.4557)\sigma_0$



Given $(x_0, y_0) \in \mathcal{D}, \sigma_0 > 0, T > 0$,
 choose an admissible control $\sigma(t) : [0, T] \rightarrow [0, \sigma_0]$ (15)
 to minimize $J = -x_\infty(x(T), y(T), \sigma_0)$
 subject to (2).

This problem can be reformulated as a minimum-time control problem.

Lemma 2 Let $\sigma^*(t)$ be an optimal control for (15), and let $(x^*(T), y^*(T))$ denote the corresponding terminal state. Then there is no admissible control that reaches $(x^*(T), y^*(T))$ from (x_0, y_0) before time T .

Proof Suppose there were a control $\hat{\sigma}(t)$ that leads to $(x(t^*), y(t^*)) = (x^*(T), y^*(T))$ for some $t^* < T$. Then we could obtain a smaller value of J in (15) by using $\hat{\sigma}$ up to time t^* combined with the choice $\sigma(t) = 0$ for $t > t^*$. This contradicts the optimality of $\sigma^*(t)$. □

Furthermore, the optimal control must be a bang-bang control.

Lemma 3 Let $\sigma(t)$ be an optimal control for (15). Then

$$\sigma(t) = \begin{cases} 0 & \lambda_1(t) < \lambda_2(t) \\ \sigma_0 & \lambda_1(t) > \lambda_2(t) \end{cases} \tag{16}$$

where $\lambda_{1,2}(t)$ are given by (11).

Proof From (10) with $L = 0$, we have

$$\frac{\partial H}{\partial \sigma} = (\lambda_2(t) - \lambda_1(t))\gamma y(t)x(t).$$

The optimality condition then implies (16) except at points where $\partial H/\partial \sigma = 0$ (see e.g. Lenhart and Workman (2007) Ch. 17.

It remains to show that there are no singular arcs. Since $x(t), y(t) > 0$ for $t < \infty$, we have that $\partial H/\partial \sigma = 0$ if and only if $\lambda_1 = \lambda_2$. Suppose (by way of contradiction)

that the latter condition holds on an open interval. Then on that interval we would have (by (11) with $L = 0$):

$$\lambda_1'(t) = (\lambda_1 - \lambda_2)\gamma\sigma y = 0 \tag{17}$$

$$\lambda_2'(t) = (\lambda_1 - \lambda_2)\gamma\sigma x = 0. \tag{18}$$

By continuity, this would imply that $\lambda_1(t) = \lambda_2(t)$ over the whole interval $[0, T]$, and in particular at time T . But then (11c)-(11d) gives

$$\left(1 - \frac{1}{x(T)\sigma_0}\right)\lambda_2(T) = \lambda_2(T).$$

We know from (7) that $\lambda_2(T) \neq 0$, so this is a contradiction. □

This motivates the following lemma.

Lemma 4 *Let (x_0, y_0) and (x_1, y_1) be given such that $x_0, x_1 \geq 1/\sigma_0$ and $x_\infty(x_0, y_0, \sigma_0) \geq x_\infty(x_1, y_1, \sigma_0)$. Let $\sigma(t)$ be a bang-bang control such that $(x(t_1; x_0, y_0, \sigma(t)), y(t_1; x_0, y_0, \sigma(t))) = (x_1, y_1)$ for some $t_1 \geq 0$. Then the minimum value of t_1 is achieved by taking*

$$\sigma(t) = \begin{cases} \sigma_0 & t < t^* \\ 0 & t^* \leq t \leq t_1, \end{cases} \tag{19}$$

where t^* satisfies $x(t^*; x_0, y_0, \sigma_0) = x_1$.

Proof Since $\sigma(t)$ is a bang-bang control, the trajectory $(x(t; \sigma(t)), y(t; \sigma(t)))$ consists of a sequence of segments each of which is a solution of (2) with $\sigma = 0$ (traveling directly downward) or with $\sigma = \sigma_0$ (traveling along a contour of x_∞). Some trajectories of this type are illustrated in Fig. 3. Notice that each trajectory must traverse the same distance in the x -direction; since $x'(t) = -\beta xy$ this travel is faster at larger y values. Meanwhile, the total length of all the downward ($\sigma = 0$) segments is the same for any trajectory, and since for these segments $y'(t) = -\gamma y$, travel is again faster at larger y values. The control given in the lemma makes all these traversals at the largest possible values of y , so it arrives in the shortest time. □

Combining these three lemmas, we obtain the following.

