International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Address: IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria
Email: repository@iiasa.ac.at

Working paper

Optimal Transmission Rate in a Basic Three-Compartment Epidemic Model

Elena Rovenskaya (rovenska@iiasa.ac.at)
Sergey Orlov (orlov@iiasa.ac.at)

WP-21-002

Approved by:

Albert van Jaarsveld
Director General and Chief Executive Officer
4 February 2021

Table of contents

ZVR 524808900

Disclaimer:
The authors gratefully acknowledge funding from IIASA and the National Member Organizations that support the institute (The Austrian Academy of Sciences; The Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES); The National Natural Science Foundation of China (NSFC); The Academy of Scientific Research and Technology (ASRT), Egypt; The Finnish Committee for IIASA; The Association for the Advancement of IIASA, Germany; The Technology Information, Forecasting and Assessment Council (TIFAC), India; The Indonesian National Committee for IIASA; The Iran National Science Foundation (INSF); The Israel Committee for IIASA; The Japan Committee for IIASA; The National Research Foundation of Korea (NRF); The Mexican National Committee for IIASA; The Research Council of Norway (RCN); The Russian Academy of Sciences (RAS); Ministry of Education, Science, Research and Sport, Slovakia; The National Research Foundation (NRF), South Africa; The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS); The Ukrainian Academy of Sciences; The Research Councils of the UK; The National Academy of Sciences (NAS), USA; The Vietnam Academy of Science and Technology (VAST).

Abstract

In this paper, we consider a communicable disease whose dynamics is described by a simple three-compartment epidemic model. The epidemic spread begins with a small number of infected persons brought into a totally susceptible population. We assume that the central planner can introduce policy measures to reduce the transmission rate and hence the basic reproduction number over some time interval in order to minimize the total number of infected individuals over the entire time horizon of the model. Our model is arguably the simplest possible model that can describe an epidemic spread and its control. This enables analytical treatment of the model and the derivation of the closed-form solution. We make the following conclusions. We observe that the total number of infected individuals depends only on the lowered basic reproduction number and the increment in the number of recovered individuals over the period of the policy measures' application; it does not depend, for example, on the number of infected or recovered at the time of the policy inception. Since our model does not include policy costs, we obtain that in order to minimize the total number of infected individuals, the central planner should impose infinitely long policy measures. There is a set of optimal measures: Measures from this set are characterized by the tradeoff between the lowered basic reproduction number and the inception time of the measures. Since infinitely long application of measures is likely to be infeasible in the real world, we derive measures of finite duration which are near-to-optimal, that is measures which lead to the total number of infected individuals that is higher than the optimal number by a small positive ε. Our analysis of these near-to-optimal measures suggests that a robust way to ensure a reasonable duration of measures and at the same time to achieve that they indeed lead to an ε-optimal outcome is as follows. The central planner could choose the optimal policy inception time and the optimal lowered basic reproduction number, and then keep these measures for as long as it is feasible. The duration of this period will determine the number of infected individuals that will be achieved and the corresponding ε.

About the authors

Elena Rovenskaya is the Program Director of the Advancing Systems Analysis Program at IIASA. She is also a Research Scholar at the Optimal Control Department of the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia.

Sergey Orlov is a Research Scholar of the Exploratory Modeling Of Human-Natural Systems Research Group in the Advancing Systems Analysis Program at IIASA. He is also a Lecturer at the Optimal Control Department of the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia.

1. Introduction

The COVID-19 pandemic boosted modelers' interest to analyze compartmental models in epidemiology. Rooted in the Kermack–McKendrick theory that was developed in the late 1920s (Kermack et al., 1927), these models split the population into several compartments based on the status of individuals with regard to the contagious disease under study. The most commonly used compartment model in epidemiology is the SIR model, which operates with three compartments – susceptible (S), infected (I), and recovered (R). The main process represented in this model is the flow of individuals from the compartment of susceptible to the compartment of infected. This flow depends on the stocks of individuals in both compartments, as well as on the transmission rate.

The transmission rate is a key parameter that defines the dynamics of an epidemic. It is a product of the contact rate of individuals and the risk of infection. The risk of infection depends on the disease type as well as on the immunity of individuals, and thus can be reduced through, for example, vaccination. In the absence of the vaccine, as it was the case with the COVID-19 outbreak, policy measures to confine the disease should target the contact rate of individuals; these typically include isolation, social distancing, and quarantine. These measures save people's lives and reduce negative health effects – but they create a disutility to the entire population which ranges from a simple inconvenience to significant economic losses. Ideally, measures that reduce the contact rate (and hence, the transmission rate) should be optimally phased-in and phased-out through the disease dynamics to make the maximum effect on the public health.

In this paper, we consider a communicable disease whose dynamics is described by the SIR model over an infinite time horizon. We assume that the government can introduce policy measures to reduce the transmission rate over some time interval within the considered time horizon. Thus, the transmission rate is modeled as a piecewise constant function with two switching points. The beginning and the end of the time interval over which the policy measure is active, as well as the lowered transmission rate in this time interval are control parameters in the model. The objective function to be minimized is the total number of infected individuals over the entire time horizon of the model.

Several optimal control problems for compartmental models in epidemiology were proposed and investigated before COVID-19 (Grigorieva et al., 2016; Hansen and Day, 2011; Rowthorn et al., 2009). The ongoing pandemic created a fertile ground for such analyses and thus a number of publications have appeared very recently, which consider models of optimal control of an epidemic (Mandal et al., 2020; Perkins and España, 2020; Tsay et al., 2020). These models attempt to analyze tradeoffs between stricter epidemic control policy measures and the economic losses that such measures entail. In terms of epidemic control policies, different papers focus on different policy outcomes, such as the number of deaths due to COVID-19 (Perkins and España, 2020), or the peak number of infected (Tsay et al., 2020), or the total number of infected (Mandal et al., 2020). As such models are too complex to be solved analytically, these publications present some numerical solutions, which give only a limited information on the entire picture.

In our paper, we take a different approach. We focus exclusively on the health impact of the disease. First, we derive a policy measure that minimizes the number of infected individuals and analyze its potential to confine the epidemic. Naturally, in the absence of a tradeoff, such an optimal policy measure prescribes to maintain as-low-as-possible transmission rate infinitely long. While this is infeasible in reality, the optimal number of infected individuals in this case can serve as a benchmark to which other, near-to-optimal policies could be related. In particular, we analyze solutions, which deliver the total number of infected individuals that is higher than the benchmark value by some small ε (as fraction of the entire population). In this case, the policy application interval can be finite. We analyze the tradeoffs between the duration and the strictness

of the policy measure for such near-optimal solutions. Information on these tradeoffs can inform decisions where other objective functions than the number of infected individuals are part of the decision making.

2. Model

2.1. Epidemic dynamics in time

In this section, we briefly describe the dynamics of the basic SIR model (Kermack et al., 1927). We assume that the population stays constant over time and that all infected individuals recover. Hence, the compartment stocks can be normalized by the total population. Let $S(\tau)$, $I(\tau)$, and $R(\tau)$ stand for the normalized stocks of the susceptible, infected, and recovered individuals at time moment $\tau \geq 0$. The dynamics of the contagious disease propagation within the population is described by the following initial value problem (IVP):

$$\begin{aligned} \dot{S}(\tau) &= -\beta S(\tau) I(\tau), & S(0) &= S_0, \\ \dot{I}(\tau) &= \beta S(\tau) I(\tau) - \gamma I(\tau), & I(0) &= I_0, \\ \dot{R}(\tau) &= \gamma I(\tau), & R(0) &= R_0. \end{aligned} \tag{1}$$

Here, nonnegative initial values S_0, I_0, R_0 are such that $S_0 + I_0 + R_0 = 1$; parameter $\beta > 0$ is the transmission rate and parameter $\gamma > 0$ is the recovery rate, which is the inverse of the average infectious period. It is during the time period $1/\gamma$ from the moment of getting infected that the individual can transmit the infection to another (susceptible) individual.

We introduce a linear transformation of time that allows to reduce the dimensionality of the parameter space as follows: $t = \gamma\tau$. Now, considering phase variables S, I, and R as functions of $t \geq 0$, we obtain the following equations equivalent to (1):

$$\begin{aligned} \dot{S}(t) &= -\mu S(t) I(t), & S(0) &= S_0, \\ \dot{I}(t) &= \mu S(t) I(t) - I(t), & I(0) &= I_0, \\ \dot{R}(t) &= I(t), & R(0) &= R_0. \end{aligned} \tag{2}$$

Here $\mu = \beta/\gamma$ – which in literature is often denoted as $\mathcal{R}_0$ – is the basic reproduction number describing the average number of secondary cases produced by an average infectious individual in a totally susceptible population over the duration of time when s/he remains infectious. In (2), $\mathcal{R}_0 = \mu$, and we use μ instead of $\mathcal{R}_0$ as a notation for the basic reproduction number simply to make formulas look more elegantly.

"Time" t is a unitless variable; it can be understood as the number of individual infectious periods stacked one after another. Let us point that such a transformation of time in the SIR model that leads to (2) was used before in, for example, (Bertozzi et al., 2020) and (Khoshnaw et al., 2017).

We introduce the following assumptions.

Assumption A1. We consider an epidemiological process that begins with a positive number of infected individuals and no recovered individuals, that is $R_0 = 0$ and $I_0 = 1 - S_0 > 0$.

Assumption A2. $\mathcal{R}_e(0) = S_0\mathcal{R}_0 = S_0\mu > 1$, where $\mathcal{R}_e(0)$ is the initial effective reproduction number that is the average number of secondary cases produced by an average infectious individual in a population with S_0 susceptible individuals over the duration of time when s/he remains infectious.

