Numerical treatment of a class of optimal control problems arising in economics

Etienne Emmrich[†] **and Horst Schmitt**[‡]

Version April 11, 2005

Abstract The approximate solution of the problem of controlling an initial value problem for a linear system of autonomous ordinary differential equations is considered. The corresponding homogeneous solution to the differential equation is assumed to be non-expansive and the inhomogeneity is a linear function of the control variable that is constant along *a priori* given sub-intervals. The optimal control minimises a convex functional that depends, possibly in a nonlinear way, on the solution of the differential equation. Infinite time horizons are allowed.

In view of the piecewise constant control, the corresponding Lagrangian can be split into the sum of Lagrangians acting on sub-intervals. The two algorithms suggested are based upon an iterative process that takes advantage of this splitting as well as of the explicit solution to the differential constraints.

Convergence results are provided under suitable assumptions on the problem's data. Finally, numerical tests for a model of global warming demonstrate the performance of the algorithms.

Keywords Optimal control, ordinary differential equation, iteration, convergence, global warming

MSC (2000) 65K05, 91B76, 34H05, 49M05

1 Introduction

Time continuous discounted control problems with infinite time horizon of the form

$$\min_{u\in U}\int_0^\infty \mathrm{e}^{-rt} f(\boldsymbol{x}(t),u(t))\,dt \quad \text{such that} \quad \dot{\boldsymbol{x}} = \boldsymbol{g}(\boldsymbol{x},u)\,,\ \boldsymbol{x}(0)=\boldsymbol{x_0}\,, \qquad (1.1)$$

are usually considered with respect to the space U of measurable or piecewise continuous control functions $u\colon \mathbb{R}_0^+ \to \Omega \subseteq \mathbb{R}$. However, the control function might also be vector-valued (cf. [7] for more details). Here, f and $\boldsymbol{g}$ are given functions, $\boldsymbol{x}=\boldsymbol{x}(t)$ is the time-dependent state with prescribed

[†] Technische Universität Berlin, Institut für Mathematik, Straße des 17. Juni 136, 10623 Berlin, Germany, eMail: emmrich@math.tu-berlin.de

[‡] AOK Sachsen-Anhalt, Lüneburger Straße 4, 39106 Magdeburg, Germany, eMail: horst.schmitt@san.aok.de

initial state $\boldsymbol{x}_0$, and $r > 0$ is the discount rate. Reducing U to the linear space of piecewise constant functions $u = u(t)$ with $u(t) \equiv u_i \in \Omega \subseteq \mathbb{R}$ on *a priori* given time intervals $[t_i, t_{i+1})$ $(i = 0, \dots, N)$ leads to discrete problems of the type

$$\min_{u_i \in \Omega} \sum_{i=0}^{N} f_i(\boldsymbol{x}_i, u_i) \quad \text{such that} \quad \boldsymbol{x}_{i+1} = \widetilde{\boldsymbol{g}}_i(\boldsymbol{x}_i, u_i) \quad (i = 0, \dots, N-1) \tag{1.2}$$

with f_i, $\widetilde{\boldsymbol{g}}_i$, $\boldsymbol{x}_0$ given, where $N \in \mathbb{N} \cup \{\infty\}$ might be finite or infinite. Such problems arise in particular when solving the continuous problem (1.1) numerically by discretisation in time (cf. [2], [3]).

On the other hand, in many applications, a discrete model seems *a priori* to be more appropriate than a time continuous one: As for economical problems in general prices, which are normally constant along some time intervals, have to be controlled, it is natural to assume the control variable to be piecewise constant. So economic decisions as for instance price fixing or dividend distribution are taken at fixed discrete points in time (daily, weekly, yearly ...). Furthermore, an infinite time horizon seems to be typical.

A typical example for such a problem is the Nordhaus model of global warming as proposed in [10], which will be considered in Section 5. The aim is, loosely spoken, to minimise the additional costs arising from the greenhouse effect by controlling the energy prices. The dynamics of the greenhouse effect is described by an initial-value problem for a system of ordinary differential equations, and the objective function incorporates the discounted welfare function and costs.

Solving problems of the type (1.1) or (1.2) numerically can be based upon the corresponding discrete Hamilton-Jacobi-Bellman equation (cf. [5])

$$V_\tau(\boldsymbol{x}) = \sup_{u \in U} \{(1 - r\tau) V_\tau(\boldsymbol{\phi}_\tau(\boldsymbol{x}, u)) - \tau f(\boldsymbol{x}, u)\},$$

where τ denotes the constant time step size, $V_\tau(\boldsymbol{x})$ is the optimal value of the functional depending on the initial state $\boldsymbol{x}_0 = \boldsymbol{x}$, $1 - r\tau$ is an approximation for $\mathrm{e}^{-r\tau}$ with the discount rate r, $\boldsymbol{\phi}_\tau(\boldsymbol{x}, u)$ is the state at point $t = \tau$ with the initial state $\boldsymbol{x}$ after controlling the system by u, and f is the cost function.

However, this equation is not discrete in the state variable and solving it requires appropriate (adaptive) grid schemes. The complexity thus increases considerably with the dimension of $\boldsymbol{x}$. Furthermore, non-autonomous problems or non-equidistant time partitions need additional considerations.

In this paper, we develop an iterative numerical scheme to construct approximate solutions using piecewise constant control functions. The algorithms presented are successfully tested for the above-mentioned model of global warming.

Let a finite partition of the (finite or infinite) time interval $[0, t_{N+1})$ ($t_{N+1} \in \mathbb{R} \cup \{\infty\}$, $N \in \mathbb{N}$) be given via

$$0 = t_0 < t_1 < \cdots < t_N < t_{N+1}\,, \quad \tau_i := t_{i+1} - t_i\,, \quad \tau_{\max} := \max_{i=0,\dots,N} \tau_i\,.$$

The space U of control functions then is assumed to consist of functions that are constant on each sub-interval $[t_i, t_{i+1})$ ($i = 0, \dots, N$). This setup leads to a control problem of the type (1.2) with finite N and might be considered as a particular nonlinear optimisation problem (cf. [9]). The states are assumed to be time-dependent with values in $\mathbb{R}^d$ ($d \in \mathbb{N}$). Throughout this paper, elements of $\mathbb{R}^d$ will be always column vectors and typed boldfaced.

We consider

Problem 1.1 *For given α, $r > 0$, a convex and twice continuously differentiable function $z : \mathbb{R}^d \to \mathbb{R}$, and the initial state $\boldsymbol{x}_0 \in \mathbb{R}^d$, find control variables $u_i^* \in \mathbb{R}$ ($i = 0\,, \dots,\, N$) minimising the functional*

$$J(u_0, \dots, u_N) := \sum_{i=0}^{N} \int_{t_i}^{t_{i+1}} \mathrm{e}^{-rt} \left(\frac{\alpha}{2}\, u_i^2 + z\left(\boldsymbol{\phi}(t; t_i, \boldsymbol{x}_i, u_i)\right) \right) dt\,, \tag{1.3}$$

where

$$\boldsymbol{x}_{i+1} = \boldsymbol{\phi}(t_{i+1}; t_i, \boldsymbol{x}_i, u_i)\,, \quad i = 0\,, \dots,\, N-1\,. \tag{1.4}$$

Here, $\boldsymbol{\phi}(\cdot; s, \boldsymbol{y}, v) : [s, \infty) \to \mathbb{R}^d$ denotes, for given $s \in [0, \infty)$, $\boldsymbol{y} \in \mathbb{R}^d$, and $v \in \mathbb{R}$, the solution to the linear, non-homogeneous, autonomous initial value problem

$$\dot{\boldsymbol{\phi}}(t) = A\boldsymbol{\phi}(t) + \boldsymbol{a}v + \boldsymbol{b} \quad (t > s)\,, \quad \boldsymbol{\phi}(s) = \boldsymbol{y}\,, \tag{1.5}$$

where $A \in \mathbb{R}^{d \times d}$, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$. With Duhamel's principle, we have

$$\boldsymbol{\phi}(t; s, \boldsymbol{y}, v) = \mathrm{e}^{(t-s)A}\boldsymbol{y} + \int_s^t \mathrm{e}^{(t-\sigma)A} d\sigma\, (\boldsymbol{a}v + \boldsymbol{b})\,.$$

Obviously, $\boldsymbol{\phi}(\cdot; s, \boldsymbol{y}, v)$ is a smooth function in all its arguments. Moreover, it holds

$$\mathrm{D}_{\boldsymbol{y}}\boldsymbol{\phi}(t; s, \boldsymbol{y}, v) = \mathrm{e}^{(t-s)A}\,, \quad \mathrm{D}_v\boldsymbol{\phi}(t; s, \boldsymbol{y}, v) = \int_s^t \mathrm{e}^{(t-\sigma)A} d\sigma\, \boldsymbol{a}\,, \tag{1.6}$$
$$\mathrm{D}_{vv}\boldsymbol{\phi}(t; s, \boldsymbol{y}, v) = 0\,, \quad \mathrm{D}_{\boldsymbol{y}v}\boldsymbol{\phi}(t; s, \boldsymbol{y}, v) = 0\,,$$

where D_v denotes the first derivative with respect to v, $\mathrm{D}_{\boldsymbol{y}}$ denotes the gradient (which is always thought to be a row vector) with respect to $\boldsymbol{y}$, and thus $\mathrm{D}_{\boldsymbol{y}}\boldsymbol{\phi}$ is the Jacobian. Furthermore, it is $\mathrm{D}_{vv} \equiv \mathrm{D}_v\mathrm{D}_v$, $\mathrm{D}_{\boldsymbol{y}v} \equiv \mathrm{D}_{\boldsymbol{y}}\mathrm{D}_v$.

