TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Operations Research and Control Systems

SWM ORCOS

Numerical Approximations in Optimal Control of a Class of Heterogeneous Systems

Vladimir M. Veliov

Research Report 2015-01

February, 2015

Operations Research and Control Systems
Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology

Research Unit ORCOS
Argentinierstraße 8/E105-4,
1040 Vienna, Austria
E-mail: orcos@tuwien.ac.at

Numerical Approximations in Optimal Control of a Class of Heterogeneous Systems[*]

V.M. Veliov

Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Austria, veliov@tuwien.ac.at

Abstract

The paper presents a numerical procedure for solving a class of optimal control problems for heterogeneous systems. The latter are described by parameterized systems of ordinary differential equations, coupled by integrals along the parameter space. Such problems arise in economics, demography, epidemiology, management of biological resources, etc. The numerical procedure include discretization and a gradient projection method for solving the resulting discrete problem. A main point of the paper is the performed error analysis, which is based on the property of metric regularity of the system of necessary optimality conditions associated with the considered problem.

Keywords: optimal control, distributed systems, numerical methods, epidemiology

1 Introduction

In principle, heterogeneous control systems, as described in [14], include age/size-structured systems, advection-reaction systems, epidemiological models for heterogeneous populations, and a variety of economic models involving agents with diverse individual features. In this paper we present a numerical approach for solving optimal control problems for such systems, focusing on the following special class of heterogeneous systems:

$$\dot{x}(t,p) = f(p, x(t,p), y(t,p), u(t,p)), \quad x(0,p) = x_0(p), \tag{1}$$

$$y(t,p) = \int_P g(p,q,x(t,q),u(t,q))\,\mathrm{d}q. \tag{2}$$

Here $t \in [0,T]$ is interpreted as time, "dot" means differentiation with respect to t, p is a scalar parameter taking values in an interval $P = [0,\Pi]$. The state variables $x : [0,T] \times P \to \mathbf{R}^n$, $y : [0,T] \times P \to \mathbf{R}^m$, and the control variable $u : [0,T] \times P \to U \subset \mathbf{R}^r$ belong to functional spaces specified below in such a way that together with appropriate assumptions for the functions $f : P \times \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^r \to \mathbf{R}^n$ and $g : P \times P \times \mathbf{R}^n \times \mathbf{R}^r \to \mathbf{R}^m$ equations (1), (2) make sense for a given initial condition x_0.

We associate with system (1), (2) the following optimal control problem:

$$\min_{x,y,u} \left\{ \int_P l(x(T,p))\,\mathrm{d}p \; + \; \int_0^T \int_P L(p, x(t,p), y(t,p), u(t,p))\,\mathrm{d}p\,\mathrm{d}t \right\}, \tag{3}$$

[*] Supported by the Austrian Scienece Foundation (FWF) Grant P 24125-N13.

$$u(t, p) \in U, \tag{4}$$

where l and L are scalar functions.

Numerous particular optimal control models of the above type can be found in the literature (see e.g. [4, 14, 11] and the bibliography therein), but the main applications we have in mind are in the dynamics of populations, where p is interpreted either as a genotype projection or as some other individual-specific indicator. In these considerations $x(t, \cdot)$ represents the density of the (multi-dimensional) population along the heterogeneity space P, and y is an aggregated variable (coupling equations (1) and (2) together) that represents "externalities" influencing the dynamics.

System (1), (2) does not directly cover age-structured population systems that play a crucial role in population dynamics and economics ([3, 9]). However, modifications of the subsequent considerations apply also to optimal control of such systems. Modifications of the presented approximation scheme and the corresponding error analysis for more general systems (such as size-structured systems or advection-reaction systems) is possible, but requires additional non-straightforward work.

We mention that an explicit dependence of the data on the time t is suppressed only for shortness. Similarly, the function g may depend on y in a sufficiently "regular" way (see [15, (A3)]). If a non-distributed control $v : [0, T] \to V$ is involved in the problem, then minor modifications are needed, as explained in Section 6.

The aim of this paper is to present a numerical procedure for solving optimal control problems of the type of (1)–(4). The numerical procedure proposed below employs the Euler discretization scheme for approximation of Pontryagin's type necessary optimality conditions for the problem. The latter involve differential equations, integral relations, and an inclusion (representing the condition of maximization of the Hamiltonian), that is, a system of generalized equations. A gradient projection technique is applied for solving this system of generalized equations. A main point of this paper is the error analysis, which provides an error estimate based on a "metric regularity assumption" for the system of optimality conditions.

We mention that in our computational tools we always use second order discretization schemes, rather than Euler's one. The reason for which we use exactly a second order scheme (not first or higher than second order) is explained in the discussions in Section 7. However, we base our exposition on the Euler scheme due to: (i) better readability that allows to grasp the idea; (ii) to prove a second order accuracy only on the assumption of metric regularity mentioned above is an open question.

The paper is organized as follows. In the next section we give a particular example from epidemiology. In Section 3 we present some preliminary material – assumptions, strict formulation of the problem, known optimality conditions. Section 4 presents the numerical method based on discretization and a gradient projection procedure. Section 5 is devoted to the error analysis. Some extensions and discussions are given in the two final sections.

2 An example from epidemiology

Models describing the spread of infectious diseases in heterogeneous populations are well known (see e.g. [4, Chapter 6], [11]). The one below is a typical representative, where, however, a control is involved (interpreted as prevention), thus an optimal control problem can be considered.