Assumptions A1 and A2 together mean that system (1) spans the whole lifecycle of an epidemic of a new disease. Assumption A1 means that initially, the population is close to be totally susceptible with a small number of initially infected individuals. There are no recovered individuals yet. The infection begins to spread as the initially infected individuals contact with the susceptible ones and transmit the disease to them. In this way, the number of infected individuals begins to grow (Assumption A2 implies a positive gradient of $I(t)$ in a

right-hand side neighborhood of $t = 0$). Over time, the number of susceptible individuals decreases which leads to a decrease of the effective reproduction number, $\mathcal{R}_e(t) = S(t)\mu$. This decelerates growth of the number of infected and eventually growth turns into decline as more and more individuals recover. This turning point occurs when the number of susceptible $S(t)$ reaches $1/\mu$. Thus, in a well-mixed population, once the probability that a randomly selected individual is susceptible becomes lower than the threshold value $1/\mu$, the number of infected individuals begins to decline. This situation is often referred to as herd immunity (Hethcote, 2000).

Lemma 1 formally establishes some useful properties of the solution to the IVP (2).

Lemma 1. *Under Assumptions A1 and A2, solution* $\big(S(\cdot), I(\cdot), R(\cdot)\big)$ *to the IVP* (2) *over the time interval* $t \in [0, \infty)$ *has the following properties*:

1. $S(t) \geq 0, I(t) \geq 0, R(t) \geq 0$ *and* $S(t) + I(t) + R(t) = 1$ *for all* $t \in [0, \infty)$;
2. $S(\cdot)$ *is a strictly decreasing and* $R(\cdot)$ *is a strictly increasing functions of time* t *over the entire* $t \in [0, \infty)$;
3. *As* $t \to \infty$, $\big(S(\cdot), I(\cdot), R(\cdot)\big)$ *converges to a fixed point* $(S_\infty, I_\infty, R_\infty)$ *such that* $S_\infty = 1 - R_\infty$, $I_\infty = 0$, and $0 < R_\infty \leq 1$. *This means that* $R(t) \in [0, R_\infty)$ *for* $t \in [0, \infty)$.

Equilibrium in statement 3 of Lemma 1 is sometimes referred to as endemic equilibrium (Choisy et al., 2007). The proof of Lemma 1 is in the Appendix.

2.2. Epidemic dynamics in the phase space

According to Lemma 1, $R(t)$ is a monotonic, strictly increasing function of t that maps the time interval $[0, \infty)$ into interval $R \in [0, R_\infty)$. Thus, there exists an inverse of this mapping, $t(R)$ – a monotonic, strictly increasing function of R that maps $R \in [0, R_\infty)$ back to $t \in [0, \infty)$. Using this bijection, we will consider phase variables S and I as functions of $R \in [0, R_\infty)$. To eliminate the time dimension in (2), we divide each of the first two equations by the third one and obtain the following equations

$$\begin{aligned} &\frac{dS}{dR}(R) = -\mu S(R), \qquad S(R_0) = S_0, \\ &\frac{dI}{dR}(R) = \mu S(R) - 1, \quad I(R_0) = 1 - S_0 \end{aligned} \tag{3}$$

over $R \in [0, R_\infty)$. The advantage of this transition is that the IVP (3) has a closed-form solution as follows:

$$S(R) = S_0 e^{-\mu R}, \tag{4}$$

$$I(R) = 1 - R - S_0 e^{-\mu R}. \tag{5}$$

Lemma 2 establishes some useful properties of function $I(\cdot)$ (5).

Lemma 2. *Consider* R_∞ *from Lemma 1. Under assumptions A1 and A2,* R_∞ *is a unique root of equation* $I(R) = 0$ *over* $R \in [0, \infty)$ *where* $I(R)$ *is given by* (5). *Moreover,* $I(R) > 0$ *for* $R \in [0, R_\infty)$ *and* $I(R) < 0$ *for* $R > R_\infty$.

Solving the transcendental equation $I(R) = 0$ where $I(R)$ is given by (5), we obtain the exact formula for R_∞ is as follows:

$$R_\infty = 1 + \mu^{-1} W(-\mu S_0 e^{-\mu}) \tag{6}$$

where $W(\cdot)$ is the principal branch of the Lambert function (Corless et al., 1996). In what follows, unless indicated differently, we will use "Lambert function" to refer to its principal branch.

Building on Lemma 2, Lemma 3 presents the equivalence between the IVP (2) and the IVP (3) formally.

Lemma 3. *Let assumptions A1 and A2 hold.*

- *Let* $\left(S^{(2)}(t), I^{(2)}(t), R^{(2)}(t)\right)$ *be the solution to the IVP* (2) *over* $t \in [0, \infty)$.
- *Let* $\left(S^{(3)}(R), I^{(3)}(R)\right)$ *be the solution to the IVP* (3) *over* $R \in [0, R_\infty)$ *as defined by* (4) *and* (5)*, where* R_∞ *is a unique root of equation* $I^{(3)}(R) = 0$ *as given by* (6)*.*

Then

- $S^{(2)}(t) \equiv S^{(3)}\left(R^{(2)}(t)\right)$ *and* $I^{(2)}(t) \equiv I^{(3)}\left(R^{(2)}(t)\right)$ *over* $t \in [0, \infty)$.
- $S^{(3)}(R) \equiv S^{(2)}\left(t^{(2)}(R)\right)$ *and* $I^{(3)}(R) \equiv I^{(2)}\left(t^{(2)}(R)\right)$ *over* $R \in [0, R_\infty)$.
- $R^{(2)}(\cdot)$ *satisfies to the following IVP:*

$$\dot{R}(t) = 1 - R(t) - S_0 e^{-\mu R(t)}, \qquad R(0) = 0,$$

and $R^{(2)}(t) \to R_\infty$ *as* $t \to \infty$*.*

- $t^{(2)}(\cdot)$ *is inverse to* $R^{(2)}(\cdot)$ *and satisfies to the following IVP:*

$$\dot{t}(R) = (1 - R - S_0 e^{-\mu R})^{-1}, \qquad t(0) = 0,$$

and $t^{(2)}(R) \to \infty$ *as* $R \to R_\infty$*.*

2.3. Epidemic containment policy

We consider an outbreak of a contagious disease with $\mathcal{R}_0 = \mu_0 > 1$ and assume that the social planner aims to contain the outbreak by introducing an epidemic containment policy (ECP) over a certain period of time. We focus on ECPs which aim to decrease the transmission rate β – these include measures such as social distancing, quarantine etc. A decrease of the transmission rate leads to a proportional decrease of the basic reproduction number $\mu = \beta/\gamma$. Hence, in our model we assume that an ECP is applied over a time segment $[t_c, t_c + \Delta t_c]$ to reduce the basic reproduction number from the original μ_0 to some lower μ_c. Then the disease dynamics is described by the IVP (2) in which μ – instead of being constant – is of the following form

$$\mu = \mu(t) = \begin{cases} \mu_0 & \text{if } t \in [0, t_c), \\ \mu_c \in [0, \mu_0] & \text{if } t \in [t_c, t_c + \Delta t_c), \\ \mu_0 & \text{if } t \in [t_c + \Delta t_c, \infty). \end{cases} \tag{7}$$

The social planner can choose three control parameters of the ECP: the inception time, $0 \le t_c < \infty$, the effectiveness of the control measures, i.e., the lowered basic reproduction number, $0 \le \mu_c \le \mu_0$, and the duration, $0 \le \Delta t_c < \infty$. Denote the set of all admissible ECPs as

$Z = \{(t_c, \mu_c, \Delta t_c) : 0 \le t_c < \infty, 0 \le \mu_c \le \mu_0, 0 \le \Delta t_c < \infty\} \subset R^3$.

The choice of the ECP control parameters defines the evolution of the phase variables $\left(S(\cdot), I(\cdot), R(\cdot)\right)$ over time t. The piecewise constant $\mu(\cdot)$ (7) allows splitting the corresponding IVP (2) into a series of three IVPs each with a constant μ. Due to this fact, the solution to this IVP has following properties.

Lemma 4. *For any triple* $(t_c, \mu_c, \Delta t_c) \in Z$*, under assumptions A1 and A2, the corresponding solution* $\left(S(\cdot), I(\cdot), R(\cdot)\right)$ *to the IVP* (2) *over the time interval* $t \in [0, \infty)$ *has properties 1, 2, and 3 from Lemma 1.*

In the same way as above, we eliminate time and proceed from the IVP (2) with the piecewise constant μ (7) to the IVP (3) in which μ is of the following form

$$\mu = \mu(R) = \begin{cases} \mu_0 & \text{if } R \in [0, R_c), \\ \mu_c \in [0, \mu_0] & \text{if } R \in [R_c, R_c + \Delta R_c), \\ \mu_0 & \text{if } R \in [R_c + \Delta R_c, R_\infty). \end{cases} \tag{8}$$

By this construction, $R_c = R(t_c)$ and $R_c + \Delta R_c = R(t_c + \Delta t_c)$.

Thus, the inception and the duration of the ECP are now marked not in terms of time t but in terms of the corresponding levels of the recovered individuals. Hence, an ECP in model (3), (8) consists of the level of the

recovered individuals which should be reached to kick off an ECP, $0 \le R_c < R_\infty$, the lowered basic reproduction number, $0 \le \mu_c \le \mu_0$, as well as the increment in the number of recovered that should be achieved to end the application of the measures, $0 \le \Delta R_c < R_\infty - R_c$. Denote the set of all *admissible* ECPs as

$$\Omega = \{(R_c, \mu_c, \Delta R_c): 0 \le R_c < R_\infty, 0 \le \mu_c \le \mu_0, 0 \le \Delta R_c < R_\infty - R_c\} \subset R^3.$$

Using (4) and (5), we obtain the closed-form solution to the IVP (3), (8) as follows:

$$S(R) = \begin{cases} S_0 e^{-\mu_0 R} & \text{if } R \in [0, R_c), \\ S_0 e^{-\mu_0 R_c} e^{-\mu_c (R - R_c)} & \text{if } R \in [R_c, R_c + \Delta R_c), \\ S_0 e^{-\mu_c \Delta R_c} e^{-\mu_0 (R - \Delta R_c)} & \text{if } R \in [R_c + \Delta R_c, R_\infty). \end{cases} \quad (9)$$

$$I(R) = 1 - R - S(R). \quad (10)$$

Analogously to Lemma 2, Lemma 5 establishes some useful properties of function $I(\cdot)$ (10) in this case.