It immediately follows from (1.3) and the properties of $\boldsymbol{\phi}$ that the Hessian of J is a diagonal matrix with the diagonal entries

$$\mathrm{D}_{u_k u_k} J(u_0, \ldots, u_N) = \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} \left(\alpha + \right.$$

$$\left. + \, \mathrm{D}_{u_k}\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k)^{\mathsf{T}} \mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z \left(\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k)\right) \mathrm{D}_{u_k}\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k)\right) dt \,.$$

As the smooth function z is assumed to be convex, its Hessian matrix $\mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z$ is positive semi-definite. Since $\alpha > 0$, this shows that J is strongly convex.

We shall make the following structural assumptions:

(A1) There are constants $c \geq 1$, $\lambda \geq 0$ such that

$$\|\mathrm{e}^{tA}\| \leq c\,\mathrm{e}^{-\lambda t} \qquad \forall t \geq 0 \,.$$

(A2) For every $R > 0$ there is a constant $\kappa(R) > 0$ such that

$$\|\mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z(\boldsymbol{\phi})\| \leq \kappa(R) \qquad \forall \boldsymbol{\phi} \in \mathbb{R}^d \,, \|\boldsymbol{\phi}\| \leq R \,.$$

(A3) If $\lambda = 0$ in (A1) then $t_{N+1} < \infty$.

Here, $\|\cdot\|$ denotes the Euclidian and spectral norm, respectively.

Note that (A1) is fulfilled if all eigenvalues of A have non-positive real part and if purely imaginary eigenvalues are simple. The constant c is given by $c = \|P\|\,\|P^{-1}\|$ where P transforms A into Jordan's normal form. If the matrix A is normal ($A^{\mathsf{T}}A = AA^{\mathsf{T}}$) then $c = 1$ (cf. [1]).

The restriction to a finite time interval in the case $\lambda = 0$ (Assumption (A3)) is not necessary but simplifies the analysis. However, the case $\lambda = 0$ and infinite time t_{N+1} requires a more intrusive assumption than (A2) that may lead to additional restrictions on the problem's data. We consider this case in more detail in Section 4, where we replace (A2) and (A3) by

(A4) There is some $K > 0$ such that for $i = 0, \ldots, N$ and arbitrary $\boldsymbol{y} \in \mathbb{R}^d$ and $v \in \mathbb{R}$

$$\int_{t_i}^{t_{i+1}} (t - t_i)^q \mathrm{e}^{-rt} \left\|\mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z \left(\boldsymbol{\phi}(t; t_i, \boldsymbol{y}, v)\right)\right\| dt \leq K \,, \quad q \in \{1, 2\} \,, \tag{1.7}$$

$$\int_{t_i}^{t_{i+1}} t\mathrm{e}^{-rt} \left\|\mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z \left(\boldsymbol{\phi}(t; t_i, \boldsymbol{y}, v)\right)\right\| dt \leq K \,. \tag{1.8}$$

Optimal control problems of that type often arise in economics. As examples, we may consider the taxation of carbon dioxide emissions in the context of the greenhouse effect (cf. [10] and Section 5) and the optimisation of the health care expenditure in the last months of life (cf. [4]).

With the multipliers $\boldsymbol{p}_i \in \mathbb{R}^d$ ($i = 0, \ldots, N$), where $\boldsymbol{p}_N := 0$, the Lagrangian corresponding to Problem 1.1 reads as

$$\mathcal{L}(u_0, \ldots, u_N, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_N, \boldsymbol{p}_0, \ldots, \boldsymbol{p}_N) := \sum_{i=0}^{N} \mathcal{L}_i(u_i, \boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \boldsymbol{p}_i) \tag{1.9a}$$

where

$$\begin{aligned}\mathcal{L}_i(u_i, \boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \boldsymbol{p}_i) :=& \int_{t_i}^{t_{i+1}} \mathrm{e}^{-rt} \left(\frac{\alpha}{2}\, u_i^2 + z\left(\boldsymbol{\phi}(t; t_i, \boldsymbol{x}_i, u_i)\right)\right) dt \\ &+ \boldsymbol{p}_i^{\mathsf{T}} \left(\boldsymbol{\phi}(t_{i+1}; t_i, \boldsymbol{x}_i, u_i) - \boldsymbol{x}_{i+1}\right), \; i = 0, \ldots, N-1\,, \qquad (1.9\text{b}) \\ \mathcal{L}_N(u_N, \boldsymbol{x}_N) :=& \int_{t_N}^{t_{N+1}} \mathrm{e}^{-rt} \left(\frac{\alpha}{2}\, u_i^2 + z\left(\boldsymbol{\phi}(t; t_i, \boldsymbol{x}_i, u_i)\right)\right) dt\,. \qquad (1.9\text{c})\end{aligned}$$

For brevity, we omit the arguments of $\mathcal{L}$ in the following. In view of the strong convexity of J, Problem 1.1 is equivalent to the system of first order conditions

$$\mathrm{D}_{u_k}\mathcal{L} = 0\,, \quad \mathrm{D}_{\boldsymbol{x}_k}\mathcal{L} = 0\,, \quad \mathrm{D}_{\boldsymbol{p}_k}\mathcal{L} = 0 \quad (k = 0, \ldots, N)\,.$$

With (1.9) and (1.6), we have

$$\begin{aligned}\mathrm{D}_{u_k}\mathcal{L} = \mathrm{D}_{u_k}\mathcal{L}_k(u_k, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}, \boldsymbol{p}_k) &= \frac{\alpha}{r}\, \mathrm{e}^{-rt_k} \left(1 - \mathrm{e}^{-r\tau_k}\right) u_k \\ &+ \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} \mathrm{D}_{\boldsymbol{\phi}} z\left(\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k)\right) \int_{t_k}^{t} \mathrm{e}^{(t-\sigma)A} d\sigma dt\, \boldsymbol{a} \\ &+ \boldsymbol{p}_k^{\mathsf{T}} \int_{t_k}^{t_{k+1}} \mathrm{e}^{(t_{k+1}-\sigma)A} d\sigma\, \boldsymbol{a}\,, \end{aligned} \tag{1.10}$$

$$\begin{aligned}\mathrm{D}_{\boldsymbol{x}_k}\mathcal{L} &= \mathrm{D}_{\boldsymbol{x}_k}\mathcal{L}_{k-1}(u_{k-1}, \boldsymbol{x}_{k-1}, \boldsymbol{x}_k, \boldsymbol{p}_{k-1}) + \mathrm{D}_{\boldsymbol{x}_k}\mathcal{L}_k(u_k, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}, \boldsymbol{p}_k) \\ &= -\boldsymbol{p}_{k-1}^{\mathsf{T}} + \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} \mathrm{D}_{\boldsymbol{\phi}} z\left(\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k)\right) \mathrm{e}^{(t-t_k)A} dt + \boldsymbol{p}_k^{\mathsf{T}}\, \mathrm{e}^{\tau_k A}\,, \end{aligned} \tag{1.11}$$

$$\mathrm{D}_{\boldsymbol{p}_k}\mathcal{L} = \mathrm{D}_{\boldsymbol{p}_k}\mathcal{L}_k(u_k, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}, \boldsymbol{p}_k) = \left(\boldsymbol{\phi}(t_{k+1}; t_k, \boldsymbol{x}_k, u_k) - \boldsymbol{x}_{k+1}\right)^{\mathsf{T}}\,. \tag{1.12}$$

Note that $\mathcal{L}_k(u_k, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}, \boldsymbol{p}_k)$ depends on $\boldsymbol{x}_{k+1}$ but $\mathrm{D}_{u_k}\mathcal{L}_k(u_k, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}, \boldsymbol{p}_k)$ does not. So we can omit the argument $\boldsymbol{x}_{k+1}$ in $\mathrm{D}_{u_k}\mathcal{L}_k$. Moreover, it is $\mathrm{D}_{\boldsymbol{p}_N}\mathcal{L} \equiv 0$.