Theorem 2 *Any optimal control for (15) is of the form (19) with $t_1 = T$.*

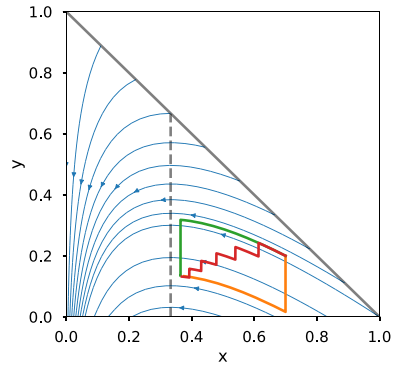
Proof By Lemmas 2 and 3, the optimal control must be bang-bang and must solve the optimal-time problem. Then Lemma 4 applies and gives the stated result. □

We can now give the solution of (15).

Theorem 3 *The optimal control for (15) is unique and is given by*

$$\sigma(t) = \begin{cases} \sigma_0 & t < t^* \\ 0 & t^* \leq t \leq T, \end{cases} \tag{20}$$

Fig. 3 Three different paths between two states, each obtained with a bang-bang control. The top (green) path arrives in the shortest time (color figure online)



where

$$t^* = 0 \text{ if } x_0 \leq \frac{1}{\sigma_0(1 - e^{-\gamma T})}, \tag{21}$$

and otherwise t^* is the unique solution of

$$x(t^*; \sigma_0, x_0, y_0) = \frac{1}{\sigma_0(1 - e^{-\gamma(T-t^*)})}. \tag{22}$$

Proof First, suppose $x(0) \leq 1/\sigma_0$. The claimed optimal control gives $x(T) = x_0$, whereas any other control will give $x(T) < x_0$. Similarly, we see from (2) that the optimal control gives $y(T) = e^{-\gamma T} y_0$ and any other control will lead to a larger value of $y(T)$. Since x_∞ is a decreasing function of y and (for $x < 1/\sigma_0$) an increasing function of x , the proposed control is optimal in this case.

Now suppose $x(0) > 1/\sigma_0$. We reformulate the objective as follows. From (7) we see that x_∞ is a strictly monotone increasing function of μ , so that maximizing x_∞ is equivalent to maximizing μ . Now

$$\begin{aligned} \mu'(t) &= (x'(t) - \sigma_0 x(t)(x'(t) + y'(t)))e^{-\sigma_0(x(t)+y(t))} \\ &= (\sigma_0 - \sigma(t))\gamma x(t)y(t)e^{-\sigma_0(x(t)+y(t))} \\ &= \gamma y(t)(\sigma_0 - \sigma(t))\mu(t). \end{aligned}$$

Thus

$$\mu(t) = \exp\left(\gamma \int_0^t y(\tau)(\sigma_0 - \sigma(\tau))d\tau\right) \mu(0).$$

Thus, maximizing $x_\infty(T)$ is equivalent to maximizing

$$I := \int_0^T y(\tau)(\sigma_0 - \sigma(\tau))d\tau.$$

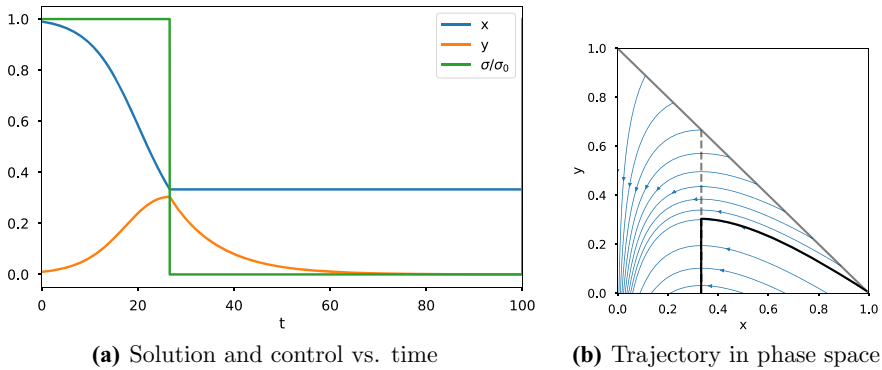


Fig. 4 Typical optimal solution. Here $(x(0), y(0)) = (0.99, 0.01)$, $\beta = 0.3$, and $\gamma = 0.1$

From Theorem 2 we have that

$$\begin{aligned}
 I &= \int_{t^*}^T y(\tau)\sigma_0 d\tau \\
 &= \frac{\sigma_0}{\gamma} y(t^*) \left(1 - e^{-\gamma(T-t^*)}\right).
 \end{aligned}$$

Differentiating with respect to t^* gives

$$\frac{dI}{dt^*} = \sigma_0 y(t^*) \left(\sigma_0 x(t^*) (1 - e^{-\gamma(T-t^*)}) - 1\right). \tag{23}$$