Lemma 5. *Let* $(R_c, \mu_c, \Delta R_c) \in \Omega$. *Consider* R_∞ *from Lemma 4 and consider a continuation of* $S(R)$ *and* $I(R)$ *over the entire* $R \in [0, \infty)$ *where in* $[R_\infty, \infty)$ $S(R)$ *takes values according to the third line in* (9) *and* $I(R)$ *is defined by* (10)*. Under assumptions A1 and A2 and provided that*

$$I(R_c) > 0 \text{ and } I(R_c + \Delta R_c) > 0, \quad (11)$$

R_∞ *is a unique root of equation* $I(R) = 0$*. Moreover,* $I(R) > 0$ *for* $R \in [0, R_\infty)$ *and* $I(R) < 0$ *for* $R \in (R_\infty, \infty)$.

Solving the transcendental equation $I(R) = 0$ where $I(R)$ is given by (10), we obtain the exact formula for R_∞ as follows:

$$R_\infty = 1 + \mu_0^{-1} W\left(-S_0 \mu_0 e^{-\mu_0} e^{(\mu_0 - \mu_c)\Delta R_c}\right). \quad (12)$$

Interestingly, R_∞ does not depend on R_c, only on ΔR_c and μ_c.

Note that for any ECP $(R_c, \mu_c, \Delta R_c) \in \Omega$, inequality $R_\infty \le R_\infty^0$ holds, where

$$R_\infty^0 = 1 + \mu_0^{-1} W(-S_0 \mu_0 e^{-\mu_0}) \quad (13)$$

corresponds to a "degenerated" ECP, that is a policy that does not lead to any change in the disease dynamics; obviously a degenerated ECP has either $\mu_c = \mu_0$ or if $R_c = R_\infty$ or if $\Delta R_c = 0$. So, any non-degenerated ECP reduces R_∞ compared to R_∞^0.

Denote

$$\widehat{\Omega} = \{(R_c, \mu_c, \Delta R_c) \in \Omega: I(R_c) > 0, I(R_c + \Delta R_c) > 0\} \subset R^3.$$

Using (9) and (10) to obtain explicit formulas for conditions (11), we specify set $\widehat{\Omega}$ in terms of control parameters R_c, μ_c, and ΔR_c as follows:

$$0 \le R_c < 1 + \mu_0^{-1} W(-S_0 \mu_0 e^{-\mu_0}), \quad (14)$$

$$0 \le \mu_c \le \mu_0, \quad (15)$$

$$0 \le \Delta R_c < 1 + \mu_c^{-1} W\left(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c) R_c}\right) - R_c. \quad (16)$$

Note that for $\mu_c = 0,$ inequality (16) should be understood as the limit case as follows:

$$\Delta R_c < 1 + \lim_{\mu_c \to +0} \mu_c^{-1} W\left(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c) R_c}\right) - R_c = 1 - S_0 e^{-\mu_0 R_c} - R_c.$$

Lemma 6 establishes the equivalence between the IVP (2) with $\mu = \mu(t)$ as defined by (7) and the IVP (3) with $\mu = \mu(R)$ as defined by (8).

Lemma 6. *Let* $(t_c, \mu_c, \Delta t_c) \in Z$ *and* $(R_c, \mu_c, \Delta R_c) \in \widehat{\Omega}$. *Let assumptions A1 and A2 hold. Moreover,*

- *Let* $\left(S^{(2)}(t), I^{(2)}(t), R^{(2)}(t)\right)$ *be the solution to the IVP* (2), (7) *over* $t \in [0, \infty)$.
- *Let* $\left(S^{(3)}(R), I^{(3)}(R)\right)$ *be the solution to the IVP* (3), (8) *over* $R \in [0, R_\infty)$ *as defined by* (9) *and* (10)*, where* R_∞ *is a unique root of equation* $I^{(3)}(R) = 0$ *as given by* (12).

Then

- $S^{(2)}(t) \equiv S^{(3)}\left(R^{(2)}(t)\right)$ *and* $I^{(2)}(t) \equiv I^{(3)}\left(R^{(2)}(t)\right)$ *over* $t \in [0, \infty)$*, where* $t_c = t^{(2)}(R_c)$ *and* $\Delta t_c = t^{(2)}(R_c + \Delta R_c) - t^{(2)}(R_c)$.

- $S^{(3)}(R) \equiv S^{(2)}\left(t^{(2)}(R)\right)$ *and* $I^{(3)}(R) \equiv I^{(2)}\left(t^{(2)}(R)\right)$ *over* $R \in [0, R_\infty)$*, where* $R_c = R^{(2)}(t_c)$ *and* $\Delta R_c = R^{(2)}(t_c + \Delta t_c) - R^{(2)}(t_c)$*.*
- $R^{(2)}(\cdot)$ *satisfies to the following IVP:*

$$\dot{R}(t) = 1 - R(t) - S_0 e^{-\mu(t)R(t)}, \qquad R(0) = 0,$$

and $R^{(2)}(t) \to R_\infty$ *as* $t \to \infty$*.*

- $t^{(2)}(\cdot)$ *is inverse to* $R^{(2)}(\cdot)$ *and satisfies to the following IVP:*

$$\dot{t}(R) = \left(1 - R - S_0 e^{-\mu(R)R}\right)^{-1}, \qquad t(0) = 0, \tag{17}$$

and $t^{(2)}(R) \to \infty$ *as* $R \to R_\infty$*.*

2.4. Optimization problem

We assume that the aim of an epidemic containment policy is to minimize the total number of infected individuals within the entire time interval $[0, \infty)$. Let us compute this number. The outbreak dynamics begins with I_0 infected individuals, and at each time moment $t \in (0, \infty)$, there are $\mu(t)S(t)I(t)$ newly infected individuals. Eventually, all infected individuals recover and, hence, the total number of infected individuals (TNII) is

$$I_0 + \int_0^\infty \mu(t)S(t)I(t)dt = I_0 + \int_0^\infty \left[\dot{I}(t) + \dot{R}(t)\right]dt = R_\infty.$$

Thus, we consider the following optimization problem:

$$\begin{array}{l} R_\infty \to \inf \\ \text{subject to:} \\ \dot{S}(t) = -\mu(t)S(t)I(t), \qquad S(0) = S_0, \\ \dot{I}(t) = \mu(t)S(t)I(t) - I(t), \quad I(0) = 1 - S_0, \\ \dot{R}(t) = I(t), \qquad R(0) = 0. \\ (t_c, \mu_c, \Delta t_c) \in Z. \end{array} \tag{18}$$

Relation (12) provides an explicit formula for the objective function R_∞ which depends only on two parameters, μ_c and ΔR_c, and does not depend on R_c. As the Lambert function $W(\cdot)$ is strictly increasing, the search for an infimum of (12) with regard to R_c, μ_c, and ΔR_c is equivalent to the search for a supremum of the product $(\mu_0 - \mu_c)\Delta R_c$. Hence, based on Lemma 6, the following optimization problem is equivalent to problem (18):

$$\begin{array}{l} (\mu_0 - \mu_c)\Delta R_c \to \sup \\ \text{subject to:} \\ 0 \le R_c < 1 + \mu_0^{-1} W(-S_0\mu_0 e^{-\mu_0}), \\ 0 \le \mu_c \le \mu_0, \\ 0 \le \Delta R_c < 1 + \mu_c^{-1} W\left(-S_0\mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c}\right) - R_c. \end{array} \tag{19}$$

(Recall the comment to inequality (16) concerning the treatment of case $\mu_c = 0$).

3. Results

In this section, we discuss problem (19) under assumptions A1 and A2. In the figures that provide illustration to our statements, we use $\mu_0 = 1.95$, which corresponds to the average estimate of the basic reproduction number for China provided by WHO on 18 January 2020 (Liu et al., 2020). We use $S_0 = 1 - 10^{-7}$, which means that the initial number of infected individuals constitutes 0.00001% of the population. Officially, a very

close level to this value was achieved in China on 19 January 2020[1]. Qualitatively, the results presented in the figures do not depend on the choice of these parameters.

3.1. Epidemic containment policy that minimizes the total number of infected individuals

Denote $\bar{R} = \mu_0^{-1} \ln(S_0\mu_0)$ and $\bar{\mu} = (1-\mu_0^{-1})^{-1} \ln(S_0\mu_0)$.

Proposition 1. *The following one-parametric family accumulates optimal solutions to the TNII-optimization problem* (19) *under assumptions A1 and A2:*

$$\begin{aligned} & R_c^* \in [0, \bar{R}], \\ & \mu_c^* = \frac{\ln(S_0\mu_0) - \mu_0 R_c^*}{1 - \mu_0^{-1} - R_c^*}, \\ & \Delta R_c^* = 1 - \mu_0^{-1} - R_c^*, \end{aligned} \tag{20}$$

The optimal value of the objective function is then

$R_\infty^* = 1 - \mu_0^{-1} = R_c^* + \Delta R_c^*$.

Note that the optimal triples $(R_c^*, \mu_c^*, \Delta R_c^*)$ in (20) lie on the boundary of the admissible set of problem (19). The optimal solution family could likewise be parametrized by $\mu_c^* \in [0, \bar{\mu}\,]$ or by $\Delta R_c^* \in [1 - \mu_0^{-1} - \bar{R}, 1 - \mu_0^{-1}]$. The proof of Proposition 1 can be found in the Appendix.

Proposition 1 states that an ECP that minimizes the total number of infected individuals should continue until all infected people have recovered, thus the number of recovered reaches $R_\infty^* = 1 - \mu_0^{-1}$. It means that the optimal terminal number of individuals who remain susceptive is $S_\infty^* = \mu_0^{-1}$, which equals to the threshold value for herd immunity. The optimal TNII is higher for an epidemic with a higher basic reproduction number μ_0.

We illustrate the TNII-optimal solution set from Proposition 1 in Figure 1. The thick black curve depicts optimal pairs of μ_c^* and R_c^* (20). Corresponding optimal ΔR_c^* is the distance between the point on thick black curve and the vertical line $R_c = R_\infty^*$. In order to minimize the total number of infected individuals – equal to R_∞ – the epidemic containment policy should kick off any time from the beginning of the disease outbreak and before the number of recovered individuals R_c^* reaches the critical threshold value $\bar{R}$. The optimal increment of the number of recovered individuals ΔR_c^* always complements R_c^* such that $R_c^* + \Delta R_c^* = R_\infty^*$. The choice of the optimal $R_c^* \in [0, \bar{R}]$ dictates the optimal level to which the basic reproduction number needs to be lowered – $\mu_c^* \in [0, \bar{\mu}\,]$ (20). There is a tradeoff between the strictness of an ECPs and the optimal start of the policy implementation period: A later start of an optimal ECP application (i.e., a higher R_c^*) requires a larger reduction in the basic reproduction number (i.e., a lower μ_c^*).