We suggest the following iterative process for constructing an approximate solution to Problem 1.1:

Algorithm 1.1

step 0) *Let* $u_0^{(0)}, \dots, u_N^{(0)}$ *be arbitrarily given.*

step ℓ) *(*$\ell = 1, 2, \dots$*)* *For* $k = 0, 1, \dots, N$*, solve*

$$\mathrm{D}_{u_k}\mathcal{L}_k(u_k^{(\ell)}, \boldsymbol{x}_k, \boldsymbol{p}_k) = 0\,. \tag{1.13}$$

Here, $\boldsymbol{x}_k, \boldsymbol{p}_k$ *are to be computed by*

$$\boldsymbol{x}_{j+1} = \boldsymbol{\phi}(t_{j+1}; t_j, \boldsymbol{x}_j, u_j^{(\ell-1)})\,, \quad j = 0, \dots, N-1\,, \tag{1.14}$$

$\boldsymbol{x}_0$ *being the initial state of Problem 1.1,*

$$\boldsymbol{p}_{j-1}^{\mathsf{T}} = \int_{t_j}^{t_{j+1}} \mathrm{e}^{-rt}\mathrm{D}_{\boldsymbol{\phi}} z\left(\boldsymbol{\phi}(t; t_j, \boldsymbol{x}_j, u_j^{(\ell-1)})\right) \mathrm{e}^{(t-t_j)A} dt + \boldsymbol{p}_j^{\mathsf{T}}\,\mathrm{e}^{\tau_j A}\,,$$
$$j = N, \dots, k+1\,, \tag{1.15}$$

and $\boldsymbol{p}_N = 0$*.*

Note that (1.14) corresponds to (1.12) whereas (1.15) comes from (1.11). The derivative in (1.13) is given by (1.10). For solving (1.13), we only need the state $\boldsymbol{x}_k$, but for computing this state, we also need $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{k-1}$. Moreover, we have to compute $\boldsymbol{p}_k$, which makes it necessary to have $\boldsymbol{p}_N, \dots, \boldsymbol{p}_{k+1}$. Therefore, we need the states $\boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_N$, too.

Besides, we consider a slightly changed version:

Algorithm 1.2

step 0) *Let* $u_0^{(0)}, \dots, u_N^{(0)}$ *be arbitrarily given.*

step ℓ) *(*$\ell = 1, 2, \dots$*)* *For* $k = 0, 1, \dots, N$*, solve (1.13), where* $\boldsymbol{x}_k, \boldsymbol{p}_k$ *are to be computed by*

$$\boldsymbol{x}_{j+1} = \boldsymbol{\phi}(t_{j+1}; t_j, \boldsymbol{x}_j, u_j^{(\ell)})\,, \quad j = 0, \dots, k-1\,, \tag{1.16a}$$

$$\boldsymbol{x}_{j+1} = \boldsymbol{\phi}(t_{j+1}; t_j, \boldsymbol{x}_j, u_j^{(\ell-1)})\,, \quad j = k, \dots, N-1\,, \tag{1.16b}$$

$\boldsymbol{x}_0$ *being the initial state of Problem 1.1,*

$$\boldsymbol{p}_{j-1}^{\mathsf{T}} = \int_{t_j}^{t_{j+1}} \mathrm{e}^{-rt} \mathrm{D}_{\boldsymbol{\phi}} z \left(\boldsymbol{\phi}(t; t_j, \boldsymbol{x}_j, u_j^{(\ell-1)}) \right) \mathrm{e}^{(t-t_j)A} dt + \boldsymbol{p}_j^{\mathsf{T}} \, \mathrm{e}^{\tau_j A} \,, \qquad j = N, \, \ldots, \, k+1 \,, \tag{1.17}$$

and $\boldsymbol{p}_N = 0$.

The difference between Algorithm 1.1 and Algorithm 1.2 lies in using –as far as possible– the new control values in the computation of $\boldsymbol{x}_k$ in (1.16a) instead of (1.14). However, in the computation of $\boldsymbol{p}_k$, we cannot use the new values.

2 Justification of (A2) and well-definiteness of the algorithms

In order to be able to apply Assumption (A2), it has to be proved whether the arguments of $z(\cdot)$ appearing in the algorithms are uniformly bounded.

We firstly remark that both Algorithm 1.1 and Algorithm 1.2 can be extended in such a way that the control variables $u_k^{(\ell)}$ $(k = 0, \ldots, N)$ are uniformly bounded. This is natural since, in practice, the control can only range in a given interval.

Proposition 2.1 *Let* $|u_k^{(\ell)}| \leq M$ $(k = 0, \ldots, N,\ \ell = 0, 1, \ldots)$ *with some* $M > 0$ *given. Then there is a constant* $R > 0$*, depending on the partition of the time interval and the data of the problem, such that for all* $t \in [t_k, t_{k+1})$ *and* $k = 0, \ldots, N$, $\ell = 0, 1, \ldots$

$$\|\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k^{(\ell)})\| \leq R \,.$$

Here, $\boldsymbol{x}_k$ *is to be determined by (1.14) or (1.16).*

Proof For simplicity, we omit the superscript indicating the iteration. By Assumption (A1), we immediately have for $k = 0, \ldots, N$, $\ell = 0, 1, \ldots$

$$\|\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k^{(\ell)})\| \leq c\mathrm{e}^{-\lambda(t-t_k)} \|\boldsymbol{x}_k\| + c\tilde{M} \int_{t_k}^{t} \mathrm{e}^{-\lambda(t-\sigma)} d\sigma \,,$$

where $\tilde{M} := M\|\boldsymbol{a}\| + \|\boldsymbol{b}\|$.

Let $\lambda > 0$. It then follows

$$\|\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k^{(\ell)})\| \leq c \, \|\boldsymbol{x}_k\| + \frac{c\tilde{M}}{\lambda} \min\left(1, \lambda(t - t_k)\right) .$$

From (1.14) and (1.16), respectively, we find for $j = 0, \ldots, k-1$

$$\|\boldsymbol{x}_{j+1}\| \leq c\,\|\boldsymbol{x}_j\| + \frac{c\tilde{M}}{\lambda}\min\left(1, \lambda\tau_j\right)$$

that leads (with the convention $(c^k - 1)/(c-1) = k$ for $c = 1$) to

$$\|\boldsymbol{x}_k\| \leq c^k\|\boldsymbol{x}_0\| + \frac{c^k - 1}{c-1}\,\frac{c\tilde{M}}{\lambda}\min\left(1, \lambda\tau_{\max}\right).$$

We thus obtain

$$\begin{aligned}\|\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k^{(\ell)})\| &\leq c^{k+1}\|\boldsymbol{x}_0\| + \frac{c^{k+1} - 1}{c-1}\,\frac{c\tilde{M}}{\lambda}\min\left(1, \lambda\tau_{\max}\right)\\ &\leq c^{N+1}\|\boldsymbol{x}_0\| + \frac{c^{N+1} - 1}{c-1}\,\frac{c\tilde{M}}{\lambda}\min\left(1, \lambda\tau_{\max}\right) =: R\,. \qquad (2.1)\end{aligned}$$

In the case of an infinite time horizon, where $\tau_{\max} = \infty$, we may use the convention

$$\min\left(1, \lambda\tau_{\max}\right) = 1\,.$$

In the case $\lambda = 0$, we have for $k = 0, \ldots, N$

$$\|\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k^{(\ell)})\| \leq c\,\|\boldsymbol{x}_k\| + c\tilde{M}(t - t_k)$$

as well as for $j = 0, \ldots, k-1$

$$\|\boldsymbol{x}_{j+1}\| \leq c\,\|\boldsymbol{x}_j\| + c\tilde{M}\tau_j\,.$$

So we come to

$$\|\boldsymbol{x}_k\| \leq c^k\|\boldsymbol{x}_0\| + c^k\tilde{M}\,t_k\,,$$

and thus it follows

$$\begin{aligned}\|\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k^{(\ell)})\| &\leq c^{k+1}\left(\|\boldsymbol{x}_0\| + \tilde{M}\,t_k\right)\\ &\leq c^{N+1}\left(\|\boldsymbol{x}_0\| + \tilde{M}\,t_{N+1}\right) =: R\,. \qquad (2.2)\end{aligned}$$

Note that the last step requires Assumption (A3). #

From Algorithm 1.1, we see that $u_k^{(\ell)}$ depends, for fixed $k \in \{0, \ldots, N\}$, on $u_0^{(\ell-1)}, \ldots, u_N^{(\ell-1)}$. In order to determine $u_k^{(\ell)}$, we have to resolve

$$\begin{aligned}F_k\left(u_0^{(\ell-1)}, \ldots, u_N^{(\ell-1)}, u_k^{(\ell)}\right) := \mathrm{D}_{u_k}\mathcal{L}_k\Big(u_k^{(\ell)}, \boldsymbol{x}_k(u_0^{(\ell-1)}, \ldots, u_{k-1}^{(\ell-1)}),&\\ \boldsymbol{p}_k(u_0^{(\ell-1)}, \ldots, u_N^{(\ell-1)})\Big) = 0\,. \qquad (2.3)&\end{aligned}$$

However, Algorithm 1.2 requires the solution of

$$G_k\left(u_0^{(\ell-1)},\dots,u_N^{(\ell-1)},u_k^{(\ell)}\right) := \mathrm{D}_{u_k}\mathcal{L}_k\left(u_k^{(\ell)},\boldsymbol{x}_k(u_0^{(\ell)},\dots,u_{k-1}^{(\ell)}),\right.$$
$$\left.\boldsymbol{p}_k(u_0^{(\ell)},\dots,u_{k-1}^{(\ell)},u_k^{(\ell-1)},\dots,u_N^{(\ell-1)})\right) = 0\,. \tag{2.4}$$