Below $p \in [0, \Pi] =: P$ is a scalar parameter representing a trait related to the level of risk of infection of individuals having this trait (say, intensity of risky contacts, state of the immune system, personal hygiene, or a combination of the above ones). The population has a fixed size and is divided into three groups: susceptible, infected, and recovered; $S(t,p)$, $I(t,p)$, $R(t,p)$, $p \in P$. Here $S(t,\cdot)$ is the density of the susceptible individuals at time t, similarly for I and R. Thus $\int_P S(t,p)\,\mathrm{d}p$ is the size of the susceptible sub-population, etc. Moreover, a control $u(t,p) \in [v,1]$, $v \in (0,1)$, is involved, interpreted as intensity of prevention applied to susceptible individuals of treat p. The dynamics of the disease is described by the following system:

$$\begin{aligned}
\dot{S}(t,p) &= -\sigma(p)\,u(t,p)\,J(t)\,S(t,p), \quad S(0,p) = S_0(p),\\
\dot{I}(t,p) &= \sigma(p)\,u(t,p)\,J(t)\,S(t,p) - \rho I(t,p), \quad I(0,p) = I_0(p),\\
\dot{R}(t,p) &= \rho I(t,p), \quad R(0,p) = R_0(p),\\
J(t) &= \int_P \alpha(p) I(t,p)\,\mathrm{d}p,
\end{aligned}$$

where $\sigma(p)$ combines the strength of the disease with the specific level of risk of individuals of treat p (without prevention), ρ is the recovery rate, $\alpha(p)$ is the infectiousness of infected individuals of treat p. The prevention control reduces $\sigma(p)$ to $\sigma(p)u(t,p)$. Since the population size is obviously constant, $J(t)$ measures the infectiousness of the environment in which the susceptibles live, thus $\sigma(p)u(t,p)J(t)$ is the incidence rate if control $u(t,p)$ is applied.

A reasonable objective function to be minimized is

$$\int_0^T \int_P [\beta I(t,p) + c(p, u(t,p))\,S(t,p)]\,\mathrm{d}p\,\mathrm{d}t,$$

where β the economic losses from one individual being infected in a unit of time, $c(p,u)$ is the per capita expenditure of applying control u to susceptibles of trait p. Typically $c(p,u)$, $u \in [v,1]$ is strongly convex and decreasing, with $c(p,1) = 0$ (no prevention effort).

Versions of the above problem will be a subject of a separate specialized study due to the intriguing fact that it may happen (even in the case where $c(p,u) = c(u)$ is independent of p) that it is optimal to allocate more prevention to individuals of high risk, but it may also happen (depending on the data specifications) that it is optimal to put more prevention to individuals of low risk. This fact has important policy implications, especially in the case of budget constraints $\int_P u(t,p)\,\mathrm{d}p \le B$ (such constraints are not involved in our general model, but are also tractable).

3 Preliminary material

In this section we formulate the needed assumption, and first order necessary optimality conditions which will be used in the next sections. This preliminary material is basically an extraction and adaptation of results from [15] and [14].

First we define the functional spaces used below. Denote $D := [0,T] \times P$. The space $\mathcal{X}$ consists of all functions $x : D \to \mathbf{R}^n$ which are Lipschitz continuous in t, measurable in p, and the norm $\|x\| := \|x\|_{L\infty(D)} + \|\dot{x}\|_{L\infty(D)}$ is finite. By $\mathcal{X}_0$ we denote the subset of $\mathcal{X}$ consisting of those x which satisfy $x(0,\cdot) = x_0(\cdot)$. Further, $\mathcal{Y} := L_\infty(D)$ and $\mathcal{U} = \{u \in L_\infty(D) : u(t,p) \in U$ for a.e. $(t,p)\}$. Finally, $\mathcal{S} := \mathcal{X}_0 \times \mathcal{Y} \times \mathcal{U}$.

The next assumptions will be standing in all the paper.

Standing assumptions. The set $U \subset \mathbf{R}^r$ is convex and compact. The initial state x_0 belongs to $L_\infty(P)$. The functions f, g, l, L are differentiable with respect to all arguments, and the first derivatives are locally Lipschitz continuous. There is a constant $\bar{C}$ such that

$$|f| + |g| \le \bar{C}(1 + |x| + |y|) \quad \text{for every } t \in [0,T],\ p,\, q \in P,\ x \in \mathbf{R}^n,\ y \in \mathbf{R}^m,\ u \in U.$$

Given $u \in \mathcal{U}$, there is a unique pair $(x,y) \in \mathcal{X} \times \mathcal{Y}$ satisfying equations (1), (2) almost everywhere on D (see e.g. [14, Theorem 1]). The triplet $s = (x,y,u) \in \mathcal{S}$ will be called *process* and the corresponding value of the objective functional in (3) will be denoted by $J(u)$.

For a given reference process $s = (x,y,u) \in \mathcal{S}$ we define the *adjoint equation*

$$\begin{aligned}
(5)\qquad -\dot{\xi}(t,p) &= \xi(t,p)\, f_x(p, s(t,p)) + \int_P \eta(t,q)\, g_x(q,p,s(t,p))\, \mathrm{d}q + L_x(p, s(t,p)),\\
(6)\qquad & \xi(T,p) = l_x(x(T,p)),\\
(7)\qquad \eta(t,p) &= \xi(t,p)\, f_y(p, s(t,p)) + L_y(p, s(t,p)).
\end{aligned}$$

Notice that $\xi \in \mathbf{R}^n$ and $\eta \in \mathbf{R}^m$ are considered as row-vectors, while x, y and u are column vectors. The derivatives f_x, f_y, etc. are matrices of dimension $n \times n$, $n \times m$, etc.

The above linear system has a solution $(\xi,\eta) \in \mathcal{X} \times \mathcal{Y}$ on D (this follows from [14, Proposition 1]).