[1] 125 confirmed cases (WHO COVID-19 Table, 2020) with respect to the population of 1,408,526,449 people as of 1 January 2020 (China population - Countrymeters, 2020).

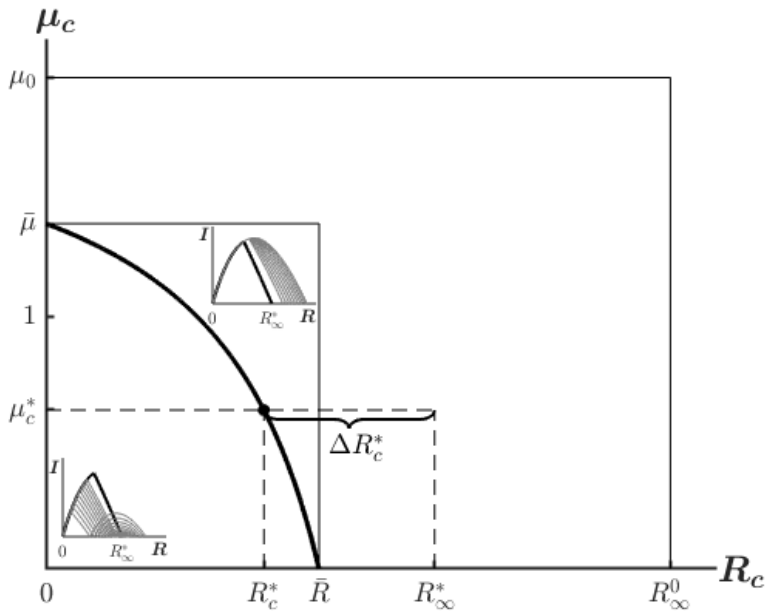

Figure 1. The space of control parameters μ_c and R_c. All optimal μ_c^* and R_c^* from the one-parametric family (20) lie on the thick black curve. Corresponding optimal ΔR_c^* is the distance between a point on the thick black curve and the vertical line $R_c = R_\infty^*$. Two small plots located above and below the thick black curve in Figure 1 depict epidemic curves for different values of control parameters; in both plots, the black curve depicts the TNII-optimal dynamics; grey curves depict illustrative dynamics for non-optimal values of μ_c and R_c – according to the placement of these plots on the Figure below or above the thick black curve.

If the central planner has not managed to start the containment measures before the number of recovered individuals exceeds threshold $\bar{R}$, s/he can still do so, however, reducing μ_c even to zero would not help achieve the optimal R_∞^* and the resultant R_∞ will be higher (area to the right from the line $R = \bar{R}$ in Figure 1). Similarly, there exists a critical threshold for the lowered basic reproduction number $\bar{\mu}$. If the central planner cannot ensure that the basic reproduction number is reduced at least to this level, starting an ECP even right at the beginning of the disease outbreak would lead to R_∞ that is higher than R_∞^* (area above the line $\mu = \bar{\mu}$ in Figure 1). A higher basic reproduction number μ_0 implies tighter conditions on R_c^* and μ_c^* that is for a higher μ_0, both $\bar{\mu}$ and $\bar{R}$ are lower.

According to Proposition 1, optimal ECPs should last until the number of recovered individuals reaches R_∞^*, that is the optimal dynamics $(S^*(R), I^*(R))$ consists of two (not three) segments (see (9), (10)). In the first segment, when $R \in [0, R_c^*)$, that is before the ECP inception, the number of infected individuals $I^*(R)$ is growing. In the second segment, where an ECP is applied, $I^*(R)$ is decreasing until it reaches zero at $R = R_\infty^*$. The thick black curves in small plots in Figure 1 depict corresponding epidemic curves.

If the basic reproduction number during an ECP is reduced more than the optimal level, that is if μ_c is lower than μ_c^*, and/or if an ECP starts at some R_c that is earlier than optimal R_c^*, there will be a second wave of the epidemic. A set of grey curves in the small plot in area below the black thick curve in Figure 1 illustrates exemplary epidemic curves corresponding to different pairs of R_c and μ_c from this area, while ΔR_c takes the highest hypothetically possible value (see (16)). As a too strict and/or too early application of an ECP leaves too many susceptible individuals in the population, each such trajectory contains a second wave, which can be of different magnitude. These dynamics result in a higher total number of infected individuals R_∞ than optimal R_∞^*.

If on the contrary, an ECP is less strict and/or kicks off later than prescribed by the optimal relation, the number of infected individuals will grow to higher levels than in the optimal scenario. A set of grey curves in the small plot in area above the black thick curve in Figure 1 illustrates exemplary trajectories corresponding to different pairs of R_c and μ_c from this area, while ΔR_c takes the highest hypothetically possible value (see (16)). In both cases $I(R)$ overshoots R_∞^* and leads to higher values of R_∞.

In terms of time t, an optimal ECP should start at $t_c^* = t(R_c^*)$ where $t(\cdot)$ is the solution to the IVP (17) (see Lemma 6). According to Proposition 1, an optimal ECP must continue until the number of recovered individuals reaches R_∞^*, hence, the duration of such an ECP in time t is infinite: $\Delta t_c^* = t(R_\infty^*) - t(R_c^*) = \infty$. This contradicts our formal assumption that $\Delta t_c^* < \infty$. Indeed, in reality, if a reduction of the basic reproduction number is intended to be achieved through a reduction of the transmission rate, an infinitely long ECP is not viable: To decrease the transmission rate, the government implements various measures restricting contacts between people from pausing activities of different economic sectors (education, gastronomy, entertainment, etc.) to constraining mobility of people (e.g., a ban to leave homes except for certain reasons etc.). All such measures compromise people's well-being and can lead to massive economic losses, so naturally, the central planner would like them to continue for as short time as possible, and certainly not forever. The central planner therefore may wish to implement near-to-optimal measures, so as to achieve the total number of infected individuals close to the optimal one, but by means of a finite-long ECP. In the next section, we explore such near-to-optimal controls in problem (19).

3.2. Near-to-TNII-optimal epidemic control policies

Consider a sufficiently small $\varepsilon > 0$ and consider a target value $R_\infty^\varepsilon = R_\infty^* + \varepsilon$. We aim to derive an epidemic control policy that allows to reach the ε-optimal TNII value R_∞^ε within a finite time. We are interested in near-to-optimal policies, so the deviation from the R_∞^* should indeed be small. By all means, the ε-optimal level of the infected individuals cannot exceed the number of recovered individuals in case of a degenerated ECP, equal to R_∞^0 (13). Note that $R_\infty^\varepsilon \le R_\infty^0$ as long as
$\varepsilon \le \mu_0^{-1}\big(1 + W(-S_0\mu_0 e^{-\mu_0})\big) = \varepsilon^0$.

In what follows, we will derive ECP triples $(R_c, \mu_c, \Delta R_c) \in \widehat{\Omega}$, which ensure that $R(t)$ reaches R_∞^ε as $t \to \infty$; we denote the control parameters of such ε-optimal ECPs as $R_c^\varepsilon, \mu_c^\varepsilon$, and ΔR_c^ε. Further, let us denote $S_\infty^\varepsilon = 1 - R_\infty^\varepsilon = \mu_0^{-1} - \varepsilon$, $\sigma_\infty^\varepsilon = S_0/S_\infty^\varepsilon$, and $\mu_\infty^\varepsilon = S_\infty^\varepsilon \mu_0$.

Proposition 2. *Under assumptions A1 and A2, for all* $\varepsilon \in (0, \varepsilon^0]$, *an* ε*-optimal triple* $(R_c^\varepsilon, \mu_c^\varepsilon, \Delta R_c^\varepsilon) \in \widehat{\Omega}^\varepsilon \subset \widehat{\Omega}$, *where*

$$\widehat{\Omega}^\varepsilon = \left\{(R_c^\varepsilon, \mu_c^\varepsilon, \Delta R_c^\varepsilon) : R_c^\varepsilon \in (\underline{R}_c^\varepsilon, \bar{R}_c^\varepsilon), R_c^\varepsilon \ge 0, \mu_c^\varepsilon \in [0, \bar{\mu}_c^\varepsilon), \Delta R_c^\varepsilon = \frac{R_\infty^\varepsilon \mu_0 - \ln \sigma_\infty^\varepsilon}{\mu_0 - \mu_c^\varepsilon}\right\},$$

$$\underline{R}_c^\varepsilon = 1 - \Delta R_c^\varepsilon + \mu_0^{-1} W_{-1}\big(-\mu_\infty^\varepsilon e^{-\mu_\infty^\varepsilon}\big),$$

$$\bar{R}_c^\varepsilon = \frac{\ln \sigma_\infty^\varepsilon - R_\infty^\varepsilon \mu_c^\varepsilon}{\mu_0 - \mu_c^\varepsilon},$$

$$\bar{\mu}_c^\varepsilon = (R_\infty^\varepsilon)^{-1} \ln \sigma_\infty^\varepsilon.$$

Here $W_{-1}(\cdot)$ *is the second branch of Lambert function different from the principal branch* $W(\cdot)$; *it is defined in* $\left[-\frac{1}{e}, 0\right)$ *and takes values in* $(-\infty, -1]$.

The proof of Proposition 2 can be found in the Appendix.