Theorem 2.1 *Algorithms 1.1 and 1.2 are well-defined.*

Proof In order to show the unique solvability of (2.3), we shall apply the implicit function theorem. Consider for $v_0,\dots,v_N$ fixed

$$F_k = F_k\left(v_0,\dots,v_N,u_k\right).$$

We then have with (1.10) and (1.6)

$$\mathrm{D}_{u_k}F_k\left(v_0,\dots,v_N,u_k\right) = \frac{\alpha}{r}\,\mathrm{e}^{-rt_k}\left(1-\mathrm{e}^{-r\tau_k}\right) + \int_{t_k}^{t_{k+1}}\mathrm{e}^{-rt}\boldsymbol{a}^\mathsf{T}\times$$
$$\times\int_{t_k}^{t}\mathrm{e}^{(t-\sigma)A^\mathsf{T}}d\sigma\,\mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}}z\left(\boldsymbol{\phi}(t;t_k,\boldsymbol{x}_k,u_k)\right)\int_{t_k}^{t}\mathrm{e}^{(t-\sigma)A}d\sigma\boldsymbol{a}\,dt\,. \tag{2.5}$$

The first term of the right-hand side is obviously positive. By virtue of the convexity of $z = z(\boldsymbol{\phi})$, the second term is nonnegative. Thus we have

$$\mathrm{D}_{u_k}F_k\left(v_0,\dots,v_N,u_k\right) \neq 0\,.$$

It remains to show that $\mathrm{D}_{u_k}F_k\left(v_0,\dots,v_N,u_k\right)$ is finite: Under Assumption (A2), it follows from (2.5) and Proposition 2.1 with some $\kappa=\kappa(R)$

$$\mathrm{D}_{u_k}F_k\left(v_0,\dots,v_N,u_k\right) \le \frac{\alpha}{r}\,\mathrm{e}^{-rt_k}\left(1-\mathrm{e}^{-r\tau_k}\right)$$
$$+\kappa c^2\|\boldsymbol{a}\|^2\int_{t_k}^{t_{k+1}}\mathrm{e}^{-rt}\left(\int_{t_k}^{t}\mathrm{e}^{-\lambda(t-\sigma)}d\sigma\right)^2 dt < \infty\,. \tag{2.6}$$

The proof for Algorithm 1.2 follows the same arguments. #

Let us remark that the well-definiteness relies upon $\mathrm{D}_{u_k}F_k = \mathrm{D}_{u_ku_k}\mathcal{L} > 0$, which is equivalent to the strict convexity of J. Moreover, since J is strongly convex, Problem (1.1) possesses a unique solution.

3 Convergence

Theorem 3.1 *Let* $(u_0^*,\,\dots,\,u_N^*)$ *be the solution to Problem 1.1 and let* $\left\{\left(u_0^{(\ell)},\,\dots,\,u_N^{(\ell)}\right)\right\}_{\ell\in\mathbb{N}}$ *be generated by Algorithm 1.1. It then holds, under Assumptions (A1), (A2), and (A3),*

$$\max_{i=0,\dots,N}\left|u_i^{(\ell+1)}-u_i^*\right| \le \rho\max_{i=0,\dots,N}\left|u_i^{(\ell)}-u_i^*\right|,\quad \ell=0,\,1,\,2,\,\dots \tag{3.1}$$

with

$$\rho := \frac{c^3 \kappa \|\boldsymbol{a}\|^2 \, t_N}{\alpha r} \left(1 - \frac{r \tau_{\max}}{\mathrm{e}^{r\tau_{\max}} - 1} \right) + \frac{c^{N+2} \kappa \|\boldsymbol{a}\|^2}{\alpha r^2} \left(1 + r t_{N+1}\right) \left(1 - \mathrm{e}^{-r t_{N+1}}\right) \tag{3.2}$$

in the case $\lambda = 0$ *and*

$$\rho := \frac{c^3 \kappa \|\boldsymbol{a}\|^2}{4 \alpha \lambda^2} \left(1 - \mathrm{e}^{-\lambda t_N}\right) + \frac{c^{N+2} \kappa \|\boldsymbol{a}\|^2}{\alpha r \lambda} \left(1 - \mathrm{e}^{-r t_{N+1}}\right) \tag{3.3}$$

in the case $\lambda > 0$*.*

Proof Due to Theorem 2.1, there exist functions $f_k : \mathbb{R}^{N+1} \to \mathbb{R}$ ($k = 0 \ldots, N$) with

$$u_k = f_k(v_0, \ldots, v_N) \;\Leftrightarrow\; F_k(v_0, \ldots, v_N, u_k) = 0\,,$$

and these are continuously differentiable as we can infer from the proof of Theorem 2.1. We will show that

$$\max_{k=0,\ldots,N} \sum_{j=0}^{N} \left| \mathrm{D}_{v_j} f_k(v_0, \ldots, v_N) \right| \le \rho$$

holds true for all $v_0, \ldots, v_N \in \mathbb{R}$. The assertion then follows because of

$$\begin{aligned} \left| u_k^{(\ell+1)} - u_k^* \right| &= \left| f_k(u_0^{(\ell)}, \ldots, u_N^{(\ell)}) - f_k(u_0^*, \ldots, u_N^*) \right| \\ &= \left| \sum_{j=0}^{N} \mathrm{D}_{u_j} f_k(\bar{u}_0, \ldots, \bar{u}_N)\, (u_j^{(\ell)} - u_j^*) \right| \\ &\le \sum_{j=0}^{N} \left| \mathrm{D}_{u_j} f_k(\bar{u}_0, \ldots, \bar{u}_N) \right| \max_{i=0,\ldots,N} \left| u_i^{(\ell)} - u_i^* \right| , \end{aligned}$$

where $(\bar{u}_0, \ldots, \bar{u}_N)$ is a point lying on the line connecting $(u_0^{(\ell)}, \ldots, u_N^{(\ell)})$ and $(u_0^*, \ldots, u_N^*)$.

It is

$$\mathrm{D}_{v_j} f_k(v_0, \ldots, v_N) = - \frac{\mathrm{D}_{v_j} F_k(v_0, \ldots, v_N, u_k)}{\mathrm{D}_{u_k} F_k(v_0, \ldots, v_N, u_k)}\,, \quad j,\, k = 0,\, \ldots\,, N\,. \tag{3.4}$$

Because of (2.3), we find

$$\mathrm{D}_{v_j} F_k(v_0, \ldots, v_N, u_k) = \mathrm{D}_{\boldsymbol{x}_k u_k} \mathcal{L}_k \mathrm{D}_{v_j} \boldsymbol{x}_k + \mathrm{D}_{\boldsymbol{p}_k u_k} \mathcal{L}_k \mathrm{D}_{v_j} \boldsymbol{p}_k\,, \tag{3.5}$$

where $\mathcal{L}_k = \mathcal{L}_k(u_k, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}, \boldsymbol{p}_k)$ (but $\mathrm{D}_{u_k}\mathcal{L}_k = \mathrm{D}_{u_k}\mathcal{L}_k(u_k, \boldsymbol{x}_k, \boldsymbol{p}_k)$), $\boldsymbol{x}_k = \boldsymbol{x}_k(v_0, \ldots, v_{k-1})$, and $\boldsymbol{p}_k = \boldsymbol{p}_k(v_0, \ldots, v_N)$. Together with (2.5), we obtain

$$\sum_{j=0}^{N} \left|\mathrm{D}_{v_j} f_k(v_0, \ldots, v_N)\right| \le \frac{r}{\alpha}\left(\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}}\right)^{-1} \times$$
$$\times \left(\|\mathrm{D}_{\boldsymbol{x}_k u_k}\mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{x}_k\right\| + \|\mathrm{D}_{\boldsymbol{p}_k u_k}\mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_k\right\| \right). \tag{3.6}$$

With (1.10) and (1.6), we observe that

$$\mathrm{D}_{\boldsymbol{x}_k u_k}\mathcal{L}_k = \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt}\boldsymbol{a}^{\mathsf{T}} \int_{t_k}^{t} \mathrm{e}^{(t-\sigma)A^{\mathsf{T}}} d\sigma\, \mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z(\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k))\mathrm{e}^{(t-t_k)A} dt\,, \tag{3.7}$$

$$\mathrm{D}_{\boldsymbol{p}_k u_k}\mathcal{L}_k = \boldsymbol{a}^{\mathsf{T}} \int_{t_k}^{t_{k+1}} \mathrm{e}^{(t_{k+1}-\sigma)A^{\mathsf{T}}} d\sigma\,. \tag{3.8}$$

From (1.14), we immediately have $\mathrm{D}_{v_j}\boldsymbol{x}_k = 0$ if $j \ge k$ as well as

$$\mathrm{D}_{v_j}\boldsymbol{x}_{j+1} = \int_{t_j}^{t_{j+1}} \mathrm{e}^{(t_{j+1}-\sigma)A} d\sigma\, \boldsymbol{a}\,.$$