If the inequality in (21) is satisfied then this has no zero and I is maximized by taking $t^* = 0$. If the condition in (21) is not satisfied, then setting the right hand side of (23) equal to zero yields the condition (22). By checking the second derivative, it is easily confirmed that this is indeed a maximum. \square

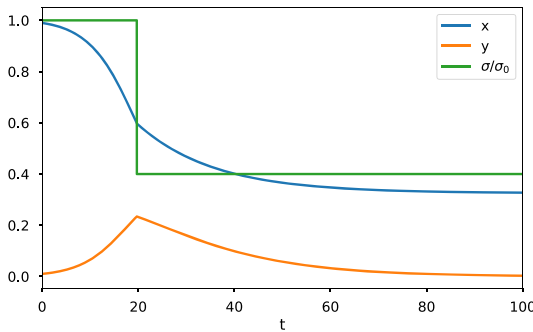
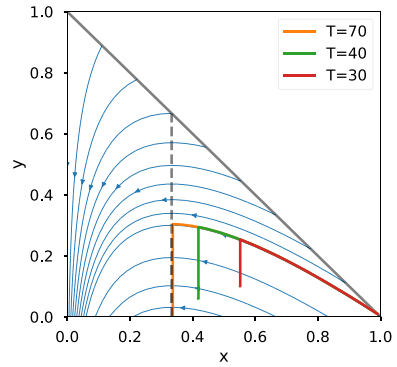
We remark that the above result apparently cannot be obtained via standard sufficiency conditions based on Pontryagin’s maximum principle, due to the nonconvexity of the right hand side of the SIR system (2).

Some optimal solutions for particular instances of (15) are shown in Figs. 4 and 5, all with the same initial data and parameters β, γ but with different final times T . Allowing for a longer intervention (larger T) makes it possible to reach a more optimal value of x_∞ .

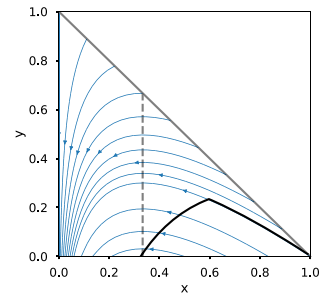
In real-world scenarios, it may not be possible to apply the maximum control $\sigma(t) = 0$. Suppose that in place of (3) we impose $\sigma_{\min} \leq \sigma(t) \leq \sigma_0$. In this case the optimal control is still bang-bang with a single switching time. In Fig. 6, we show an optimal solution when $\sigma(t) \geq 0.4\sigma_0$ is imposed.

The result above can also be obtained via the Hamilton-Jacobi-Bellman (HJB) equation for (15). Here we sketch this approach. The HJB equation for $u(x, y, t)$ can be written

Fig. 5 Optimal solutions starting from the same point (0.99, 0.01) but with different final times. A larger value of T allows the system to reach a more optimal state. For all solutions, $\beta = 0.3$ and $\gamma = 0.1$



(a) Solution and control vs. time



(b) Trajectory in phase space

Fig. 6 Optimal solutions with $\sigma(t) \geq 0.4\sigma_0$. Here $(x(0), y(0)) = (0.99, 0.01)$, $\beta = 0.3$, $\gamma = 0.1$, and $T = 100$

$$u_t = \gamma y u_y - \gamma x y \min_{\sigma} ((u_y - u_x) \sigma) \tag{24a}$$

$$u(x, y, T) = -x_{\infty}(x, y, \sigma_0). \tag{24b}$$

The required minimum is obtained by taking

$$\sigma(t) = \begin{cases} 0 & u_y(x, y, t) > u_x(x, y, t) \\ \sigma_0 & u_y(x, y, t) < u_x(x, y, t). \end{cases} \tag{25}$$

From (7) we see that $u_y(x, y, T) > u_x(x, y, T)$ for all (x, y) . Thus for small enough values of $T - t$, the solution of (24) satisfies

$$u_t = \gamma y u_y(x, y, t).$$

The solution of this hyperbolic PDE is

$$u(x, y, t) = u(x, ye^{-\gamma(T-t)}, T) = -x_{\infty}(x, ye^{-\gamma(T-t)}).$$

Thus, for small enough $T - t$,

$$\begin{aligned}u_x(x, y, t) &= -\frac{\partial x_\infty}{\partial y} \left(1 - \frac{1}{x(t)\sigma_0}\right) \\u_y(x, y, t) &= -\frac{\partial x_\infty}{\partial y} e^{-\gamma(T-t)}.\end{aligned}$$

According to (25), the optimal control value will switch when $u_x = u_y$, which leads to (22). Meanwhile, substituting (25) in (24) in the case $u_y < u_x$ yields the linear hyperbolic PDE

$$u_t = \gamma y u_y - \beta x y (u_y - u_x),$$

whose characteristics are just the trajectories of the SIR system (1) illustrated in Fig. 1, which are also contours of x_∞ . It can be shown that once $u_y - u_x < 0$, this inequality will continue to hold along each such characteristic.