Proposition 2 establishes the fact that for any $\varepsilon \in (0, \varepsilon^0]$, one can choose values R_c^ε, μ_c^ε, and ΔR_c^ε from $\widehat{\Omega}^\varepsilon$ and thus achieve $R_\infty^\varepsilon = R_\infty^* + \varepsilon$. Note that if the central planner chooses a $R_c \ge 0$ such that $R_c \in (\underline{R}_c^\varepsilon, \bar{R}_c^\varepsilon)$ and a μ_c

such that $\mu_c \in [0, \bar{\mu}_c^{\varepsilon})$, but – instead of $\Delta R_c^{\varepsilon} = \frac{R_\infty^{\varepsilon}\mu_0 - \ln \sigma_\infty^{\varepsilon}}{\mu_0 - \mu_c^{\varepsilon}}$ – s/he takes some admissible ΔR_c (i.e., ΔR_c satisfying (16) where $R_c = R_c^{\varepsilon}$ and $\mu_c = \mu_c^{\varepsilon}$) such that $\Delta R_c \geq \Delta R_c^{\varepsilon}$, the corresponding total number of the infected individuals will be $R_\infty \leq R_\infty^* + \varepsilon$.

For illustrative purposes, we choose $\varepsilon = 0.1$. Panel A in Figure 2 depicts the set of ε-optimal ECPs $\widehat{\Omega}^{\varepsilon}$. Any triple $(R_c^{\varepsilon}, \mu_c^{\varepsilon}, \Delta R_c^{\varepsilon})$ from the presented surface delivers the ε-optimal TNII value $R_\infty^{\varepsilon} = R_\infty^* + \varepsilon$. As ΔR_c^{ε} does not depend on R_c^{ε}, for a constant μ_c^{ε}, ΔR_c^{ε} is constant. ΔR_c^{ε} gets higher for higher μ_c^{ε} with an increasing return to scale. Thus, a less strict ECP requires reaching a disproportionally larger increment of the fraction of recovered individuals in the population to exit the ECP. If the central planner chooses an ECP such that the corresponding point $(R_c, \mu_c, \Delta R_c)$ is admissible (satisfies (16)) and lies above the ε-optimal surface, the resultant TNII value will be lower than R_∞^{ε}.

Recall that the need to introduce near-to-optimal ECPs arose from the fact that the ECP that minimized the total number of infected individuals requires an infinitely long application of the optimal policy, which may be infeasible. Panels B and C in Figure 3 present trough-like shapes which illustrate the ECP duration Δt_c^{ε} in terms of time t (measured as the number of infectious periods) in an ε-optimal policy as a function of μ_c^{ε} and R_c^{ε} and as a function of μ_c^{ε} and t_c^{ε} correspondingly. Any triple $(R_c^{\varepsilon}, \mu_c^{\varepsilon}, \Delta t_c^{\varepsilon})$ and triple $(t_c^{\varepsilon}, \mu_c^{\varepsilon}, \Delta t_c^{\varepsilon})$ from the presented surfaces deliver the ε-optimal TNII value $R_\infty^{\varepsilon} = R_\infty^* + \varepsilon$. In these plots, ΔR_c^{ε} and R_c^{ε} are translated into Δt_c^{ε} and t_c^{ε} according to Lemma 6 where $\mu_c = \mu_c^{\varepsilon}$, $t_c = t_c^{\varepsilon}$, $R_c = R_c^{\varepsilon}$, $\Delta R_c = \Delta R_c^{\varepsilon}$, $\Delta t_c = \Delta t_c^{\varepsilon}$. If the central planner chooses an ECP such that the corresponding points $(R_c, \mu_c, \Delta t_c)$ and $(t_c, \mu_c, \Delta t_c)$ are admissible and lie above the corresponding ε-optimal surfaces in Panels B and C, the resultant TNII value will be lower than R_∞^{ε}.

As can be seen from comparison of Panel A with Panels B and C, the translation of ΔR_c^{ε} into Δt_c^{ε} on the vertical axis is non-trivial. First, it features a new dependence as Δt_c^{ε} depends not only on μ_c^{ε}, but also on R_c^{ε} (and hence, on t_c^{ε}). Let us first consider the dependence of Δt_c^{ε} on μ_c^{ε} and R_c^{ε} depicted in Panel B. This dependence highly non-linear: Transverse sections of the trough-shaped surface in Panel B have a U shape with some relatively flat area around the minimum point. As μ_c^{ε} and R_c^{ε} move away this minimum point and approach the boundaries of the trough-shaped surface, Δt_c^{ε} increases steeply and tends to the infinity. The dependence of Δt_c^{ε} on μ_c^{ε} and t_c^{ε} depicted in Panel C has qualitatively similar.

The shortest ECP duration Δt_c^{ε} – in this case, equal to 2.6 infectious periods – is achieved at the edge of the trough corresponding to $\mu_c^{\varepsilon} = 0$ (see Panel B). It requires $R_c^{\varepsilon} = 0.34$, which means that the ECP should be kicked off once about 1/3 of the population have already had the disease. In terms of time, it means that this policy should start in 16.7 infectious periods from the beginning of the disease outbreak (see Panel C). In this case, the eventual ε-optimal fraction of recovered is $R_\infty^{\varepsilon} = 0.59$ which is less than 2/3, while minimum possible fraction of recovered that would require an infinitely long ECP is $R_\infty^* = 0.49$. Generally, if the central planner wants to minimize the duration of an ECP, s/he should wait until a sufficiently high number of individuals get the disease and recover from it, and then introduce strictest possible measures.

Note that for any $\varepsilon \in (0, \varepsilon^0]$, line

$$\begin{aligned} & R_c \in [0, \bar{R}], \\ & \mu_c = \frac{\ln(S_0\mu_0) - \mu_0 R_c}{1 - \mu_0^{-1} - R_c}, \\ & \Delta R_c = \frac{\mu_0 R_\infty^{\varepsilon} - \ln \sigma_\infty^{\varepsilon}}{\mu_0 - \mu_c} \end{aligned} \tag{21}$$

fully belongs to $\widehat{\Omega}^{\varepsilon}$ (see details in the Appendix). The projection of this line onto the $R\mu$-plane coincides with the projection of the optimal solution $(R_c^*, \mu_c^*, \Delta R_c^*)$ (see Proposition 1 and formula (20)) onto the same plane. This means that one feasible policy for the central planner is to initiate an ECP with optimal μ_c^* and R_c^*, however, with a duration $\Delta R_c < \Delta R_c^*$ that is finite in time. By regulating this duration, the central planner

would be able to adjust R_∞ and, if so desired, to bring it to a certain $R_\infty^\varepsilon = R_\infty^* + \varepsilon$. The relatively flat areas that we discussed above are located around this curve and this supports robustness of this policy recommendation.

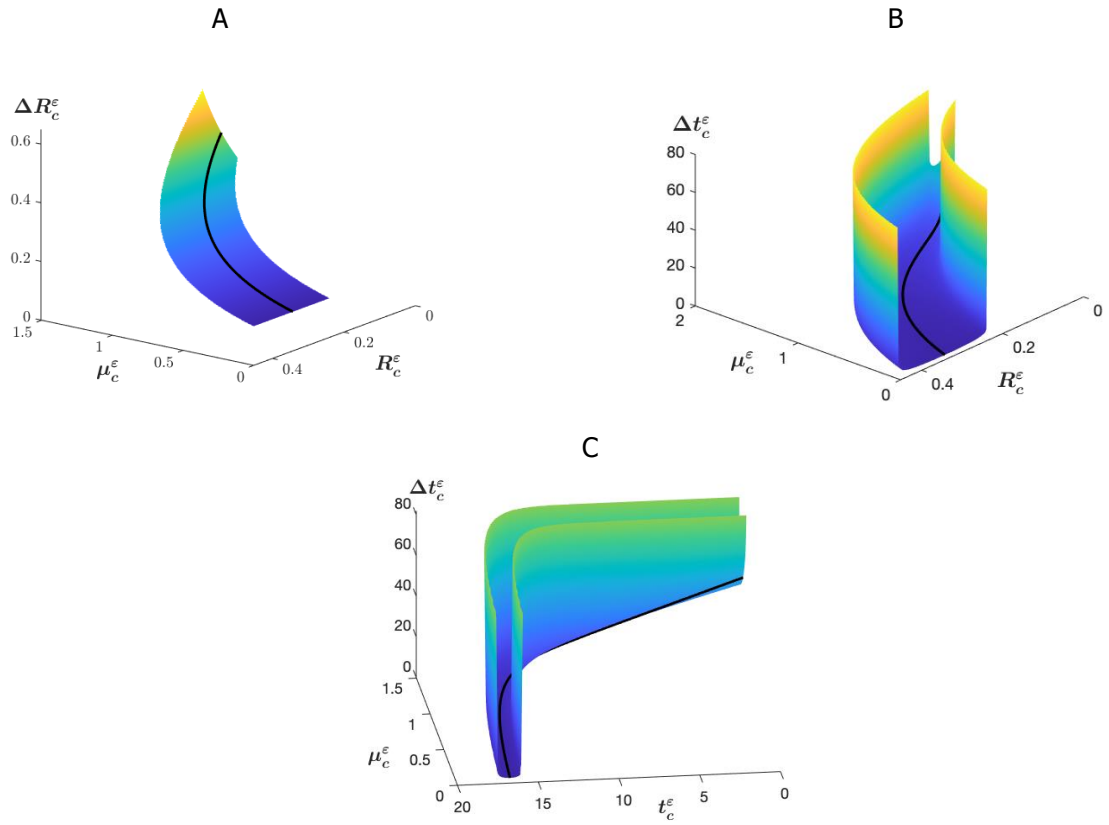

Figure 2. Near-to-optimal solutions for $\varepsilon = 0.1$. The optimal TNII is $R_\infty^* = 0.49$ and thus $R_\infty^\varepsilon = 0.59$. Panel A depicts a set of ε-optimal triples $(R_c^\varepsilon, \mu_c^\varepsilon, \Delta R_c^\varepsilon)$, i.e., set $\widehat{\Omega}^\varepsilon$. In Panel B, the dependence of Δt_c^ε on R_c^ε and μ_c^ε is depicted. Panel C presents a dependence of Δt_c^ε on t_c^ε and μ_c^ε. Black curves in all plots depict line (21). In Panels B and C, ΔR_c^ε and R_c^ε are translated into Δt_c^ε and t_c^ε according to Lemma 6 where $\mu_c = \mu_c^\varepsilon$, $t_c = t_c^\varepsilon$, $R_c = R_c^\varepsilon$, $\Delta R_c = \Delta R_c^\varepsilon$, $\Delta t_c = \Delta t_c^\varepsilon$.