Moreover, it is for $j < k-1$

$$\begin{aligned}\mathrm{D}_{v_j}\boldsymbol{x}_k = \mathrm{D}_{\boldsymbol{x}_{k-1}}&\boldsymbol{\phi}(t_k; t_{k-1}, \boldsymbol{x}_{k-1}, v_{k-1}) \times \ldots \\ &\ldots \times \mathrm{D}_{\boldsymbol{x}_{j+1}}\boldsymbol{\phi}(t_{j+2}; t_{j+1}, \boldsymbol{x}_{j+1}, v_{j+1})\mathrm{D}_{v_j}\boldsymbol{x}_{j+1}\,,\end{aligned}$$

and so we come up with

$$\mathrm{D}_{v_j}\boldsymbol{x}_k = \begin{cases} \displaystyle\int_{t_j}^{t_{j+1}} \mathrm{e}^{(t_k-\sigma)A} d\sigma\, \boldsymbol{a} & \text{for } j < k\,, \\ 0 & \text{for } j \ge k\,. \end{cases} \tag{3.9}$$

From (1.15), we see that for $l = N, \ldots, k+1$ with $\boldsymbol{p}_N = 0$

$$\mathrm{D}_{v_j}\boldsymbol{p}_{l-1} = \int_{t_l}^{t_{l+1}} \mathrm{e}^{-rt}\mathrm{e}^{(t-t_l)A^{\mathsf{T}}} \left(\mathrm{D}_{v_j\boldsymbol{\phi}} z\left(\boldsymbol{\phi}(t; t_l, \boldsymbol{x}_l, v_l)\right)\right)^{\mathsf{T}} dt + \mathrm{e}^{\tau_l A^{\mathsf{T}}}\mathrm{D}_{v_j}\boldsymbol{p}_l\,. \tag{3.10}$$

Furthermore, we have

$$\left(\mathrm{D}_{v_j \boldsymbol{\phi}} z\left(\boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l)\right)\right)^{\mathsf{T}} = \mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z\left(\boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l)\right) \mathrm{D}_{v_j} \boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l) \quad (3.11)$$

as well as (by (1.6) and (3.9))

$$\begin{aligned}\mathrm{D}_{v_j} \boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l) &= \mathrm{D}_{\boldsymbol{x}_l} \boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l)\, \mathrm{D}_{v_j} \boldsymbol{x}_l + \mathrm{D}_{v_l}\, \boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l) \mathrm{D}_{v_j} v_l \\ &= \begin{cases} \displaystyle\int_{t_j}^{t_{j+1}} \mathrm{e}^{(t-\sigma)A} d\sigma\, \boldsymbol{a} & \text{for } j < l\,, \\ \displaystyle\int_{t_j}^{t} \mathrm{e}^{(t-\sigma)A} d\sigma\, \boldsymbol{a} & \text{for } j = l\,, \\ 0 & \text{for } j > l\,. \end{cases} \end{aligned} \quad (3.12)$$

Under Assumption (A1) and (A2), it follows from (3.7)

$$\|\mathrm{D}_{\boldsymbol{x}_k u_k} \mathcal{L}_k\| \le c^2 \kappa \|\boldsymbol{a}\| \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} \int_{t_k}^{t} \mathrm{e}^{-\lambda(t-\sigma)} d\sigma\, \mathrm{e}^{-\lambda(t-t_k)} dt\,,$$

and from (3.9)

$$\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j} \boldsymbol{x}_k\right\| \le c \|\boldsymbol{a}\| \int_{0}^{t_k} \mathrm{e}^{-\lambda(t_k-\sigma)} d\sigma\,.$$

Hence, we have for $\lambda = 0$

$$\begin{aligned}\|\mathrm{D}_{\boldsymbol{x}_k u_k} \mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j} \boldsymbol{x}_k\right\| &\le c^3 \kappa \|\boldsymbol{a}\|^2\, t_k \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} (t-t_k) dt \\ &= c^3 \kappa \|\boldsymbol{a}\|^2\, \frac{t_k}{r^2} \left(\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}} - r\tau_k \mathrm{e}^{-rt_{k+1}}\right). \end{aligned} \quad (3.13)$$

For $\lambda > 0$, we find

$$\begin{aligned}&\|\mathrm{D}_{\boldsymbol{x}_k u_k} \mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j} \boldsymbol{x}_k\right\| \\ &\le \frac{c^3 \kappa \|\boldsymbol{a}\|^2}{\lambda^2} \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} \left(1 - \mathrm{e}^{-\lambda(t-t_k)}\right) \mathrm{e}^{-\lambda(t-t_k)} dt \left(1 - \mathrm{e}^{-\lambda t_k}\right).\end{aligned}$$

Since the function $t \mapsto \left(1 - \mathrm{e}^{-\lambda(t-t_k)}\right) \mathrm{e}^{-\lambda(t-t_k)}$ takes its maximum value $1/4$ at $t = t_k + (\ln 2)/\lambda$, we have

$$\|\mathrm{D}_{\boldsymbol{x}_k u_k} \mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j} \boldsymbol{x}_k\right\| \le \frac{c^3 \kappa \|\boldsymbol{a}\|^2}{4r\lambda^2} \left(\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}}\right) \left(1 - \mathrm{e}^{-\lambda t_N}\right). \quad (3.14)$$

Because of (3.8), we have for $k = 0, \dots, N$

$$\|\mathrm{D}_{\boldsymbol{p}_k u_k}\mathcal{L}_k\| \le c\|\boldsymbol{a}\| \begin{cases} \tau_k & \text{for } \lambda = 0\,, \\ \frac{1}{\lambda}\left(1 - \mathrm{e}^{-\lambda\tau_k}\right) & \text{for } \lambda > 0\,. \end{cases} \tag{3.15}$$

From (3.10) and (3.12), we see that for $l = N, \dots, 1$

$$\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_{l-1}\right\| \le c^2\kappa\|\boldsymbol{a}\| \int_{t_l}^{t_{l+1}} \mathrm{e}^{-rt-\lambda(t-t_l)} \int_0^t \mathrm{e}^{-\lambda(t-\sigma)} d\sigma\, dt$$
$$+ c\mathrm{e}^{-\lambda\tau_l} \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_l\right\|. \tag{3.16}$$

For $\lambda = 0$, it follows

$$\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_{l-1}\right\| \le c^2\kappa\|\boldsymbol{a}\| \int_{t_l}^{t_{l+1}} t\mathrm{e}^{-rt}dt + c\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_l\right\|,$$

and so we obtain for $k = 0, \dots, N$ since $\boldsymbol{p}_N = 0$

$$\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_k\right\| \le c^{N-k+1}\kappa\|\boldsymbol{a}\| \int_{t_{k+1}}^{t_{N+1}} t\mathrm{e}^{-rt}dt$$
$$= c^{N-k+1}\kappa\|\boldsymbol{a}\| \frac{1}{r^2}\left((1+rt_{k+1})\mathrm{e}^{-rt_{k+1}} - (1+rt_{N+1})\mathrm{e}^{-rt_{N+1}}\right).$$

We thus find

$$\|\mathrm{D}_{\boldsymbol{p}_k u_k}\mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_k\right\|$$
$$\le c^{N+2}\kappa\|\boldsymbol{a}\|^2 \frac{\tau_k}{r^2}\left((1+rt_{k+1})\mathrm{e}^{-rt_{k+1}} - (1+rt_{N+1})\mathrm{e}^{-rt_{N+1}}\right). \tag{3.17}$$

For $\lambda > 0$, we have

$$\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_{l-1}\right\| \le \frac{c^2\kappa\|\boldsymbol{a}\|}{\lambda} \int_{t_l}^{t_{l+1}} \mathrm{e}^{-rt-\lambda(t-t_l)}\left(1 - \mathrm{e}^{-\lambda t}\right) dt$$
$$+\ c\mathrm{e}^{-\lambda\tau_l} \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_l\right\|$$
$$\le \frac{c^2\kappa\|\boldsymbol{a}\|}{\lambda} \int_{t_l}^{t_{l+1}} \mathrm{e}^{-rt}dt + c\mathrm{e}^{-\lambda\tau_l} \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_l\right\|,$$

and thus

$$\sum_{j=0}^{N} \left\| \mathrm{D}_{v_j} \boldsymbol{p}_{l-1} \right\| \leq \frac{c^2 \kappa \|\boldsymbol{a}\|}{\lambda} \left(\int_{t_l}^{t_{l+1}} \mathrm{e}^{-rt} dt + \right.$$

$$+ c\mathrm{e}^{-\lambda(t_{l+1}-t_l)} \int_{t_{l+1}}^{t_{l+2}} \mathrm{e}^{-rt} dt + c^2 \mathrm{e}^{-\lambda(t_{l+2}-t_l)} \int_{t_{l+2}}^{t_{l+3}} \mathrm{e}^{-rt} dt +$$

$$\left. + \cdots + c^{N-l} \mathrm{e}^{-\lambda(t_N - t_l)} \int_{t_N}^{t_{N+1}} \mathrm{e}^{-rt} dt \right) .$$

So we have

$$\begin{aligned} \sum_{j=0}^{N} \left\| \mathrm{D}_{v_j} \boldsymbol{p}_k \right\| &\leq \frac{c^{N+1} \kappa \|\boldsymbol{a}\|}{\lambda} \int_{t_{k+1}}^{t_{N+1}} \mathrm{e}^{-rt} dt \\ &= \frac{c^{N+1} \kappa \|\boldsymbol{a}\|}{r\lambda} \left(\mathrm{e}^{-rt_{k+1}} - \mathrm{e}^{-rt_{N+1}} \right) \end{aligned}$$

and obtain for $k = 0, \ldots, N$

$$\|\mathrm{D}_{\boldsymbol{p}_k u_k} \mathcal{L}_k\| \sum_{j=0}^{N} \left\| \mathrm{D}_{v_j} \boldsymbol{p}_k \right\| \leq \frac{c^{N+2} \kappa \|\boldsymbol{a}\|^2}{r\lambda^2} \left(1 - \mathrm{e}^{-\lambda\tau_k}\right) \left(\mathrm{e}^{-rt_{k+1}} - \mathrm{e}^{-rt_{N+1}} \right) . \tag{3.18}$$