Further we sometimes abridge $z := (s,\xi,\eta) \in \mathcal{Z} := \mathcal{S} \times \mathcal{X} \times \mathcal{Y} = \mathcal{X}_0 \times \mathcal{Y} \times \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$. Notice that if $z \in \mathcal{Z}$ then $z(t,\cdot) \in L_\infty(P)$. For a function $z = (x,y,u,\xi,\eta) \in L_\infty(P)$ (as it is usual, we overload the notation z, since here z does not depend on t) we define the *Hamiltonian*

$$(8)\qquad H(z) = \int_P \left[\xi(p) f(p,s(p)) + L(p,s(p)) + \eta(p) \int_P g(p,q,s(q))\, \mathrm{d}q\right] \mathrm{d}p.$$

The Hamiltonian is differentiable in z and its derivative has a functional representation, which is also an element of $L_\infty(P)$, namely,

$$(9)\qquad H_s(p,z) = \xi(p)\, f_s(p,s(p)) + L_s(p,s(p)) + \int_P \eta(q)\, g_s(q,p,s(p))\, \mathrm{d}q,$$

$$(10)\qquad H_\xi(p,z) = f(p,s(p)), \quad H_\eta(p,z) = \int_P g(p,q,s(q))\, \mathrm{d}q.$$

The following proposition follows from the second assertion of Theorem 3 in [14] (Pontryagin's type necessary optimality conditions).

Proposition 1 *Let $\hat{s} \in \mathcal{S}$ be an optimal solution of problem (1)–(4). Let $(\hat{\xi},\hat{\eta}) \in \mathcal{X} \times \mathcal{Y}$ be the corresponding unique solution of the adjoint system (5)–(7). Then for almost every $(t,p) \in D$*

$$(11)\qquad H_u(\hat{s}(t,\cdot), \hat{\xi}(t,\cdot), \hat{\eta}(t,\cdot))(p) \in -N_U(\hat{u}(t,p)),$$

where

$$N_U(v) = \begin{cases} \emptyset & \text{if } v \notin U,\\ \{\nu \in \mathbf{R}^r :\ \langle \nu, w - v\rangle \le 0 \ \text{ for all } \ w \in U\} & \text{if } v \in U. \end{cases}$$

is the normal cone to U at v.

Due to expressions (9), (10) one can rewrite the primal-adjoint system (1)–(2), (5)–(7) in Hamiltonian form. Then Proposition 1 can be reformulated as follows.

Proposition 2 *Let $\hat{s} \in \mathcal{S}$ be an optimal solution of problem (1)–(4). Then there exists a pair $(\hat{\xi}, \hat{\eta}) \in \mathcal{X} \times \mathcal{Y}$ such that $\hat{z} := (\hat{s}, \hat{\xi}, \hat{\eta}) \in \mathcal{Z}$ satisfies the following system of generalized*[1] *equations:*

$$\begin{aligned}
&(12) && 0 = -\dot{x}(t,p) + H_\xi(p, z(t,\cdot)),\\
&(13) && 0 = -y(t,p) + H_\eta(p, z(t,\cdot)),\\
&(14) && 0 = \dot{\xi}(t,p) + H_x(p, z(t,\cdot)),\\
&(15) && 0 = -\xi(T,p) + l_x(x(T,p)),\\
&(16) && 0 = -\eta(t,p) + H_y(p, z(t,\cdot)),\\
&(17) && 0 \in H_u(p, z(t,\cdot)) + N_U(u(t,p)).
\end{aligned}$$

The right-hand side of the above system is defined in $\mathcal{Z}$, while the images are in $\mathcal{G} := \mathcal{Y} \times \mathcal{Y} \times \mathcal{Y} \times L_\infty(P) \times \mathcal{Y} \times \mathcal{Y}$. Thus the right-hand side defines a set-valued (due to the last inclusion) mapping $F : \mathcal{Z} \rightrightarrows \mathcal{G}$ and system (12)–(17) takes the form

$$0 \in F(z). \tag{18}$$

This inclusion represent the set of necessary optimality conditions for problem (1)–(4) and our aim below will be to find numerically some $\hat{z} \in \mathcal{Z}$ that is close to the set of solutions of (18).

4 Numerical procedure

The numerical approach presented in this paper consists (as usual in optimal control theory) of two parts. The first part is to pass to a discretized version of problem (1)–(4), then to solve the resulting finite dimensional optimization problem and to interpret the results in the terms of the original problem. In practice we use for discretization of the differential equations a second order Runge-Kutta scheme (the Heun scheme) and for integration we use the corresponding quadrature formula (the trapezoidal rule). However, for more transparency in this paper we employ the Euler scheme and the rectangular integration rule. A version of the gradient projection method in the control space is used for solving the mathematical programming problem, where the adjoin system is used in the calculation of the gradients.

4.1 Discretization of problem (1)–(4)

Let N and M be (presumably large) natural numbers, $h := T/N$, $\tau := \Pi/M$, $t_k := kh$, $k = 0, \ldots, N$, $p_i = i\tau$, $i = 0, \ldots, M$. The box with vertices (t_k, p_i), (t_{k+1}, p_i), (t_{k+1}, p_{i+1}), (t_k, p_{i+1}) in D will be denoted by Q_{ki}, $k = 0, \ldots, N-1$, $i = 0, \ldots, M-1$. In order to understand the discrete version of system (1)–(4) below, we clarify that x_{ki}, y_{ki}, ξ_{ki}, η_{ki} will be interpreted as approximations of $x(t_k, p_i)$, ... $\eta(t_k, p_i)$, respectively, while $u_{k,i}$ will be interpreted as a constant approximation of u in the box Q_{ki}, (In fact, the index k of y_{ki} and η_{ki} will vary only from 0 to $N-1$ and from 1 to N, respectively.) This interpretation of u_{ki} is important in the error analysis, due to the presence of inclusion (17) in the system of necessary optimality conditions, where the

[1] The term "generalized" refers to the fact that the last relation is an inclusion rather than an equation.

normal cone $N_U(u)$ may be (and usually is) a discontinuous set-valued mapping, therefore it is important to view the control u as constant on every box Q_{ki}.