Additionally, in Figure 3 we demonstrate some two-dimensional projections of sets presented in Figure 2 and how they change with the change of ε. (Three exemplary values of ε are used for illustraiton: 0.05; 0.1; and 0.15). The black curve depicts the projection of the optimal solution (20). Clearly, the projection of near-to-optimal sets $\widehat{\Omega}^\varepsilon$ onto $R_c^\varepsilon \mu_c^\varepsilon$-plane as well as their translation onto $t_c^\varepsilon \mu_c^\varepsilon$-plane expand with the increase of ε (see Panels A and C) allowing for more space for maneuvering. Simultaneously, the increase of ε allows for shorter corresponding ECP durations ΔR_c^ε (see Panel B).

From Panel A we can see that lower values of μ_c^ε provide broader ranges of values for the fraction of recovered at which an ε-optimal ECP should start. For example, for $\varepsilon = 0.1$ and $\mu_c^\varepsilon = 0$ this range is between 0.24 and 0.45. However, in terms of time t, this range is rather narrow, as can be seen from Panel C. In the above example, t_c ranges from 16 to 17.5 (numbers of infectious periods). This is because the translation of R_c into t_c is non-linear, especially for lower values of μ_c^ε: The epidemic dynamics proceeds very fast when approaching the peak of the epidemic curve. Since lower ECP durations ΔR_c^ε require lower values of μ_c^ε, if the central planner wants to achieve R_∞^ε with a shorter ECP, s/he has a rather short time period during which an ECP should kick off.

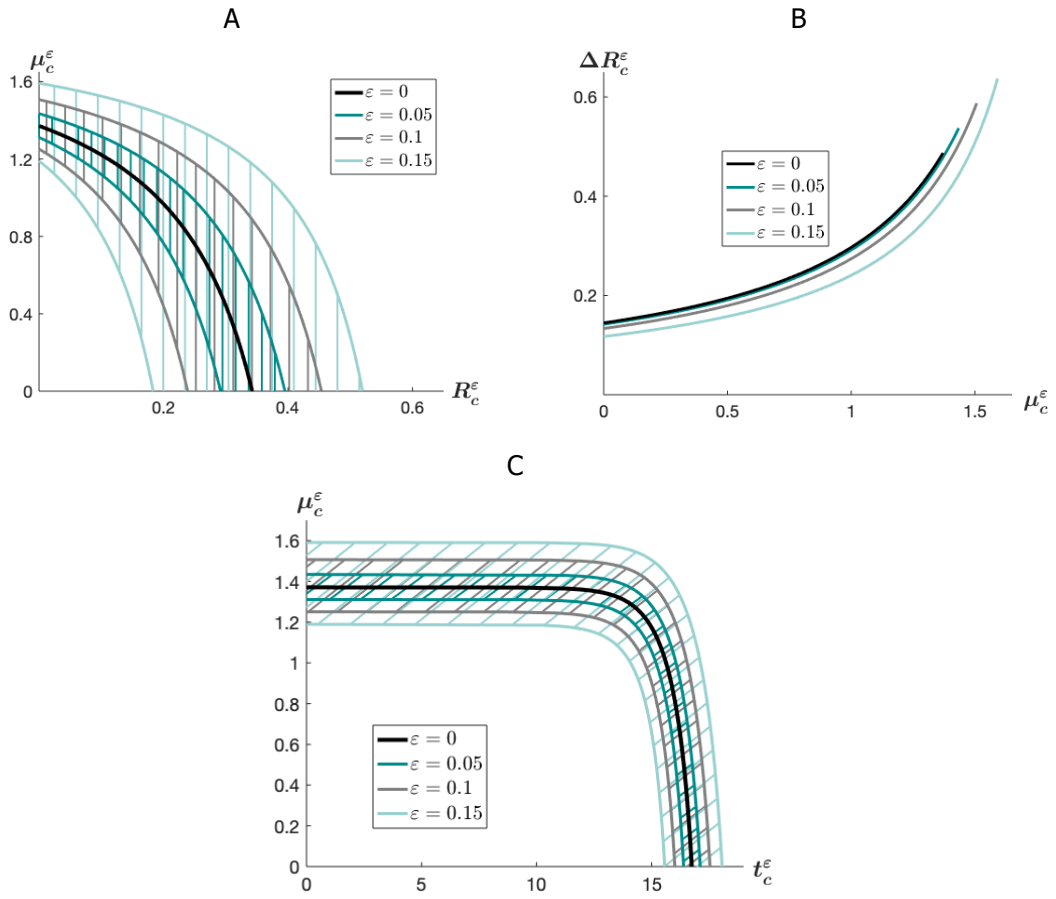

Figure 3. Panels A and B present orthogonal projections of $\widehat{\Omega}^\varepsilon$ onto $R_c^\varepsilon\mu_c^\varepsilon$- and $\mu_c^\varepsilon\Delta R_c^\varepsilon$- planes for $\varepsilon = 0.05$, $\varepsilon = 0.1$, $\varepsilon = 0.15$ (hatched). Panel C presents sets from Panel A where R_c^ε is translated into t_c^ε. Black line on both plots indicates corresponding orthogonal projections of the optimal solution (20) ($\varepsilon = 0$). Optimal $R_\infty^* = 0.49$ and thus $R_\infty^{0.05} = 0.54$, $R_\infty^{0.1} = 0.59$, and $R_\infty^{0.15} = 0.64$.

4. Conclusions and discussion

In this paper, we investigated a possibility to minimize the total number of infected individuals by applying an epidemic control policy in a compartmental epidemic model. We considered an epidemic control policy (ECP) that, starting from a certain time moment, reduces the basic reproduction number to a certain value and keeps this value constant over a certain period of time. Hence, the following three parameters are at the discretion of the central planner: the ECP inception time, the lowered basic reproduction number and the duration of the ECP. In real life, the reduction of the basic reproduction number can be achieved through various measures, for example, social distancing and pausing selected economic and social activities to reduce the rate of contacts between people as well as through progress in disease prevention. All these measures reduce the transmission rate of the disease. In the model that we analyze in this paper, we do not consider economic or other costs of these measures. We also do not discuss the feasibility of some of the policies which we consider mathematically admissible, for example, the feasibility of a significant reduction of the basic reproduction number way below 1, potentially even down to 0, and the feasibility of a very early inception of an epidemic control policy, potentially even as early as the beginning of the epidemic.

Hence, our model is arguably the simplest possible model that can describe an epidemic spread. This enables its analytical treatment and the derivation of the closed-form solutions for optimal and near-to-optimal controls. We make the following conclusions.

First, we observe that the total number of infected individuals over the entire time horizon of the disease dynamics depends only on the lowered basic reproduction number and the increment in the number of recovered individuals over the period of the ECP application; it does not depend, for example, on the number of infected or recovered at the time of the ECP inception.

Second, since ECPs do not imply any costs in our model, we naturally obtain that in order to minimize the total number of infected individuals, the central planner should impose an infinitely long ECP. There is a set of optimal ECPs; each ECP from this set yields the same (minimal) total number of infected individuals.

Third, ECPs from this optimal solution set are characterized by the tradeoff between the optimal lowered basic reproduction number and the optimal ECP inception time. Namely, stricter ECPs, that is ECPs with a lower value of the reduced basic reproduction number, require a later implementation time. Deviations from the set of optimal ECPs lead to higher total number of infected individuals in the model. In particular, too strict ECP and/or too early implementation later leads to a second wave of epidemic. There is no second wave if a too mild ECP is implemented and/or if an ECP is implemented too late, but in both cases, a higher level of the total number of infected individuals will be observed.

Fourth, the range of optimal lowered basic reproduction numbers and the range of optimal ECP inception times are narrower for a disease with a higher (original) basic reproduction number. This means that the more expansive is the propagation of the disease, the stricter and earlier measures are needed.

Fifth, an infinite ECP duration is likely to be infeasible in the real world; to address this problem, we derived ECPs of finite duration which are near-to-optimal, that is ECPs which lead to the total number of infected individuals that is higher than the optimal number by a small positive ε. Again, for each ε there is a set of near-to-optimal ECPs each yielding the same total number of infected individuals.

Sixth, a higher target number of infected individuals to be achieved allows for wider ranges of ECP inception times and the lowered basic reproduction numbers to choose from. Also, in this case, the ECP duration can be shorter.

Seventh, all pairs of optimal ECP inception times and lowered basic reproduction numbers belong to the set of near-to-optimal ECPs for any ε. This means that in order to achieve an ε-optimal number of recovered individuals, the central planner can always choose an optimal pair of these two parameters and then s/he only needs to determine the duration of the ECP application sufficient to reach an ε-optimal number of infected individuals. Or, put it differently, the social planner can choose a pair of optimal ECP inception time and lowered basic reproduction number and then keep this ECP for as long as it is feasible. The duration of this period will determine the number of infected individuals that will be achieved and the corresponding ε. This approach provides a robust way to ensure that the ECP duration is reasonable. Deviating from this approach contains the risk to select a pair of the other two control parameters which will require very long ECP.

The model presented in this paper is indeed the simplest possible epidemic control model. Despite its simplicity, it reveals a number of interesting and non-trivial qualitative observations as discussed above. Disease-caused mortality is not included in the model which simplifies the mathematical derivations. In the presence of mortality, formulas would become much more cumbersome, however, we do not expect that the results will be qualitatively different. Another possible extension to make the model more realistic could be to consider an age-structured population in which transmission and recovery rates depend on the individual's age. This dependence has been critical in case of COVID-19. It would also be interesting to consider other objective functions, for example, a meaningful objective function could be the number of infected individuals

at the peak of the epidemic curve. This number would characterize the maximum load on hospitals and the central planner would like to keep it well below the existing capacity. Taking into consideration costs of implementing an epidemic control policy is another very significant extension that can offer a wealth of new results. We leave some of these extensions to our next studies.

References

Bertozzi, A.L., Franco, E., Mohler, G., Short, M.B., Sledge, D., 2020. The challenges of modeling and forecasting the spread of COVID-19. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.2006520117

China population - Countrymeters [WWW Document], 2020. URL https://countrymeters.info/en/China (accessed 8.22.20).