In the case $\lambda = 0$, (3.6) reads, together with (3.13) and (3.17), as

$$\sum_{j=0}^{N} \left| \mathrm{D}_{v_j} f_k(v_0, \ldots, v_N) \right| \leq c^3 \kappa \|\boldsymbol{a}\|^2 \frac{t_k}{\alpha r} \left(1 - \frac{r\tau_k}{\mathrm{e}^{r\tau_k} - 1} \right) +$$

$$+ c^{N+2} \kappa \|\boldsymbol{a}\|^2 \frac{\tau_k}{\alpha r} \frac{1 + rt_{k+1} - (1 + rt_{N+1}) \mathrm{e}^{-r(t_{N+1} - t_{k+1})}}{\mathrm{e}^{r\tau_k} - 1} . \tag{3.19}$$

Since

$$\frac{r\tau_{\max}}{\mathrm{e}^{r\tau_{\max}} - 1} \leq \frac{r\tau_k}{\mathrm{e}^{r\tau_k} - 1} < 1 , \tag{3.20}$$

it follows

$$\sum_{j=0}^{N} \left| \mathrm{D}_{v_j} f_k(v_0, \ldots, v_N) \right| \leq c^3 \kappa \|\boldsymbol{a}\|^2 \frac{t_N}{\alpha r} \left(1 - \frac{r\tau_{\max}}{\mathrm{e}^{r\tau_{\max}} - 1} \right) +$$

$$+ c^{N+2} \kappa \|\boldsymbol{a}\|^2 \frac{1}{\alpha r^2} \left(1 + rt_{N+1} - (1 + rt_{N+1}) \mathrm{e}^{-rt_{N+1}} \right) ,$$

which leads to (3.2).

In the case $\lambda > 0$, (3.6) reads, together with (3.14) and (3.18), as

$$\sum_{j=0}^{N} \left|\mathrm{D}_{v_j} f_k(v_0, \ldots, v_N)\right| \le \frac{c^3 \kappa \|\boldsymbol{a}\|^2}{4\alpha\lambda^2} \left(1 - \mathrm{e}^{-\lambda t_N}\right) + \\ + \frac{c^{N+2}\kappa\|\boldsymbol{a}\|^2}{\alpha\lambda^2} \left(1 - \mathrm{e}^{-\lambda\tau_k}\right) \frac{1 - \mathrm{e}^{-r(t_{N+1} - t_{k+1})}}{\mathrm{e}^{r\tau_k} - 1} .$$

Since

$$1 - \mathrm{e}^{-\lambda\tau_k} \le \lambda\tau_k ,$$

it follows with (3.20)

$$\sum_{j=0}^{N} \left|\mathrm{D}_{v_j} f_k(v_0, \ldots, v_N)\right| \le \frac{c^3 \kappa \|\boldsymbol{a}\|^2}{4\alpha\lambda^2} \left(1 - \mathrm{e}^{-\lambda t_N}\right) + \\ + \frac{c^{N+2}\kappa\|\boldsymbol{a}\|^2}{\alpha r\lambda} \left(1 - \mathrm{e}^{-r t_{N+1}}\right) ,$$

which leads to (3.3). #

Remark 3.1 *If $\rho < 1$ then Algorithm 1.1 is convergent. This is only fulfilled if, in particular, κ and t_{N+1} are small or α, r, and λ are large. As estimate (3.6) can be improved if z is strongly convex (cf. relation (2.5) for the denominator in (3.4)), the estimates (3.2) and (3.3) for ρ can then be improved, too. However, this needs to have concrete information on z.*

Theorem 3.1 also includes the case $\lambda > 0$ and $t_{N+1} = \infty$. We then use the convention $\mathrm{e}^{-\lambda t_{N+1}} = \mathrm{e}^{-r t_{N+1}} = 0$.

Theorem 3.2 *Let $(u_0^*, \ldots, u_N^*)$ be the solution to Problem 1.1 and let $\left\{\left(u_0^{(\ell)}, \ldots, u_N^{(\ell)}\right)\right\}_{\ell\in\mathbb{N}}$ be generated by Algorithm 1.2. It then holds, under Assumptions (A1), (A2), and (A3), estimate (3.1) with*

$$\rho := \frac{c^{N+2}\kappa\|\boldsymbol{a}\|^2}{\alpha r^2} \left(1 + r\tau_{\max} - (1 + r t_{N+1})\mathrm{e}^{-r t_{N+1}}\right) \tag{3.21}$$

in the case $\lambda = 0$ and

$$\rho := \frac{c^{N+2}\kappa\|\boldsymbol{a}\|^2}{\alpha r\lambda} \left(1 - \mathrm{e}^{-r t_{N+1}}\right) \tag{3.22}$$

in the case $\lambda > 0$.

Proof The first part is the same as in the proof of Theorem 3.1 when replacing f_k by g_k. Because of (2.4), we have

$$G_k(v_0,\dots,v_N,u_k) = \mathrm{D}_{u_k}\boldsymbol{\mathcal{L}}_k\left(u_k,\boldsymbol{x}_k(u_0,\dots,u_{k-1}),\boldsymbol{p}_k(u_0,\dots,u_{k-1},v_k,\dots,v_N)\right)$$

and, therefore, analogously to (2.5)

$$\begin{aligned}\mathrm{D}_{u_k}G_k\left(v_0,\dots,v_N,u_k\right) &= \frac{\alpha}{r}\,\mathrm{e}^{-rt_k}\left(1-\mathrm{e}^{-r\tau_k}\right) + \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt}\boldsymbol{a}^{\mathsf{T}} \times \\ \times \int_{t_k}^{t} \mathrm{e}^{(t-\sigma)A^{\mathsf{T}}}d\sigma\, &\mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}}z\left(\boldsymbol{\phi}(t;t_k,\boldsymbol{x}_k,u_k)\right)\int_{t_k}^{t}\mathrm{e}^{(t-\sigma)A}d\sigma\boldsymbol{a}\,dt \\ &\geq \frac{\alpha}{r}\,\mathrm{e}^{-rt_k}\left(1-\mathrm{e}^{-r\tau_k}\right).\end{aligned}$$

In opposite to (3.5), it holds

$$\mathrm{D}_{v_j}G_k(v_0,\dots,v_N,u_k) = \mathrm{D}_{\boldsymbol{p}_k u_k}\boldsymbol{\mathcal{L}}_k\mathrm{D}_{v_j}\boldsymbol{p}_k\,,$$

whereas (3.8), (3.10), (3.11), and (3.15) remain valid. We have to estimate

$$\sum_{j=0}^{N}\left|\mathrm{D}_{v_j}g_k(v_0,\dots,v_N)\right| \leq \frac{r}{\alpha}\,\left(\mathrm{e}^{-rt_k}-\mathrm{e}^{-rt_{k+1}}\right)^{-1}\|\mathrm{D}_{\boldsymbol{p}_k u_k}\boldsymbol{\mathcal{L}}_k\|\sum_{j=0}^{N}\left\|\mathrm{D}_{v_j}\boldsymbol{p}_k\right\|. \tag{3.23}$$

Because of (1.16), we have $\boldsymbol{x}_k = \boldsymbol{x}_k\left(u_0,\dots,u_{k-1}\right)$ and

$$\boldsymbol{x}_l = \boldsymbol{x}_l\left(u_0,\dots,u_{k-1},v_k,\dots,v_{l-1}\right),\quad l=k+1,\dots,N\,.$$

It follows $\mathrm{D}_{v_j}\boldsymbol{x}_l = 0$ for $j=0,\dots,k-1$ and for $j=l,\dots,N$. From (1.16b) and (1.6), we find for $j=k,\dots,l-1$

$$\begin{aligned}\mathrm{D}_{v_j}\boldsymbol{x}_l &= \mathrm{e}^{\tau_{l-1}A}\mathrm{D}_{v_j}\boldsymbol{x}_{l-1} + \int_{t_{l-1}}^{t_l}\mathrm{e}^{(t_l-\sigma)A}d\sigma\,\boldsymbol{a}\,\mathrm{D}_{v_j}v_{l-1} \\ &= \mathrm{e}^{(t_l-t_k)A}\mathrm{D}_{v_j}\boldsymbol{x}_k + \sum_{\mu=0}^{l-k-1}\int_{t_{k+\mu}}^{t_{k+1+\mu}}\mathrm{e}^{(t_l-\sigma)A}d\sigma\,\boldsymbol{a}\,\mathrm{D}_{v_j}v_{k+\mu} \\ &= \int_{t_j}^{t_{j+1}}\mathrm{e}^{(t_l-\sigma)A}d\sigma\,\boldsymbol{a}\,.\end{aligned}$$