4 Optimal control with $L \neq 0$

We now consider the case of a non-zero Lagrangian, which allows us to account for factors like the economic cost of intervention or heightened risks caused by hospital overflow. We formulate the Hamilton-Jacobi-Bellman (HJB) equation for this problem and apply an upwind numerical method to compute approximate solutions. The numerical solutions obtained via the HJB equation have also been checked in each case against solutions of the BVP given in Sect. 2.3, and found to agree within numerical errors.

Because the Lagrangian in this section is not a linear function, the solution is not bang-bang, and instead varies smoothly (except when it reaches the minimum or maximum allowable value).

4.1 Quadratic running cost of control

We now attempt to account for the economic cost of intervention. Quantification of the cost of measures like closing schools and businesses is a challenging problem in economic modeling, and well outside the scope of the present work. Based on the general idea that both the cost and the marginal cost will increase with the degree of contact reduction, we take for simplicity

$$L(x(t), y(t), \sigma(t)) = c_2 \left(1 - \frac{\sigma(t)}{\sigma_0}\right)^2.$$

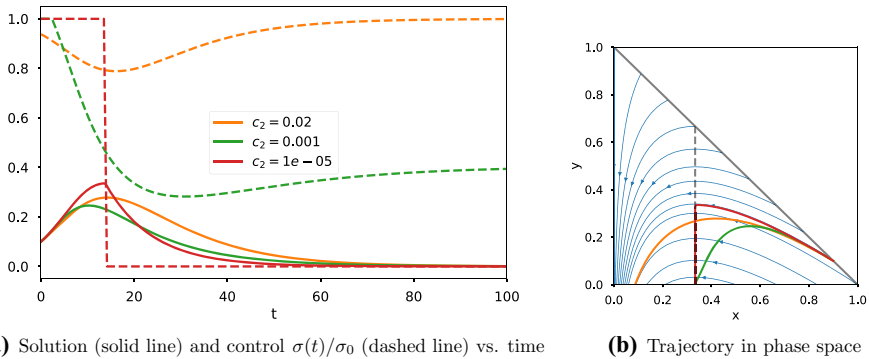


Fig. 7 Optimal solutions with different running cost. Here $(x(0), y(0)) = (0.9, 0.1)$, $\beta = 0.3$, $\gamma = 0.1$, and $T = 100$

The HJB equation for (6) is then

$$u_t - \gamma y u_y = - \min_{0 \leq \sigma \leq \sigma_0} \left((u_y - u_x) \gamma x y \sigma(t) + c_2 \left(1 - \frac{\sigma(t)}{\sigma_0} \right)^2 \right) \tag{26a}$$

$$u(x, y, T) = -x_\infty(x, y, \sigma_0). \tag{26b}$$

The minimum in (26a) is obtained with

$$\sigma(t) = \sigma_0 \min \left(1, \max \left(0, \left(1 - \frac{\sigma_0 \gamma}{2c_2} x y (u_y - u_x) \right) \right) \right). \tag{27}$$

We approximate the solution of (26)-(27) using a second-order finite volume discretization with the PyClaw software (Ketcheson et al. 2012, 2013; Mandli et al. 2016). For details, the reader is referred to the reproducibility repository that contains the code for all examples in this paper (Ketcheson 2021).

Numerical solutions for a range of values of c_2 are shown in Fig. 7. The values of c_2 used here are chosen merely to illustrate the range of possible behaviors. Notice that the strength of the control $\sigma(t)$ and the number of infected at certain times vary non-monotonically with c_2 . Indeed, the optimal control $\sigma(t)$ up to around day 15 is simply σ_0 in both limits $c_2 \rightarrow \infty$ and $c_2 \rightarrow 0$, whereas for intermediate values of c_2 some intervention is imposed in this period.

4.2 Minimizing hospital overflow

The optimal solutions above may be unsatisfactory in practice, since the number of people simultaneously infected at certain times may be too great for all of them to receive adequate medical care. This is a major concern with respect to the current COVID-19 crisis. A natural objective is to keep the number of infected below some threshold, corresponding for instance to the number of hospital beds. We thus consider the Lagrangian