Choisy, M., Guégan, J.-F., Rohani, P., 2007. Mathematical Modeling of Infectious Diseases Dynamics, in: Encyclopedia of Infectious Diseases. John Wiley & Sons, Ltd, pp. 379–404. https://doi.org/10.1002/9780470114209.ch22

Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J., Knuth, D.E., 1996. On the LambertW function. Adv. Comput. Math. 5, 329–359. https://doi.org/10.1007/BF02124750

Grigorieva, E.V., Khailov, E.N., Korobeinikov, A., 2016. Optimal Control for a SIR Epidemic Model with Nonlinear Incidence Rate. Math. Model. Nat. Phenom. 11, 89–104. https://doi.org/10.1051/mmnp/201611407

Hansen, E., Day, T., 2011. Optimal control of epidemics with limited resources. J. Math. Biol. 62, 423–451. https://doi.org/10.1007/s00285-010-0341-0

Hethcote, H.W., 2000. The Mathematics of Infectious Diseases. SIAM Rev. 42, 599–653.

Kermack, W.O., McKendrick, A.G., Walker, G.T., 1927. A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character 115, 700–721. https://doi.org/10.1098/rspa.1927.0118

Khoshnaw, S.H.A., Mohammad, N.A., Salih, R.H., 2017. Identifying Critical Parameters in SIR Model for Spread of Disease. Open J. Model. Simul. 05, 32. https://doi.org/10.4236/ojmsi.2017.51003

Liu, Y., Gayle, A.A., Wilder-Smith, A., Rocklöv, J., 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. J. Travel Med. 27. https://doi.org/10.1093/jtm/taaa021

Mandal, M., Jana, S., Nandi, S.K., Khatua, A., Adak, S., Kar, T.K., 2020. A model based study on the dynamics of COVID-19: Prediction and control. Chaos Solitons Fractals 136, 109889. https://doi.org/10.1016/j.chaos.2020.109889

Perkins, T.A., España, G., 2020. Optimal Control of the COVID-19 Pandemic with Non-pharmaceutical Interventions. Bull. Math. Biol. 82, 118. https://doi.org/10.1007/s11538-020-00795-y

Rowthorn, R.E., Laxminarayan, R., Gilligan, C.A., 2009. Optimal control of epidemics in metapopulations. J. R. Soc. Interface 6, 1135–1144. https://doi.org/10.1098/rsif.2008.0402

Tsay, C., Lejarza, F., Stadtherr, M.A., Baldea, M., 2020. Modeling, state estimation, and optimal control for the US COVID-19 outbreak. Sci. Rep. 10, 10711. https://doi.org/10.1038/s41598-020-67459-8

WHO COVID-19 Table [WWW Document], 2020. URL https://covid19.who.int/table (accessed 8.22.20).

Appendix

Proof of Lemma 1.

From the first and the second differential equations (DEs) of (2), we obtain that

$S(t) = S_0 e^{-\mu \int_0^t I(s)\mathrm{d}s}$ and $I(t) = I_0 e^{\mu \int_0^t S(s)\mathrm{d}s - t}, t \geq 0.$

Hence, due to assumption A1, $S(t) > 0$ and $I(t) > 0$, and $S(\cdot)$ is decreasing, $t \geq 0$.

Based on this, from the third DE of (2) we conclude that $R(\cdot)$ is increasing and hence $R(t) > 0$, $t > 0$.

Summing up all three DEs of (2) gives $\dot{S}(t) + \dot{I}(t) + \dot{R}(t) = 0$, and hence $S(t) + I(t) + R(t) = 1$, $t \geq 0$.

Therefore, all phase variables are bounded and take positive values over $(0, \infty)$. This proves the first and second statements of Lemma 1.

Let us prove the third statement of Lemma 1. As mentioned above, all state variables of system (2) are bounded from above by 1 and from below by 0. Being monotonic and continuous, $S(\cdot)$ and $R(\cdot)$ necessarily converge to some $S_\infty \in [0,1)$ and $R_\infty \in (0,1]$ as $t \to \infty$. Then, thanks to equality $S(t) + I(t) + R(t) = 1$ for all $t \in [0, \infty)$, as $t \to \infty$ function $I(\cdot)$ also converges to some $I_\infty \in [0,1]$. As point $(S_\infty, I_\infty, R_\infty)$ must be an equilibrium point, we conclude that $I_\infty = 0$ and then $S_\infty = 1 - R_\infty$. The third statement of Lemma 1 is proven.

Proof of Lemma 2.

Consider (5) over $[0, \infty)$, whose first- and second-order derivatives are as follows:

$$I'(R) = -1 + \mu S_0 e^{-\mu R}, \qquad I''(R) = -\mu^2 S_0 e^{-\mu R} < 0.$$

Obviously, $I(\cdot)$ is strictly concave, and $R^* = \mu^{-1} \ln S_0 \mu > 0$ is the unique maximum point of function $I(\cdot)$. Hence, $I(\cdot)$ is increasing over $[0, R^*)$ and decreasing over (R^*, ∞). Since $I(0) > 0$, $I(R^*) > 0$ and $I(R) \to -\infty$ as $R \to \infty$, there exists a unique root $R_\infty > R^* > 0$ of equation $I(R) = 0$ over $R \in [0, \infty)$, and $I(R) > 0$ for $R \in [0, R_\infty)$ and $I(R) < 0$ for $R > R_\infty$. The value R_∞ coincides with the one from Lemma 1 thanks to the established connection between systems (2) and (3).

Proof of Lemma 3.

The one-to-one correspondence between solutions of IVPs (2) and (3), presented in the first two statements of Lemma 3, will be proven if mapping $t \to R$ is proven to be a diffeomorphism. Let us prove this. Substituting (5) in the equation of IVP (2), we obtain a Cauchy problem, whose solution defines mapping $t \to R$:

$$\dot{R}(t) = 1 - R(t) - S_0 e^{-\mu R(t)}, \qquad R(0) = 0.$$

A simple analysis of the right-hand side of this equation leads to the result that $t \to R$ is a strictly increasing smooth function, and if $t \to \infty$ then $R \to R_\infty$. Hence, mapping $t \to R$ has a smooth inverse, provided by the corresponding Cauchy problem

$$\dot{t}(R) = (1 - R - S_0 e^{-\mu R})^{-1}, \qquad t(0) = 0,$$

and, therefore, it is a diffeomorphism.

Proof of Lemma 4.

To analyze system (2) under (7), we consider three consequent IVPs of type (2) over intervals $t \in [0, t_c)$, $t \in [t_c, t_c + \Delta t_c)$, and $t \in [t_c + \Delta t_c, \infty)$ respectively, each of which with a constant μ as specified in (7) and an appropriate initial condition to guarantee a continuous conjunction of the solution parts. All three IVPs are analogous to (2) and thus their combined solution has all properties of system (2) established in Lemma 1.

Proof of Lemma 5.

Obviously, after substitution of (9) into (10), over each of the three segments $[0, R_c), [R_c, R_c + \Delta R_c)$ and $[R_c + \Delta R_c, R_\infty)$ function $I(\cdot)$ has the same behavior as function $I(\cdot)$ from Lemma 2, namely, it is strictly concave and tends to minus infinity as $R \to \infty$ (see proof of Lemma 2). Therefore, analogously to Lemma 2, relations (11) ensure that there exists unique value $R_\infty > R_c + \Delta R_c$ such that $I(R_\infty) = 0$, and $I(R) > 0$, $R \in$

$[0, R_\infty)$ and $I(R) < 0$, $R \in (R_\infty, \infty)$. The value R_∞ coincides with the one from Lemma 4 thanks to the established connection between systems (2) and (3).

Proof of Lemma 6.

The proof of Lemma 6 is fully analogous to the proof of Lemma 3.

Proof of Proposition 1.

The objective function $(\mu_0 - \mu_c)\Delta R_c$ has no local extremum in $\widehat{\Omega}$ and, hence, it achieves it supremum at one of its boundaries. As function $(\mu_0 - \mu_c)\Delta R_c$ is strictly increasing in ΔR_c, it achieves its supremum at the right boundary of the last inequality in (19), that is:

$$\Delta R_c = 1 + \mu_c^{-1} W\left(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c}\right) - R_c. \qquad (22)$$

Substituting (22) into the objective function $(\mu_0 - \mu_c)\Delta R_c$, we obtain its representation as a function of μ_c and R_c as follows:

$$F(R_c, \mu_c) = (\mu_0 - \mu_c)\left(1 + \mu_c^{-1} W\left(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c}\right) - R_c\right), \qquad (23)$$

which is to be maximized subject to (14) and (15).

This maximization problem has a solution either at a stationary point of $F(\cdot,\cdot)$ or at the boundary of the domain in R^2 defined by (14) and (15).

First, let us look for and analyze stationary points of $F(\cdot,\cdot)$. Consider the case of $\mu_c = 0$. Then function (23) takes the form

$F(R_c, 0) = \mu_0(1 - S_0 e^{-\mu_0 R_c} - R_c)$.

Obviously, it is strictly concave in R_c and has a unique positive maximum point

$$\bar{R} = \mu_0^{-1} \ln S_0 \mu_0.$$

Let us prove that $\bar{R} < 1 + \mu_0^{-1} W(-S_0 \mu_0 e^{-\mu_0})$. The following inequality is true:

$$-1 < W(-S_0 \mu_0 e^{-\mu_0}) < 0.$$

Then

$1 < -\frac{1}{W(-S_0 \mu_0 e^{-\mu_0})} \Leftrightarrow S_0 \mu_0 < e^{\mu_0} \frac{-S_0 \mu_0 e^{-\mu_0}}{W(-S_0 \mu_0 e^{-\mu_0})}$.

Employing the definition of the Lambert function, we further simplify this to the following

$S_0 \mu_0 < e^{\mu_0 + W(-S_0 \mu_0 e^{-\mu_0})}$.