So the identity (3.12) becomes

$$\begin{aligned}
\mathrm{D}_{v_j}\boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l) &= \mathrm{D}_{\boldsymbol{x}_l}\boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l)\,\mathrm{D}_{v_j}\boldsymbol{x}_l + \mathrm{D}_{v_l}\,\boldsymbol{\phi}(t;t_l,\boldsymbol{x}_l,v_l)\mathrm{D}_{v_j}v_l \\
&= \begin{cases} \displaystyle\int_{t_j}^{t_{j+1}} \mathrm{e}^{(t-\sigma)A} d\sigma\,\boldsymbol{a} & \text{for } j=k,\dots,l-1\,, \\ \displaystyle\int_{t_j}^{t} \mathrm{e}^{(t-\sigma)A} d\sigma\,\boldsymbol{a} & \text{for } j=l\,, \\ 0 & \text{else} \end{cases}
\end{aligned} \tag{3.24}$$

for $l=k+1,\dots,N$, and we obtain from (3.10) and (3.11)

$$\begin{aligned}
\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_{l-1}\right\| \le c^2\kappa\|\boldsymbol{a}\| \int_{t_l}^{t_{l+1}} \mathrm{e}^{-rt-\lambda(t-t_l)} \int_{t_k}^{t} \mathrm{e}^{-\lambda(t-\sigma)} d\sigma\,dt \\
+ c\mathrm{e}^{-\lambda\tau_l} \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_l\right\|\,.
\end{aligned} \tag{3.25}$$

Note the difference between (3.16) and (3.25): the second integral is taken over $(0,t)$ and (t_k,t), respectively. It follows for $\lambda=0$

$$\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_{l-1}\right\| \le c^{N-l+2}\kappa\|\boldsymbol{a}\| \int_{t_l}^{t_{N+1}} (t-t_k)\mathrm{e}^{-rt}\,dt$$

and, therefore, with $l=k+1$ using (3.15)

$$\begin{aligned}
\|\mathrm{D}_{\boldsymbol{p}_k u_k}\mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_k\right\| &\le c^{N-k+2}\kappa\|\boldsymbol{a}\|^2\tau_k \int_{t_{k+1}}^{t_{N+1}} (t-t_k)\mathrm{e}^{-rt}\,dt \\
&\le c^{N+2}\kappa\|\boldsymbol{a}\|^2\,\frac{\tau_k}{r^2}\left((1+r\tau_k)\mathrm{e}^{-rt_{k+1}} - (1+r(t_{N+1}-t_k))\mathrm{e}^{-rt_{N+1}}\right)
\end{aligned}$$

instead of (3.17). We find

$$\begin{aligned}
&\sum_{j=0}^{N} \left|\mathrm{D}_{v_j} g_k(v_0,\dots,v_N)\right| \\
&\le c^{N+2}\kappa\|\boldsymbol{a}\|^2\,\frac{\tau_k}{\alpha r}\,\frac{1+r\tau_k-(1+r(t_{N+1}-t_k))\mathrm{e}^{-r(t_{N+1}-t_{k+1})}}{\mathrm{e}^{r\tau_k}-1}\,.
\end{aligned}$$

This gives, together with (3.20) and

$$(1+r(t_{N+1}-t_k))\mathrm{e}^{-r(t_{N+1}-t_k)} \ge (1+rt_{N+1})\mathrm{e}^{-rt_{N+1}}\,,$$

the assertion.

If $\lambda > 0$, we do not make use of the difference between (3.16) and (3.25), and so we arrive –as in the proof of Theorem 3.1– at (3.18). This gives

$$\sum_{j=0}^{N} \left| \mathrm{D}_{v_j} g_k(v_0, \ldots, v_N) \right| \le \frac{c^{N+2}\kappa \|\boldsymbol{a}\|^2}{\alpha\lambda^2} \left(1 - \mathrm{e}^{-\lambda\tau_k}\right) \frac{1 - \mathrm{e}^{-r(t_{N+1} - t_{k+1})}}{\mathrm{e}^{r\tau_k} - 1} \,,$$

which proves, together with (3.20), the assertion. #

Remark 3.1 also applies with respect to Algorithm 1.2 and Theorem 3.2.

4 Infinite time horizon in the case $\lambda = 0$

As we have already mentioned, an infinite time interval is of particular interest in applied problems even if $\lambda = 0$. In this case, we need to replace Assumption (A2), which can be no longer justified since the boundedness of $\boldsymbol{\phi}$ is not at hand, and Assumption (A3) by the more refined criterion (A4). In the following, let

$$\tau_{\max,0} := \max_{i=0,\ldots,N-1} \tau_i \,.$$

Theorem 4.1 *Let $\lambda = 0$ in (A1). Under Assumption (A4), Theorem 2.1 then remains valid. Furthermore, Theorem 3.1 holds true with*

$$\rho := \frac{c^3 \|\boldsymbol{a}\|^2 K r}{\alpha} \left(t_N + \frac{c^N - 1}{c - 1}\, \tau_{\max,0} \right) \max_{k=0,\ldots,N} \left(\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}} \right)^{-1} \tag{4.1}$$

whereas Theorem 3.2 holds true with

$$\rho := \frac{c^3 \|\boldsymbol{a}\|^2 K r}{\alpha} \, \frac{c^N - 1}{c - 1}\, \tau_{\max,0} \max_{k=0,\ldots,N} \left(\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}} \right)^{-1} \,, \tag{4.2}$$

and (1.8) need not to be assumed.

Proof The proof of Theorem 2.1 has to be slightly changed since (2.6) is no longer true: We replace the estimate (2.6) by

$$\mathrm{D}_{u_k} F_k \left(v_0, \ldots, v_N, u_k \right) \le \frac{\alpha}{r}\, \mathrm{e}^{-rt_k} \left(1 - \mathrm{e}^{-r\tau_k} \right) +$$
$$+ c^2 \|\boldsymbol{a}\|^2 \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} (t - t_k)^2 \left\| \mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z \left(\boldsymbol{\phi}(t; t_k, \boldsymbol{x}_k, u_k) \right) \right\| \, dt < \infty \,, \tag{4.3}$$

which follows from (2.5) and (A4).

In order to prove the convergence of Algorithm 1.1, we firstly observe that, instead of (3.13), from (3.7) and (3.9)

$$\|\mathrm{D}_{\boldsymbol{x}_k u_k}\mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{x}_k\right\| \leq c^3\|\boldsymbol{a}\|^2 K t_k \leq c^3\|\boldsymbol{a}\|^2 K t_N \tag{4.4}$$

follows. Inequality (3.15) remains true for $k = 0, \ldots, N-1$. Since $\boldsymbol{p}_N = 0$, $\mathcal{L}_N$ does not depend on $\boldsymbol{p}_N$ and so $\mathrm{D}_{\boldsymbol{p}_N u_N}\mathcal{L}_N = 0$. From (3.10) and (3.12), we find with (A4)

$$\begin{aligned}\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_{l-1}\right\| &\leq c^2\|\boldsymbol{a}\| \int_{t_l}^{t_{l+1}} t\mathrm{e}^{-rt}\|\mathrm{D}_{\boldsymbol{\phi\phi}} z(\boldsymbol{\phi}(t; t_l, \boldsymbol{x}_l, v_l))\| dt + c\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_l\right\| \\ &\leq c^2\|\boldsymbol{a}\|K + c\sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_l\right\|,\end{aligned}$$

and so we obtain for $k = 0, \ldots, N$ since $\boldsymbol{p}_N = 0$

$$\|\mathrm{D}_{\boldsymbol{p}_k u_k}\mathcal{L}_k\| \sum_{j=0}^{N} \left\|\mathrm{D}_{v_j}\boldsymbol{p}_k\right\| \leq c^3\|\boldsymbol{a}\|^2 K \,\frac{c^N - 1}{c-1}\, \tau_{\max,0} \tag{4.5}$$

instead of (3.17).

By virtue of (3.6), we finally come up with

$$\max_{k=0,\ldots,N} \sum_{j=0}^{N} \left|\mathrm{D}_{v_j} f_k(v_0, \ldots, v_N)\right| \leq$$

$$\frac{c^3\|\boldsymbol{a}\|^2 K r}{\alpha} \left(t_N + \frac{c^N - 1}{c-1}\tau_{\max,0}\right) \max_{k=0,\ldots,N} \left(\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}}\right)^{-1} \tag{4.6}$$

instead of (3.19), and the assertion follows.

The proof of the convergence of Algorithm 1.2 in the case $\lambda = 0$ with infinite time horizon follows the same arguments as in the proof of Theorem 3.2 with the observations just made (compare (3.23) with (3.6)). However, due to the difference between (3.16) and (3.25), (1.8) need not to be satisfied. #

It should be noted that the above convergence result is rather rough: A more refined estimate can be only obtained with a more refined Assumption (A4) relying on the concrete function z and the given time partition.