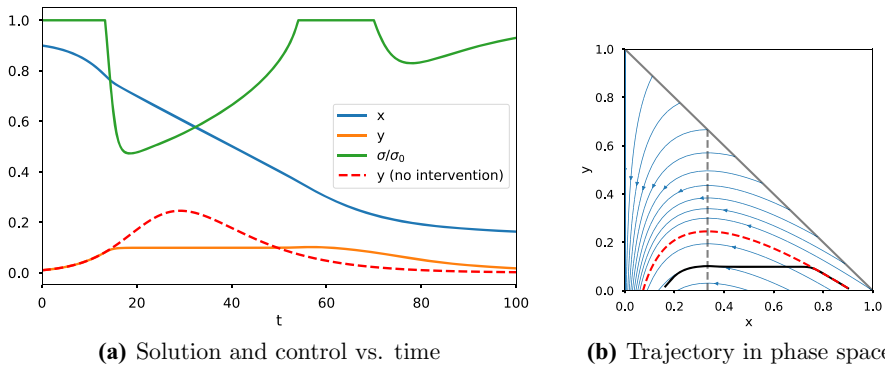


Fig. 8 Optimal solutions with cost for hospital overflow. Here $(x(0), y(0)) = (0.9, 0.01)$, $\beta = 0.3$, $\gamma = 0.1$, $T = 100$, and $y_{\max} = 0.1$. In the cost function, we take $c_2 = 10^{-2}$ and $c_3 = 100$. The dashed red line shows the result of imposing no control

$$L(x(t), y(t), \sigma(t)) = c_2 \left(1 - \frac{\sigma(t)}{\sigma_0} \right)^2 + c_3 g(y(t) - y_{\max}).$$

Here y_{\max} is the maximum number of hospital beds. The HJB equation is then

$$u_t - \gamma y u_y + c_3 g(y - y_{\max}) = - \min_{0 \leq \sigma \leq \sigma_0} \left((u_y - u_x) \gamma x y \sigma + c_2 \left(1 - \frac{\sigma}{\sigma_0} \right)^2 \right) \tag{28a}$$

$$u(x, y, T) = -x_{\infty}(x, y, \sigma_0). \tag{28b}$$

The control that achieves the minimum in (28a) is again given by (27). The function $g(v)$ should be nearly zero for $v < 0$ and increase in an approximately linear fashion for $v > 0$. For the purpose of having a tractable control problem, it is also desirable that g be differentiable. We take

$$g(v) = \frac{v}{1 + e^{-100v}}.$$

Figs. 8 and 9 show examples of solutions. Again, we choose parameter values that demonstrate the range of qualitative behaviors. In both examples, the cost of control is scaled by $c_2 = 10^{-2}$. In Fig. 8, a higher cost for hospital overflow is applied, with $c_3 = 100$. As might be expected, $y(t)$ is generally kept below y_{\max} (which is set to 0.1). The control is initially off, then turns on to avoid hospital overflow, and then turns off again. While the control is applied, it is maintained at a level that keeps the value of $y(t)$ nearly constant in time.

Figure 9 shows another example scenario in which the cost of hospital overflow is smaller, with $c_3 = 1$. In this case the hospital capacity is significantly exceeded for a short time, and the control is kept on until the final time, but the epidemiological overshoot is significantly reduced compared to the previous solution.

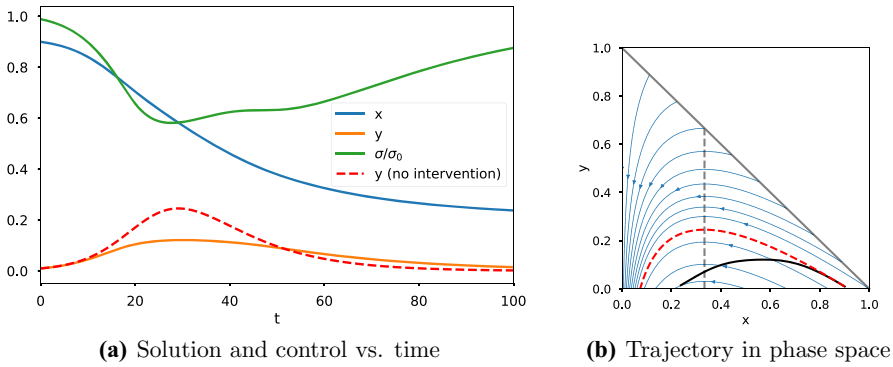


Fig. 9 Optimal solutions with smaller cost for hospital overflow. Here $(x(0), y(0)) = (0.9, 0.01)$, $\beta = 0.3$, $\gamma = 0.1$, $T = 100$, and $y_{\max} = 0.1$. In the cost function, we take $c_2 = 10^{-2}$ and $c_3 = 1$. The dashed red line shows the result of imposing no control

5 Application to the COVID-19 pandemic

The main goal of this work has been a mathematical investigation of optimal controls for the SIR model with a controlled rate of contact, as presented in the previous sections. We now present a brief illustration of the results in practical terms through application to the current COVID-19 pandemic. This application is imprecise, for several reasons: the SIR model is one of the simplest epidemiological models, and assumes homogeneous mixing among a population; the current state of susceptible and infected persons is not accurately known; and the parameters of the disease itself (i.e. γ, σ_0) are still quite uncertain. The examples in this section should be viewed only as illustrations of a few possible scenarios, and not an exhaustive or detailed study.