Taking the logarithm, we obtain

$$\ln S_0 \mu_0 < \mu_0 + W(-S_0 \mu_0 e^{-\mu_0}) \Leftrightarrow$$
$$\bar{R} = \mu_0^{-1} \ln S_0 \mu_0 < 1 + \mu_0^{-1} W(-S_0 \mu_0 e^{-\mu_0}).$$

Thus, $\bar{R}$ satisfies inequalities (14). Let us calculate the value of the objective function:

$$F(\bar{R}, 0) = \mu_0 - 1 - \ln S_0 \mu_0 > 0.$$

Now let us consider the case of $\mu_c > 0$. We calculate the first-order derivatives of $F(\cdot,\cdot)$ as follows:

$$F'_{R_c}(R_c, \mu_c) = -\frac{(\mu_0 - \mu_c)\left(W\left(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c}\right)\mu_0 + \mu_c\right)}{\mu_c\left(1 + W(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c})\right)},$$

and

$$F'_{\mu_c}(R_c, \mu_c) = -\frac{\left(-R_c \mu_c + W\left(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c}\right) + \mu_c\right)\left(W\left(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c}\right)\mu_0 + \mu_c\right)}{\mu_c^2\left(1 + W(-S_0 \mu_c e^{-\mu_c} e^{-(\mu_0 - \mu_c)R_c})\right)}.$$

The following one-parametric family (R_c^*, μ_c^*):

$$\mu_c^* = \frac{\ln(S_0 \mu_0) - \mu_0 R_c^*}{1 - \mu_0^{-1} - R_c^*}. \qquad (24)$$

Solves the system of equations $F'_{R_c}(R_c, \mu_c) = 0$, $F'_{\mu_c}(R_c, \mu_c) = 0$.

It is easy to show that $\mu_c^* \geq 0$ only for $R_c^* \in [0, \bar{R}]$. Note that if $R_c^* = \bar{R}$ then $\mu_c^* = 0$, which coincides with the case considered above. Also μ_c^* is decreasing in R_c^* and if $R_c^* = 0$, then $\mu_c^* = \frac{\ln(S_0\mu_0)}{1-\mu_0^{-1}} < \mu_0$. Thus, family (24), $R_c^* \in [0, \bar{R}]$, satisfies inequalities (14), (15).

Let us calculate

$$F(R_c^*, \mu_c^*) = (\mu_0 - \mu_c^*)\left(1 + W\left(-S_0\mu_c^* e^{-\mu_c^*} e^{-(\mu_0-\mu_c^*)R_c^*}\right) - R_c^*\right) \equiv \mu_0 - 1 - \ln S_0\mu_0 > 0.$$

Let us now find the maximal value of $F(\cdot,\cdot)$ at the boundaries of inequalities (14), (15). The case of $\mu_c = 0$ was considered above. Let us consider the case of $R_c = 0$, $\mu_c \in (0, \mu_0]$. The stationary points of function $F(0,\cdot)$ in this case are solutions to equation

$$F'_{\mu_c}(0, \mu_c) = -\frac{(W(-S_0\mu_c e^{-\mu_c}) + \mu_c)(W(-S_0\mu_c e^{-\mu_c})\mu_0 + \mu_c)}{\mu_c^2\left(1 + W(-S_0\mu_c e^{-\mu_c})\right)} = 0.$$

There exists a unique stationary point $\bar{\mu}_c = \frac{\ln(S_0\mu_0)}{1-\mu_0^{-1}}$. This point coincides with the case above, where $\mu_c^* = \bar{\mu}_c$ and $R_c^* = 0$. Hence, $F(0, \bar{\mu}_c) = \mu_0 - 1 - \ln S_0\mu_0$.

In the remaining cases of $R_c \in [0, 1 + \mu_0^{-1}W(-S_0\mu_0 e^{-\mu_0}))$, $\mu_c = \mu_0$ and $R_c = 1 + \mu_0^{-1}W(-S_0\mu_0 e^{-\mu_0})$, $\mu_c \in (0, \mu_0]$, the value of $F(\cdot,\cdot)$ is zero and, therefore, these pairs are not optimal.

Thus, the solution to the maximization problem (19) is (R_c^*, μ_c^*), $R_c^* \in [0, \bar{R}]$.

Substituting (24) into (22), we obtain

$$\Delta R_c^* = 1 + (\mu_c^*)^{-1}W\left(-S_0\mu_c^* e^{-\mu_c^*} e^{-(\mu_0-\mu_c^*)R_c^*}\right) - R_c^* = 1 - \mu_0^{-1} - R_c^*$$

and finally, $R_\infty^* = R_c^* + \Delta R_c^* = 1 - \mu_0^{-1}$.

Proof of Proposition 2.

From formula (12), where $R_\infty = R_\infty^\varepsilon$, we immediately obtain

$$\Delta R_c^\varepsilon = \frac{R_\infty^\varepsilon\mu_0 - \ln\sigma_\infty^\varepsilon}{\mu_0 - \mu_c^\varepsilon} > 0. \qquad (25)$$

Then the set $\widehat{\Omega}^\varepsilon$ of all triples $(R_c^\varepsilon, \mu_c^\varepsilon, \Delta R_c^\varepsilon)$ is defined by (14)-(16), where $(R_c, \mu_c, \Delta R_c) = (R_c^\varepsilon, \mu_c^\varepsilon, \Delta R_c^\varepsilon)$ and (25). Substituting (25) in (16), we obtain.

$$\frac{R_\infty^\varepsilon\mu_0 - \ln\sigma_\infty^\varepsilon}{\mu_0 - \mu_c^\varepsilon} < 1 + (\mu_c^\varepsilon)^{-1}W\left(-S_0\mu_c^\varepsilon e^{-\mu_c^\varepsilon} e^{-(\mu_0-\mu_c^\varepsilon)R_c^\varepsilon}\right) - R_c^\varepsilon$$

Denote

$$G(R_c^\varepsilon, \mu_c^\varepsilon) = 1 + (\mu_c^\varepsilon)^{-1}W\left(-S_0\mu_c^\varepsilon e^{-\mu_c^\varepsilon} e^{-(\mu_0-\mu_c^\varepsilon)R_c^\varepsilon}\right) - R_c^\varepsilon - \frac{R_\infty^\varepsilon\mu_0 - \ln\sigma_\infty^\varepsilon}{\mu_0 - \mu_c^\varepsilon}.$$

Then inequalities $G(R_c^\varepsilon, \mu_c^\varepsilon) > 0$, (14), and (15) define set $\widehat{\Omega}^\varepsilon$.

As long as $\mu_c^\varepsilon \in [0, \bar{\mu}_c^\varepsilon)$, equation $G(R_c^\varepsilon, \mu_c^\varepsilon) = 0$ has two solutions: $R_c^\varepsilon = \underline{R}_c^\varepsilon$ and $R_c^\varepsilon = \bar{R}_c^\varepsilon$. The derivative becomes:

$$G'_{R_c^\varepsilon}(R_c^\varepsilon, \mu_c^\varepsilon) = -\frac{\left(W\left(-S_0\mu_c^\varepsilon e^{-\mu_c^\varepsilon} e^{-(\mu_0-\mu_c^\varepsilon)R_c^\varepsilon}\right)\mu_0 + \mu_c^\varepsilon\right)}{\mu_c^\varepsilon\left(1 + W\left(-S_0\mu_c^\varepsilon e^{-\mu_c^\varepsilon} e^{-(\mu_0-\mu_c^\varepsilon)R_c^\varepsilon}\right)\right)}$$

The solution to equation $G'_{R_c^\varepsilon}(R_c^\varepsilon, \mu_c^\varepsilon) = 0$ is the following one-parametric family:

$$\mu_c^\varepsilon = \frac{\ln(S_0\mu_0) - \mu_0 R_c^\varepsilon}{1 - \mu_0^{-1} - R_c^\varepsilon}, R_c^\varepsilon \in [0, \bar{R}]. \qquad (26)$$

It coincides with formula (24).

The second derivative becomes:

$$G''_{(R_c^\varepsilon)^2}(R_c^\varepsilon, \mu_c^\varepsilon) = \frac{(\mu_0 - \mu_c)^2 W\left(-S_0 \mu_c^\varepsilon e^{-\mu_c^\varepsilon} e^{-(\mu_0 - \mu_c^\varepsilon) R_c^\varepsilon}\right)}{\mu_c \left(1 + W\left(-S_0 \mu_c^\varepsilon e^{-\mu_c^\varepsilon} e^{-(\mu_0 - \mu_c^\varepsilon) R_c^\varepsilon}\right)\right)^3} < 0.$$

Hence, $G(\cdot, \mu_c^\varepsilon)$ is concave in R_c^ε and (26) is its maximum point. Substituting (26) into $G(\cdot,\cdot)$, we obtain:

$$G(R_c^\varepsilon, \mu_c^\varepsilon) = 1 + (\mu_c^\varepsilon)^{-1} W\left(-S_0 \mu_c^\varepsilon e^{-\mu_c^\varepsilon} e^{-(\mu_0 - \mu_c^\varepsilon) R_c^\varepsilon}\right) - R_c^\varepsilon - \frac{R_\infty^\varepsilon \mu_0 - \ln \sigma_\infty^\varepsilon}{\mu_0 - \mu_c^\varepsilon} = \frac{-\gamma \ln(1 - \varepsilon \mu_0) - \mu_0 \varepsilon}{\mu_0 - \mu_c^\varepsilon} > 0.$$

Therefore, family (26) fully belongs to $\widehat{\Omega}^\varepsilon$. It is easy to show that one-parametric family (26) belongs to the set of all such $R_c^\varepsilon, \mu_c^\varepsilon$ that $R_c^\varepsilon \in (\underline{R}_c^\varepsilon, \bar{R}_c^\varepsilon), R_c^\varepsilon \geq 0, \mu_c^\varepsilon \in [0, \bar{\mu}_c^\varepsilon)$.

Aggregating the obtained above information about function $G(\cdot,\cdot)$, we finally conclude that inequalities $G(R_c^\varepsilon, \mu_c^\varepsilon) > 0$, (14)-(16) are equivalent to inequalities $R_c^\varepsilon \in (\underline{R}_c^\varepsilon, \bar{R}_c^\varepsilon), R_c^\varepsilon \geq 0, \mu_c^\varepsilon \in [0, \bar{\mu}_c^\varepsilon)$.