Correspondingly, we define the discrete analogs of the spaces $\mathcal{X}$, $\mathcal{X}_0$, $\mathcal{Y}$, $\mathcal{U}$, and $\mathcal{S}$, putting a $\tilde{}$ on top of the symbols, with the l_∞-norm in each of them. So, for example, $\tilde{\mathcal{X}} = \{x_{ki} : k = 0, \ldots, N,\ i = 0, \ldots, M-1\}$, $\tilde{\mathcal{X}_0} = \{x \in \tilde{\mathcal{X}} : x_{0,i} = x_0(p_i)\}$, etc. As above we abridge $s = (x, y, u) \in \tilde{\mathcal{S}} := \tilde{\mathcal{X}_0} \times \tilde{\mathcal{Y}} \times \tilde{\mathcal{U}}$, $z = (s, \xi, \eta) \in \tilde{\mathcal{Z}} := \tilde{\mathcal{S}} \times \tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}$.

Formally applying the Euler discretization scheme and the rectangular quadrature formula we obtain the following discretized problem:

$$\min \left\{ \tau \sum_{i=0}^{M-1} l(p_i, x_{Ni}) + h\tau \sum_{k=0}^{N-1} \sum_{i=0}^{M-1} L(p_i, x_{ki}, y_{ki}, u_{ki}) \right\} \tag{19}$$

subject to $x \in \tilde{\mathcal{X}_0}$, $u \in \tilde{\mathcal{U}}$, and the equations

$$y_{ki} = \tau \sum_{j=0}^{M-1} g(p_i, p_j, x_{kj}, u_{kj}), \quad i = 0, \ldots, M-1, \tag{20}$$

$$x_{k+1,i} = x_{ki} + hf(p_i, x_{ki}, y_{ki}, u_{ki}), \quad i = 0, \ldots, M-1, \tag{21}$$

where k runs from 0 to $N-1$. Clearly, given $u \in \tilde{\mathcal{U}}$ the iterative scheme (20), (21) produces unique sequences denoted further by $x[u]$ and $y[u]$. The objective value in (19) resulting from u is denoted by $\tilde{J}(u)$, Clearly, the spaces $\tilde{\mathcal{X}}$, ..., $\tilde{\mathcal{U}}$, the matrices (of vectors) $x[u]$, $y[u]$, and the value $\tilde{J}(u)$ depend on N and M (hence on h and τ) and this dependence will be sometimes made explicit: for example $x^{h\tau}[u]$ and $J^{h\tau}(u)$ may appear in the text to indicate it.

4.2 Gradient projection method for solving the discretized problem

In order to apply a gradient projection method for solving problem (19)–(21) we need to calculate the gradient of $\tilde{J}$. Since $\tilde{J}$ is defined in the subset $\tilde{\mathcal{U}}$ of l_∞, the derivative $\tilde{J}'(u)$, considered as a matrix (of r-dimensional vectors) acts on $\Delta u \in \tilde{\mathcal{U}}$ as

$$\tilde{J}'(u)\Delta u = h\tau \sum_{k=0}^{N-1} \sum_{i=0}^{M-1} [\tilde{J}'(u)]_{ki} \Delta u_{ki}.$$

It is a routine task to represent this derivative by involving the discrete adjoint equations corresponding to (20), (21). For this reason we define the following discrete version of the Hamiltonian H in (8): for $x = (x_0, \ldots, x_{M-1})$, $y = (y_0, \ldots, y_{M-1})$, ..., $\eta = (\eta_0, \ldots, \eta_{M-1})$ and $s = (x, y, u)$, $z = (s, \xi, \eta)$ (again we overload the symbols x, ..., η, since here they denote M-dimensional vectors instead of $(N+1) \times M$ matrices) define

$$\tilde{H}(z) := \tau \sum_{i=0}^{M-1} \left[\xi_i f(p_i, s_i) + L(p_i, s_i) + \tau \eta_i \sum_{j=0}^{M-1} g(p_i, p_j, s_j) \right].$$

The derivatives of $\tilde{H}$ with respect to z have the component-wise representation

$$\begin{aligned} \tilde{H}_s(i, z) &= \tilde{H}_s(i, s_i, \xi_i, \eta) = \xi_i f_s(p_i, s_i) + L_s(p_i, s_i) + \tau \sum_{j=0}^{M-1} \eta_j g_s(p_j, p_i, s_i), \\ \tilde{H}_\xi(i, z) &= \tilde{H}_\xi(i, s_i) = f(p_i, s_i), \qquad \tilde{H}_\eta(i, z) = \tilde{H}_\eta(i, x, u) = \tau \sum_{j=0}^{M-1} g(p_i, p_j, x_j, u_j). \end{aligned}$$

Then it is a routine task to represent

$$[\tilde{J}'(u)]_{ki} = \tilde{H}_u(i, s_{ki}, \xi_{k+1,i}, \eta_{k+1,\cdot}), \tag{22}$$

where now $u \in \tilde{\mathcal{U}}$, the components $x = x[u]$ and $y = y[u]$ of z are determined from (20), (21), and the components ξ and η satisfy for $k = 0, \dots, N-1$ and $i = 0, \dots, M-1$ the equations

$$\begin{aligned}
\xi_{N,i} &= l_x(p_i, x_{Ni}), && (23)\\
\eta_{k+1,i} &= \tilde{H}_y(i, s_{ki}, \xi_{k+1,i}), && (24)\\
\xi_{k,i} &= \xi_{k+1,i} + h\tilde{H}_x(i, s_{ki}, \xi_{k+1,i}, \eta_{k+1,\cdot}), && (25)
\end{aligned}$$

where k runs (backwards) from $N-1$ to 0 and $i = 0, \dots, M-1$. As usual, $z_{k,\cdot}$, etc., denotes the vector (of vectors) $(z_{ki})_{i=1,\dots,M-1}$. Notice also that $\tilde{H}_y$ does not depend on η, therefore the argument η is missing in (24).