We take the infectious period $\gamma^{-1} = 10$ days, and the basic reproduction number $\sigma_0 = 3.2$, based on recent estimates (Verity et al. 2020; Ying et al. 2020). To make the results easy to interpret, we use a fixed terminal cost of $c_1 z_\infty$, where we have introduced an additional scaling constant. Taking $c_1 = \alpha N$, where N is the total population being modeled and α is the infection fatality ratio, then this cost is the expected number of lives lost. Since $z_\infty = 1 - x_\infty$, this is merely a rescaling of the terminal cost used throughout this work. We take $\alpha \approx 0.006$ based on recent estimates (Verity et al. 2020; Russell et al. 2020; Wu et al. 2020).

We seek reasonable order-of-magnitude estimates for c_2 and c_3 . The value of c_3/N should be equal to the increase in probability of a given infected person dying because of the lack of medical care. We take $c_3 = N\eta$, where the fatality ratio in the absence of medical care is $\alpha + \eta$. We take $\eta \approx \alpha$, giving $c_3 = 0.006$. For y_{\max} we take values from the United States, where there are about 3 hospital beds per 1000 people, and two-thirds of them are typically occupied. Since it is estimated that about 5% of COVID-19 cases are hospitalized (Verity et al. 2020), this gives $y_{\max} = 0.02N$.

Any attempt to quantify the cost of an intervention in human lives is bound to be contentious. Whether we consider the value of a human life to be in intrinsic personal value or extrinsic economic value, we can view the cost of intervention as a reduction of the value of human lives during the intervention period. We take $c_2 = N\epsilon/d$ where

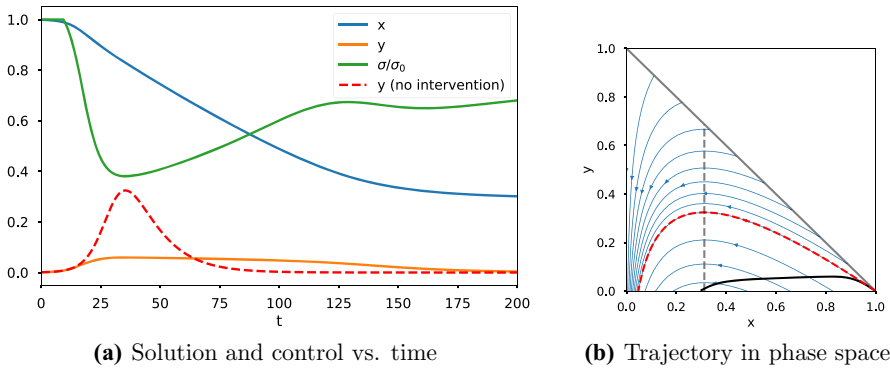


Fig. 10 Optimal control for COVID-19 with $\sigma_0 = 3.2$, $\gamma = 0.1$, $\alpha = \eta = 0.006$, $\epsilon = 0.2$, $d = 10^4$, $T = 200$, and $(x(0), y(0)) = (0.999, 0.001)$. The dashed red line shows the result of imposing no control

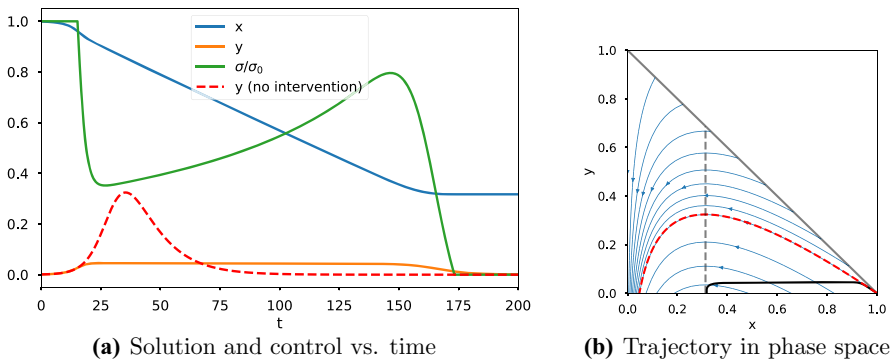


Fig. 11 Optimal control for COVID-19 with $\sigma_0 = 3.2$, $\gamma = 0.1$, $\alpha = \eta = 0.012$, $\epsilon = 0.05$, $d = 10^4$, $T = 200$, and $(x(0), y(0)) = (0.999, 0.001)$. The dashed red line shows the result of imposing no control

$d \approx 10^4$ is the number of days in a human life (more precisely, the average number of days remaining in a life claimed by the disease) and $1 - \epsilon$ is the relative value of a day spent in full isolation ($\sigma = 0$) compared to a day without intervention. Taking $\epsilon = 0.2$, we have $c_2 = 2 \times 10^{-5} N$.