Having the representation (22) of the gradient of $\tilde{J}$ we may define the following procedure. Starting from an initial guess $u \in \tilde{\mathcal{U}}$ we calculate the corresponding solution of the initial-value problem (20), (21) and obtain $(x[u], y[u])$. Then we solve (backwards) system (23)–(25), and calculate $\tilde{J}'(u)$ from (22). At this point one can use a variety of methods to calculate a next control $u_{\text{next}} \in \tilde{\mathcal{U}}$ (see e.g. [12]). One possibility is to implement a line search involving the following scalar function:

$$\varphi(\alpha) := \tilde{J}(\mathcal{P}_{\tilde{\mathcal{U}}}(u - \alpha\tilde{J}'(u))),$$

where $\mathcal{P}_{\tilde{\mathcal{U}}}$ is the projection operator on $\tilde{\mathcal{U}}$ with respect to any norm in the finite-dimensional space where u lives. The projection operator $\mathcal{P}_{\tilde{\mathcal{U}}}$ has the comfortable representation

$$[\mathcal{P}_{\tilde{\mathcal{U}}}(v)]_{ki} = \mathcal{P}_U(v_{ki}), \tag{26}$$

where the projection on U is usually easy to calculate. The calculation of $\varphi(\alpha)$ requires to solve again (20), (21) and to calculate the objective value in (19). After making iterations defined by any method for scalar minimization of φ on $[0, \infty)$, we obtain $u_{\text{next}} \in \tilde{\mathcal{U}}$ and continue the procedure in the same manner.

The above procedure can be terminated by any of the usual stopping tests, resulting in some $\tilde{z} \in \tilde{Z}$. One reasonable criterion for the quality of the obtained "approximation" is the residual that it gives in the system of necessary optimality conditions for problem (19)–(21). Clearly, for optimality of $z \in \tilde{\mathcal{Z}}$ it is necessary that $-\tilde{J}'(u) \in N_{\tilde{\mathcal{U}}}(u)$. Having in mind the representation of $\tilde{J}'$ in (22)–(25) and (26) we formulate this necessary optimality condition as: $z \in \mathcal{Z}$ and

$$\begin{aligned}
0 &= -y(t, p_i) + \tilde{H}_\eta(i, x_{k,\cdot}, u_{k,\cdot}), && (27)\\
0 &= -\frac{x_{k+1,i} - x_{ki}}{h} + \tilde{H}_\xi(i, s_{ki}), && (28)\\
0 &= -\xi_{N,i} + l_x(p_i, x_{Ni}), && (29)\\
0 &= -\eta_{k+1,i} + \tilde{H}_y(i, s_{ki}, \xi_{k+1,i}), && (30)\\
0 &= \frac{\xi_{k+1,i} - \xi_{ki}}{h} + \tilde{H}_x(i, s_{ki}, \xi_{k+1,i}, \eta_{k+1,\cdot}), && (31)\\
0 &\in \tilde{H}_u(i, s_{ki}, \xi_{k+1,i}, \eta_{k+1,\cdot}) + N_U(u_{ki}), && (32)
\end{aligned}$$

for $k = N-1, \dots, 0$, $i = 0, \dots, M-1$. The numerically obtained $\tilde{z}$ satisfies the above system with some (calculable) residual $g[\tilde{z}] = \{g[\tilde{z}]_{ki}\}$, so that equations (inclusion) (27)–(32) are satisfied with $g[\tilde{z}]_{ki} = (g[\tilde{y}]_{ki}, g[\tilde{x}]_{ki}, \dots, g[\tilde{u}]_{ki})$ in the left-hand side.

Let us introduce a single number $\tilde{\rho} = \hat{\rho}(\tilde{z})$ that measures the size of the inaccuracy of the gradient procedure:

$$\tilde{\rho} := \max\{|g[\tilde{z}]_{ki}| : \ k = 0, \dots, N-1, \ i = 0, \dots, M-1\}.$$

In fact, if $(\tilde{x}, \tilde{y})$ solves (27), (28) for the control $\tilde{u}$ and then ξ and η solve (29) and (31) for the obtained $\tilde{s}$ (as on every step of the gradient projection procedure), then we have that

$$\tilde{\rho} = \max\{|g[\tilde{u}]_{ki}| : \ k = 0, \dots, N-1, \ i = 0, \dots, M-1\}.$$

Having in mind (22) we obtain that

$$\tilde{\rho} = \mathrm{dist}(-\tilde{J}(\tilde{u}), N_{\tilde{\mathcal{U}}}(\tilde{u})),$$

where the distance is l_∞. Since the gradient projection methods are convergent (in the sense that the above distance converges to zero), only under continuous Frechet differentiability of $\tilde{J}$ and convexity of U, and since $\tilde{\rho}$ is computable, we may assume that it can be made as small as we want.

Ones some $\tilde{z} \in \tilde{\mathcal{Z}}$ is produced by the gradient projection procedure, we define $\hat{u} \in \mathcal{U}$ as $\hat{u}(t,p) = u_{ki}$ for $(t,p) \in Q_{ki}$, $k = 0, \dots, N-1$, $i = 0, \dots, M-1$. This is the result of the overall solution procedure.

What is the relation between $\hat{u}$ and the set of optimal controls of the original problem (1)–(4)? Similar question concerns the corresponding trajectories and adjoint variables. This will be investigated in the next section.

5 Error analysis

The ideology for establishing an error estimate for the proposed numerical procedure is based on the concept of metric regularity of generalized equations such as (18) (see [7] for a comprehensive exposition). We implement it in a way similar to that in our earlier contribution [8] (dealing with ODE control problems). We first embed the approximate discrete solution $\tilde{z} = \tilde{z}^{h\tau}$ in the continuous-time space $\mathcal{Z}$, then we estimate the corresponding residual in (18) and use a metric regularity assumption for F to obtain an error estimate.