Since all terms in the cost function are proportional to N , we take $N = 1$ without loss of generality. Results for the parameter values given above are shown in Fig. 10. We see that the optimal control corresponds to a level of intervention that becomes more strict as the epidemic grows, and is gradually relaxed as the epidemic subsides. Most importantly, and in agreement with results from the examples in earlier sections, the strongest control is applied around the time of peak infection and shortly thereafter. The resulting epidemiological overshoot is very small.

An alternative scenario is shown in Fig. 11, in which we have assumed a fatality ratio and a value of η that are twice as large (in line with the highest estimates of the infection fatality ratio), as well as taking a smaller cost of intervention with $\epsilon = 0.05$. These parameters lead to stronger intervention, especially in the later phases of the epidemic. The result is almost no epidemiological overshoot.

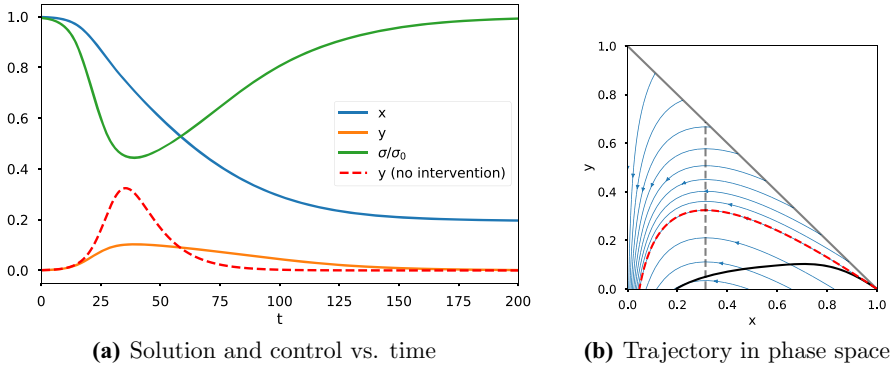


Fig. 12 Optimal control for COVID-19 with $\sigma_0 = 3.2$, $\gamma = 0.1$, $\alpha = \eta = 0.006$, $\epsilon = 0.5$, $d = 10^4$, $T = 200$, and $(x(0), y(0)) = (0.999, 0.001)$. The dashed red line shows the result of imposing no control

Finally, in Fig. 12, we repeat the first scenario but increase the cost of control by taking $\epsilon = 1$. In this case a more mild control is applied, peaking at about 35% contact reduction and concentrated around the time of the infection peak. In this case the optimal solution includes a small but significant epidemiological overshoot, and significantly exceeds the available hospital beds for a certain period of time.

6 Conclusion

We have studied, for an SIR model with a control on the rate of contact, the problem of minimizing the eventually infected population in the long-time limit, when the control can be applied only up to a finite time. In the absence of any cost of intervention, the optimal strategy is to apply no control until a certain switching time, and then apply maximum control. We have also considered other objective functions that include a running cost of control and a penalty for large numbers of simultaneous infections.

Contrary to simple intuition, it is not optimal to impose the maximum level of intervention from the earliest possible time. But real-world studies have supported this observation; a too-strong intervention may simply lead to a strong second wave of infection after the intervention is lifted, and not significantly reduce epidemiological overshoot (Bootsma and Ferguson 2007). On the other hand, intervention that starts too late or is lifted too soon may also have a negligible effect on total mortality (Bootsma and Ferguson 2007; Hatchett et al. 2007; Markel et al. 2007). The idea that intervention should possibly be delayed in order to increase its effect was also found in Ballard et al. (2017), although the objective and optimal policy found there differ from the present work.

The general results obtained here may provide insight into what optimal intervention strategies and their consequences may look like, but this should be informed by additional insight that can be gained from more detailed models. This work could form the basis of more detailed real-world application, using values of the disease parameters, costs, and effectiveness of NPIs relevant to a specific population of interest.

Although the SIR model is perhaps the simplest mathematical epidemiological model available, it has the advantage of requiring only a few parameters to be constrained. Results based on control of the SIR model could form a starting point for studying control in more complex models.