The control $\tilde{u}$ obtained by the numerical procedure in the previous section was embedded in $\mathcal{U}$ as the piece-wise constant function $\hat{u}$ (constant on every box Q_{ki}). In a similar way we proceed with $\tilde{y}$ and $\tilde{\eta}$: $\hat{y}(t,p) := \tilde{y}_{ki}$ and $\hat{\eta}(t,p) := \tilde{\eta}_{k+1,i}$ for $(t,p) \in Q_{ki}$. For x and ξ we define for $(t,p) \in Q_{ki}$

$$\hat{x}(t,p) := \tilde{x}_{ki} + \frac{t - t_k}{h}(\tilde{x}_{k+1,i} - \tilde{x}_{ki}), \quad \hat{\xi}(t,p) := \tilde{\xi}_{ki} + \frac{t - t_k}{h}(\tilde{\xi}_{k+1,i} - \tilde{\xi}_{ki}).$$

Due to Standing Assumptions, it is easy to verify that both $\|\tilde{z}\|$ and $\|\hat{z}\|$ are bounded, uniformly in N and M. Then easy calculations (involving the Lipschitz constants of f, g, L, l and their first derivatives) show that there exists a constant C_0, independent of N and M, such that

$$(33) \qquad \hat{g} \in F(\hat{z}), \qquad \text{with } \|\hat{g}\|_{\mathcal{G}} \le C_0(h + \tau + \tilde{\rho}).$$

We present the calculation for y and x, where $O(\varepsilon)$ will denote any measurable and bounded function on D such that $\|O(\varepsilon)\|_{L_\infty} \le c\varepsilon$ and the constant c is independent of N, M, and the particular approximate solution $\tilde{u}$ obtained by the gradient projection procedure. We have $|\hat{x}(t,p) - \tilde{x}_{ki}| = O(h+\tau)$ on Q_{ki}, thus for $(t,p) \in Q_{ki}$

$$\begin{aligned}
&\int_P g(p,q,\hat{x}(t,q),\hat{u}(t,q))\,\mathrm{d}q = \sum_{j=0}^{M-1} \int_{p_j}^{p_{j+1}} g(p,q,\tilde{x}_{kj},\tilde{u}_{kj})\,\mathrm{d}q + O(h) \\
&= \sum_{j=0}^{M-1} \int_{p_j}^{p_{j+1}} g(p_i,p_j,\tilde{x}_{kj},\tilde{u}_{kj})\,\mathrm{d}q + O(h+\tau) = \tilde{y}_{ki} + O(h+\tau+\tilde{\rho}) \\
&= \hat{y}(t,p) + O(h+\tau+\tilde{\rho}).
\end{aligned}$$

Similarly, for $(t,p) \in Q_{ki}$ we have

$$\begin{aligned}
\dot{\hat{x}}(t,p) &= \frac{\tilde{x}_{k+1,i} - \tilde{x}_{ki}}{h} = f(p_i,\tilde{s}_{ki}) + O(\tilde{\rho}) = f(p_i,\tilde{x}_{ki},\hat{y}(t,p),\hat{u}(t,p)) + O(\tilde{\rho}) \\
&= f(p,\hat{s}(t,p)) + O(h+\tau+\tilde{\rho}).
\end{aligned}$$

For the residual in (17) we obtain for $(t,p) \in Q_{ki}$

$$\begin{aligned}
\operatorname{dist}(H_u(p,\hat{z}(t,p)), N_U(\hat{u}(t,p))) &= \operatorname{dist}(H_u(p,\hat{z}(t,p)), N_U(\tilde{u}_{ki})) \\
&\le \operatorname{dist}(H_u(p,\tilde{z}_{ki}), N_U(\tilde{u}_{ki})) + O(h+\tau+\tilde{\rho}) \\
&= O(h+\tau+\tilde{\rho}),
\end{aligned}$$

where we substantially used that $\hat{u}$ is constant in every box Q_{ki}.

Having at hand (33) we have to establish an estimation for the distance in $\mathcal{Z}$ from $\hat{z}$ to some solution $z \in \mathcal{Z}$ of the inclusion $0 \in F(z)$ (see (18)). The instrument for that is briefly presented below based on [7, Chapter 3].

Below we denote by d the distance generated by the norm either in $\mathcal{Z}$ or in $\mathcal{G}$. We use the same symbol for the distance from point to set: $d(p,Q) = \inf_{q\in Q} d(p,q)$. As usual $F^{-1}(g) := \{z \in \mathcal{Z} : g \in F(z)\}$.

Definition 1 The set-valued mapping $F : Z \rightrightarrows \mathcal{G}$ is said to be *metrically regular* (MR) at z^* for 0 if $0 \in F(z^*)$ and there is a constant $\kappa \ge 0$ together with neighborhoods A of z^* and B of 0 such that

$$d(z, F^{-1}(g)) \le \kappa d(g, F(z)) \text{ for all } (z,g) \in A \times B.$$

The infimum of κ over all combinations of κ, A and B for which the above relation is satisfied is called the *regularity modulus* for F at z^* for 0 and denoted by $\operatorname{reg}(F; z^* \,|\, 0)$. In absence of metric regularity we set $\operatorname{reg}(F; z^* \,|\, 0) = \infty$.

Definition 2 The set-valued mapping $F : Z \rightrightarrows \mathcal{G}$ is said to be *strongly metrically regular* (SMR) at z^* for 0 if it is metrically regular at z^* for 0 with neighborhoods A, B and constant κ such that $F^{-1}(g) \cap A$ is single valued when $g \in B$.

The last definition means that $g \mapsto F^{-1}(g) \cap A$ is a Lipschitz function on B with Lipschitz constant κ.

Denote by $\mathcal{Z}^*$ the set of solutions $z^* \in \mathcal{Z}$ of inclusion (18), that is, $\mathcal{Z}^* = F^{-1}(0)$.

Clearly, the triple (κ, A, B) in the definition of MR is not unique. If F is MR at z (further we skip "for 0"), then for every $\kappa > \operatorname{reg}(F; z\,|\,0)$ there is a pair of open sets $A(z,\kappa) \ni z$ and $B(z,\kappa) \ni 0$ such that the definition of MR is fulfilled with the triple $(\kappa, A(z,\kappa), B(z,\kappa))$. Presumably, the smaller is κ, the smaller must be these neighborhoods. Given $\kappa > 0$, we fix such pairs of neighborhoods for every $z \in \mathcal{Z}$ for which $\operatorname{reg}(F; z\,|\,0) < \kappa$, else we set $A(z,\kappa) = \emptyset$, $B(z,\kappa) = \emptyset$. In particular, these sets are empty if $0 \notin F(z)$ (in that case $\operatorname{reg}(F; z\,|\,0) = +\infty$).

We do the same for the points z at which F is SMR: for every $z \in \mathcal{Z}$ we fix open sets $A^s(z,\kappa) \ni z$, $B^s(z,\kappa) \ni 0$ (possibly empty) such that $F^{-1} \cap A^s(z,\kappa)$ is single-valued and Lipschitz with constant κ on the set $B^s(z,\kappa)$. We shall use also the notation $A^s_{1/2}(z^*,\kappa) := \{z \in \mathcal{Z} : \|z - z^*\| \le d(z, \mathcal{Z} \setminus A^s(z^*,\kappa))\}$, which is nonempty whenever $A^s(z^*,\kappa)$ is nonempty.

We remind that above we have in mind the mapping F in (18), representing the Pontryagin maximum principle for our original problem.

Theorem 1 *There exists a constant C such that for every $\kappa > 0$, every natural numbers N and M, and every $\tilde{z} \in \tilde{\mathcal{Z}}$ obtained by the numerical procedure in Section 4 with discretization mesh size (N, M) and residual $\tilde{g}$, the following claims hold for the embedding $\hat{z} \in \mathcal{Z}$ of $\tilde{z}$ and the residual $\hat{g}$ in (33):*

(i) If $(\hat{z}, \hat{g}) \in A(z^*,\kappa) \times B(z^*,\kappa)$ for some $z^* \in \mathcal{Z}^*$, then

$$d(\hat{z}, Z^*) \le C\kappa\,(h + \tau + \tilde{\rho});$$

(ii) If $\hat{z} \in A^s_{1/2}(z^*,\kappa)$ and $\hat{g} \in B^s(z^*,\kappa)$ for some $z^* \in \mathcal{Z}^*$, then

$$\|\hat{z} - z^*\|_{\mathcal{Z}} \le C\kappa\,(h + \tau + \tilde{\rho}).$$

Proof. If the assumption in Claim (i) is fulfilled, then

$$d(\hat{z}, \mathcal{Z}^*) = d(\hat{z}, F^{-1}(0)) \le \kappa\, d(0, F(\hat{z})) \le \kappa\, \|\hat{g}\|_{\mathcal{G}} \le \kappa C_0 (h + \tau + \tilde{\rho}),$$

due to (33).

Let us prove Claim (ii). According to the definition of $A^s_{1/2}(z^*,\kappa)$ we have

$$\begin{aligned} d(\hat{z}, Z^*) &= \min\left\{ \inf_{z \in Z^* \cap A^s(z^*,\kappa)} \|z - \hat{z}\|,\ \inf_{z \in Z^* \setminus A^s(z^*,\kappa)} \|z - \hat{z}\| \right\} \\ &= \min\left\{ \inf_{z \in F^{-1}(0) \cap A^s(z^*,\kappa)} \|z - \hat{z}\|,\ \inf_{z \in Z^* \setminus A^s(z^*,\kappa)} \|z - \hat{z}\| \right\} \\ &= \min\left\{ \|z^* - \hat{z}\|,\ \inf_{z \in Z^* \setminus A^s(z^*,\kappa)} \|z - \hat{z}\| \right\} \\ &= \|z^* - \hat{z}\|, \end{aligned}$$

where we use that $F^{-1}(0) \cap A^s(z^*,\kappa) = \{z^*\}$. Then we repeat the argument in the proof of Claim (i). Q.E.D.

The assumption in Theorem 1 that the residual $\hat{g}$ of the numerically obtained "approximate solution" $\hat{z}$ belongs to the "region of Lipschitz stability", $B(z^*,\kappa)$, of some solution $z^* \in \mathcal{Z}^*$ of

(18) needs some comments. We do not assume even existence of a solution of problem (1)–(4), so $\mathcal{Z}^*$ may be empty. Even if $\mathcal{Z}^*$ is nonempty, the limit of (a subsequence of) $\hat{z} = \hat{z}^{h\tau}$ when h, τ converge to zero (in a topology in which the sequence $z^{h\tau}$ is precompact) the limit does not need to be a solution of (18). To ensure the latter, one needs additional conditions that we do not discuss further, since the issue is well known and similar in the ODE control context. We only mention that the assumption in Theorem 1, that $(\hat{z}^{h\tau}, \hat{g}^{h\tau}) \in A^s_{1/2}(z^*, \kappa) \times B^s(z^*, \kappa)$ for some $z^* \in \mathcal{Z}^*$ is demanding in another aspect. If this assumption is fulfilled with a fixed z^* for h, τ and $\tilde{\rho}$ converging to zero (along some sequence), then the estimation in part (ii) of the theorem implies that

$$\|\hat{u} - u^*\|_{L_\infty(D)} \le C\kappa\,(h + \tau + \tilde{\rho})$$

along this sequence. Taking into account that $\hat{u}$ is constant on every box Q_{ki}. Essentially, this implies that u^* must be (equivalent to) a Lipschitz continuous function. In several applications (including reasonable modifications of the epidemiological model in Section 2) all optimal controls may happen to be discontinuous. Such cases are not covered in the present paper. The error analysis in case of discontinuous optimal controls is currently under investigation even for simple classes of ODE optimal control problems (see e.g. [16, 10, 1, 2, 13]).