References

- Agusto FB (2013) Optimal isolation control strategies and cost-effectiveness analysis of a two-strain avian influenza model. *Biosystems* 113(3):155–164
- Ballard PG, Bean NG, Ross JV (2017) Intervention to maximise the probability of epidemic fade-out. *Math Biosci* 293:1–10
- Bootsma MCJ, Ferguson NM (2007) The effect of public health measures on the 1918 influenza pandemic in US cities. *Proc Nat Acad Sci* 104(18):7588–7593
- Clarke F (2013) *Functional analysis, calculus of variations and optimal control*, vol 264. Springer
- Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsrithaworn S, Burke DS (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437(7056):209–214
- Fister KR, Suzanne L, Scott MJ (1998) Optimizing chemotherapy in an HIV model. *Electron J Diff Eqn* 1998(32):1–12
- Greenhalgh D (1988) Some results on optimal control applied to epidemics. *Math Biosci* 88(2):125–158
- Harko T, Lobo FSN, Mak MK (2014) Exact analytical solutions of the susceptible-infected-recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Appl Math Comput* 236:184–194
- Hatchett RJ, Mecher CE, Lipsitch M (2007) Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proc Nat Acad Sci* 104(18):7582–7587
- Hethcote HW (2000) The mathematics of infectious diseases. *SIAM Rev* 42(4):599–653
- Jung E, Lenhart S, Feng Z (2002) Optimal control of treatments in a two-strain tuberculosis model. *Discret Contin Dyn Syst-B* 2(4):473
- Kar TK, Batabyal A (2011) Stability analysis and optimal control of an SIR epidemic model with vaccination. *Biosystems* 104(2–3):127–135
- Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. In: *Proceedings of the royal society of London. Series A, Containing papers of a mathematical and physical character* 115(772):700–721
- Ketcheson DI, Matteo Parsani, LeVeque RJ (2013) High-order wave propagation algorithms for hyperbolic systems. *SIAM J Sci Comput* 35(1):A351–A377
- Ketcheson DI (2021) SIR-control-code. Optimal control of an SIR epidemic through finite-time non-pharmaceutical intervention. <https://github.com/ketch/SIR-control-code>
- Ketcheson DI, Mandli KT, Ahmadi AJ, Alghamdi A, Quezada de Luna M, Parsani M, Knepley MG, Emmett M (2012) PyClaw: accessible, extensible, scalable tools for wave propagation problems. *SIAM J Sci Comput* 34(4):C210–C231
- Kirschner D, Lenhart S, Serbin S (1997) Optimal control of the chemotherapy of HIV. *Journal of Mathematical Biology* 35(7):775–792
- Lenhart S, Workman JT (2007) *Optimal control applied to biological models*. CRC press
- Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J (2020) The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med* 27(2):taaa021
- Mandli KT, Ahmadi AJ, Berger M, Calhoun D, George DL, Hadjimichael Y, Ketcheson DI, Lemoine GI, LeVeque RJ (2016) Clawpack: building an open source ecosystem for solving hyperbolic PDEs. *PeerJ Comput Sci* 2016(8):e68
- Markel H, Lipman HB, Navarro JA, Sloan A, Michalsen JR, Stern AM, Cetron MS (2007) Nonpharmaceutical interventions implemented by US cities during the 1918–1919 influenza pandemic. *JAMA* 298(6):644–654
- Pakes AG (2015) Lambert’s W meets Kermack–McKendrick epidemics. *IMA J Appl Math* 80(5):1368–1386
- Russell TW, Hellewell J, Jarvis CI, van Zandvoort K, Abbott S, Ratnayake R, Flasche S, Eggo RM, Kucharski AJ, CMMID nCov working group, et al (2020) Estimating the infection and case fatality ratio for

- COVID-19 using age-adjusted data from the outbreak on the Diamond Princess cruise ship. medRxiv. <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.12.2000256>
- Safi MA, Gumel AB (2013) Dynamics of a model with quarantine-adjusted incidence and quarantine of susceptible individuals. *J Math Anal Appl* 399(2):565–575
- Sharomi O, Malik T (2017) Optimal control in epidemiology. *Ann Oper Res* 251(1–2):55–71
- Sun H, Qiu Y, Yan H, Huang Y, Zhu Y, Chen SX (2020) Tracking and predicting COVID-19 epidemic in China Mainland. medRxiv. [https://doi.org/10.6339/JDS.202007_18\(3\).0010](https://doi.org/10.6339/JDS.202007_18(3).0010)
- Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, Cuomo-Dannenburg G, Thompson H, Walker PGT, Fu H, et al (2020) Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 20(6):669–677
- Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, Cowling BJ, Lipsitch M, Leung GM (2020) Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat Med* 26(4):506–510
- Yan X, Zou Y (2008) Optimal and sub-optimal quarantine and isolation control in SARS epidemics. *Math Comput Model* 47(1–2):235–245

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.