A question that remains open so far, is whether the generalized equation (18) has solutions $z^* \in Z^*$ at which F is MR (SMR). A necessary and sufficient condition for SMR of F at a given point z^* is given in [6] in the case of an ODE optimal control problem (where P consists of a single point and the aggregated state y is not present). The extension of this result to the problem (1)–(4) is not straightforward, although the same approach, based on the stability of the SMR property with respect to linearization ([7, Theorem 5E.1]), is still applicable. Although it is hard to check the sufficient conditions for SMR in advance, this is doable a posteriori.

6 Some extensions

We have already noticed that explicit dependence on the time t of all data involved in problem (1)–(4) does not bring any difficulty, assuming Lipschitz continuity in t.

In many applications the function g in (2) is independent of p, in which case $y(t,p) = y(t)$ is also independent of p (see the example in Section 2). Clearly, this brings only a simplification.

A version of problem (1)–(4) may involve non-distributed control variables $v(t) \in V$. This case needs small changes in the optimality conditions (12)–(17), namely, in the last one. To avoid additional notations, assume that all control variables are independent of p, that is, $u(t,p) = u(t)$. Then $\mathcal{U} = \{u \in L_\infty([0,T]) : u(t) \in U \text{ for a.e. } t\}$ and (12) changes to

$$0 \in \int_P H_u(p, z(t,\cdot))\,\mathrm{d}p + N_U(u(t)).$$

Also the second index i of u_{ki} should be deleted in the discrete considerations in sections 4 and 5, and (32) changes to

$$0 \in \sum_{i=0}^{M-1} \tilde{H}_u(i, s_{ki}, \xi_{k+1,i}, \eta_{k+1,\cdot}) + N_U(u_k).$$

7 Final discussions

As mentioned above, in practice we use second order Runge-Kutta schemes for discretization of the differential equations and integrals. One reason for which we present only an approximation procedure based on the Euler scheme, is to avoid technicalities. Another reason is, that even when second order schemes are used, the error estimation technique employed in this paper faces difficulties in obtaining second order estimations. Such are obtained for ODE problems in [5], where, however, the error estimation is based on assumptions implying (uniform in the discretization step) strong metric regularity of the discrete problem. To author's knowledge, it is an open question whether the latter follows only from SMR of the original problem. The technique from [5] can probably be extended for obtaining second order accuracy also for the distributed problems considered in the present paper, but this is a technically challenging issue.

On the other hand, applying higher than second order schemes to problems with control constraints does not bring higher order accuracy, in general, due to the discontinuity of the first derivative of the control function that usually appears at contact points with the boundary of the constraints.

References

[1] W. Alt, R. Baier, M. Gerdts, F. Lempio. Error bounds for Euler approximations of linear-quadratic control problems with bang-bang solutions. *Numerical Algebra, Control and Optimization*, **2**(3):547-570, 2012.

[2] W. Alt, R. Baier, F. Lempio, M. Gerdts. Approximations of linear control problems with bang-bang solutions. *Optimization*, **62**(1):9–32, 2013.

[3] S. Aniţa. *Analysis and Control of Age-Dependent Population Dynamics.* Mathematical Modelling Series. Springer, 2000.

[4] O. Diekmann, J.A.P. Heesterbeek. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation.* Wiley, Chichester, 2000.

[5] A.L. Dontchev, W.W. Hager, V. M. Veliov. Second order Runge-Kutta approximations in control constrained optimal control. *SIAM J. Numer. Anal.* **38**(1):202–226, 2000.

[6] A.L Dontchev and K. Malanowski. A characterization of Lipschitz stability in optimal control. In *Calculus of variations and optimal control: Technion, 1998*, A. Ioffe, S. Reich, I Shafrir, Edts., pp. 62–76, Chapman and Hall, 1999.

[7] A. K. Dontchev, R. T. Rockafellar. *Implicit Functions and Solution Mappings.* Second Edition, Springer 2014.

[8] A.L Dontchev and V.M. Veliov. Metric regularity under approximations. *Control and Cybernetics*, **38**(4):1283–1303 , 2009.

[9] G. Feichtinger, G. Tragler, and V.M. Veliov. Optimality conditions for age-structured control systems. *J. Math. Anal. Appl.*, **288**(1):47–68, 2003.

[10] U. Felgenhauer. On stability of bang-bang type controls. *SIAM J. of Control and Optimization*, **41**(6):1843-1867, 2003.

[11] R.I. Hickson, M.G. Roberts. How population heterogeneity in susceptibility and infectivity influences epidemic dynamics, *J. Theor. Biol.*, **350**:70–80, 2014.

[12] E. Polak. *Computational Methods in Optimization. A unified approach.* Academic Press, 1971.

[13] M. Quincampoix and V.M. Veliov. Metric regularity and stability of optimal control problems for linear systems. *SIAM J. Contr. Optim.*, **51**(5):4118-4137, 2013.

[14] V.M. Veliov. Optimal Control of Heterogeneous Systems: Basic Theory. *J. Math. Anal. Appl.*, **346**:227–242, 2008.

[15] V.M. Veliov. Newton's method for problems of optimal control of heterogeneous systems. *Optimization Methods and Software*, **18**(6):689–703, 2003.

[16] V.M. Veliov. Error analysis of discrete Approximation to bang bang optimal control problems: the linear case. *Control&Cybernetics*, **34**(3):967-982, 2005.