5 Numerical results for a global warming model

Wirl [10] considered the Nordhaus model of global warming in which the greenhouse effect is represented by the following initial value problem:

$$\begin{aligned} \dot{M}(t) &= -\delta\, M(t) + \beta\,\vartheta\, E(p(t))\,, \quad M(0) = M_0\,, \\ \dot{T}(t) &= \alpha(\gamma(M(t)) - T(t))\,, \quad T(0) = T_0 = 0\,. \end{aligned}$$

The first differential equation accumulates the greenhouse gas concentration in the atmosphere $M(t)$. Its dynamics is determined by the natural reduction $-\delta\, M(t)$ (the parameter $\delta \geq 0$ has been proposed by Wirl later on, although in [10], he deals with $\delta = 0$) and by the consumption of energy $E(p(t))$ that depends on the price p per unit of energy. Here, the price p plays the rôle of the control variable. Each unit of (fossil) energy induces inevitably carbon dioxide emissions. The parameter ϑ denotes the average ratio of carbon dioxide emissions per unit of fossil energy and is aggregated over the diffusion coefficient β.

A rise of the carbon dioxide concentration in the atmosphere increases stationary the earth temperature $T(t)$ according to the function $\gamma(M)$. The assessment of the function γ depends on the model used to quantify this effect (e. g. global circulation models). The parameter α comes from the linear demand function. The temperature T is modelled in terms of deviations from present temperature so that $T_0 = 0$.

The social planner maximises the discounted so-called welfare function W minus the cost C of the increase in temperature T that leads to the following objective function

$$K := \int_0^\infty \mathrm{e}^{-rt}\Big(W(p(t)) - C(T(t))\Big)\, dt\,,$$

where the welfare function $W = W(p(t))$ is the sum of consumer and producer surplus at time t. Finally, the objective function depends on the discount rate $r \in (0,1)$. For more details, we refer to [10].

For the numerical experiments, we shall focus our attention on the following class of problems that covers the above application: Let for $t \in [t_k, t_{k+1})$ $(k = 0, \ldots, N)$ the functions $x(t) = x(t; t_k, x_k, y_k, u_k)$ and $y(t) = y(t; t_k, x_k, y_k, u_k)$ be the solution to the initial value problem

$$\dot{x}(t) = -\,\zeta\, x(t) + a\, u_k + b_1\,, \quad x(t_k) = x_k\,, \tag{5.1a}$$

$$\dot{y}(t) = \eta\,(\xi\, x(t) - y(t) + b_2)\,, \quad y(t_k) = y_k\,, \tag{5.1b}$$

where $a, b_1, b_2, \xi, \eta, \zeta \in \mathbb{R}$ and $\eta > 0$, $\zeta \geq 0$. Here, u_k is the price to be controlled, which is assumed to be piecewise constant. Moreover, the consumption of energy depends upon the price in an affine-linear way. The relation between the carbon dioxide concentration and the corresponding increase of the temperature $\gamma(M(t))$ is also taken to be affine-linear.

System (5.1) coincides with (1.5) taking

$$A = \begin{pmatrix} -\zeta & 0 \\ \xi\,\eta & -\eta \end{pmatrix}, \quad \boldsymbol{a} = \begin{pmatrix} a \\ 0 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ \eta\, b_2 \end{pmatrix}.$$

Since A possesses the eigenvalues $-\zeta$ and $-\eta$, Assumption (A1) is fulfilled with $\lambda := \min\{\zeta, \eta\}$. For simplicity, let us introduce the abbreviation $\eta_k(t) := 1 - \mathrm{e}^{-\eta(t-t_k)}$ and $\zeta_k(t) := 1 - \mathrm{e}^{-\zeta(t-t_k)}$ for $t \geq t_k$. The exact solution to (5.1) is then given by

$$x(t) = \begin{cases} x_k(1-\zeta_k(t)) + (a\,u_k + b_1)\zeta^{-1}\zeta_k(t) & \text{for } \zeta > 0\,, \\ x_k + (a\,u_k + b_1)(t - t_k) & \text{for } \zeta = 0\,, \end{cases} \tag{5.2a}$$

and

$$y(t) = \begin{cases} \xi(a\,u_k + b_1)(t - t_k - \eta^{-1}\eta_k(t)) \\ \quad + y_k(1-\eta_k(t)) + (\xi\, x_k + b_2)\eta_k(t) & \text{for } \zeta = 0\,, \\ \xi\,\eta(a\,u_k + b_1)(\eta-\zeta)^{-1}(\zeta^{-1}\zeta_k(t) - \eta^{-1}\eta_k(t)) \\ \quad + y_k(1-\eta_k(t)) + b_2\,\eta_k(t) \\ \quad + \xi\,\eta\, x_k(\eta-\zeta)^{-1}(\eta_k(t) - \zeta_k(t)) & \text{for } \zeta \neq \eta\,, \\ y_k(1-\eta_k(t)) + (\xi\,\eta^{-1}(a\,u_k + b_1) + b_2)\eta_k(t) \\ \quad + \xi\,\eta(x_k - \eta^{-1}(a\,u_k + b_1))(t - t_k)(1-\eta_k(t)) & \text{for } \zeta = \eta\,. \end{cases} \tag{5.2b}$$

According to these state equations, we consider two types of objective functions:

$$I(u_0, \ldots, u_N, x_0, \ldots, x_N, y_0, \ldots, y_N) := \\ \sum_{k=0}^{N} \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} \left(\frac{\alpha}{2}\, u_k^2 + \frac{\beta}{2}\, y^2(t; t_k, x_k, y_k, u_k) \right) dt$$

and

$$J(u_0,\dots,u_N,x_0,\dots,x_N,y_0,\dots,y_N) := \\ \sum_{k=0}^{N}\int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt}\left(\frac{\alpha}{2}\,u_k^2 + \mathrm{e}^{\gamma\, y(t;t_k,x_k,y_k,u_k)}\right)dt\,.$$

The functional I appears in [10] by assuming a linear demand function and quadratic costs with respect to the rise in temperature. The objective is to maximise the sum of consumer surplus and inland revenue minus the costs for the greenhouse effect. Wirl [10] suggests further to consider a more sharp cost function, for instance costs that increase exponentially as it is modelled by the functional J.

In our setting, where only piecewise constant controls are considered, the explicit solution to the (discrete) **Problem** (P_K), $K \in \{I,J\}$,

$$\min_{\substack{u_0,\dots,u_N\\ x_0,\dots,x_N\\ y_0,\dots,y_N}} K(u_0,\dots,u_N,x_0,\dots,x_N,y_0,\dots,y_N)$$

such that for $k = 0,\dots,N-1$

$$x_{k+1} = x(t_{k+1};t_k,x_k,y_k,u_k)\,, \quad y_{k+1} = y(t_{k+1};t_k,x_k,y_k,u_k)\,, \\ u_k \in [0,M]\,, \quad M > 0 \text{ given,}$$

is not known, neither for $K = I$ nor in the case $K = J$. However, enlarging the control space to the set of piecewise continuous functions, at least the solution to the (continuous) Problem (P_I) can be determined by using the Hamilton-Jacobi-Bellman equation of continuous dynamic programming. We, therefore, refer to these different situations as the *discrete* and the *continuous* solution, respectively.

In all our experiments, the following parameter values have been used relying on [10]:

$x_0 = 2746$	$y_0 = 0$		for the initial condition,
$\zeta \in \{0; 0.0025\}$	$a = -0.04$	$b_1 = 12.5$	in the state equation (5.1a),
$\eta = 0.02$	$\xi = 0.0011$	$b_2 = -3$	in the state equation (5.1b),
$\alpha = 0.08$	$\beta = 15000$	$r = 0.03$	for the functional I,
$\alpha = 0.08$	$\gamma = 3.7066$	$r = 0.03$	for the functional J.

Note that Assumption (A4) is valid for the functional I. It remains fulfilled for the functional J if $\zeta > 0$. In the case $\zeta = 0$, we have $y(t) \le \bar{y}\, t + const$ where $\bar{y} = \xi(a\, u_N + b_1)$, and (with $\boldsymbol{\phi} = (x,y)^{\mathsf{T}}$) $\|\mathrm{D}_{\boldsymbol{\phi}\boldsymbol{\phi}} z(\boldsymbol{\phi})\| = \gamma^2 \mathrm{e}^{\gamma y(t)}$. Hence, in the case of an infinite time horizon, Assumption (A4) is also fulfilled if

$$u_N < \frac{r}{\gamma\, \xi\, a} - \frac{b_1}{a} \approx 128.553\,. \tag{5.3}$$

Using (5.2), the states $x(t)$, $y(t)$ have been calculated exactly. All integrals appearing in Algorithm 1.1 and 1.2, respectively, have been computed approximately using the composite Weddle rule with a step size $h = 0.1$. This integration formula is of order 8 and needs 7 evaluations on each subinterval of length h (cf. e. g. [8]).

The case of an infinite time horizon has been approximated using a (large) finite value for $T := t_{N+1}$. The corresponding problem $(P_K) = (P_{K(T)})$ admits a solution $u^* = u^{*,T}$ that converges to $u^{*,\infty}$ for $T \to \infty$.

A useful stopping criterion for both, the convergence of the iterates and the convergence in time,

$$u^{\ell,T} \to u^{*,T}\,, \qquad u^{*,T} \to u^{*,\infty}\,,$$

is to require a small residual R_ℓ,

$$\begin{aligned} R_\ell := \sum_{k=0}^{N} |\mathrm{D}_{u_k}\mathcal{L}_k| &+ \sum_{k=1}^{N} |\mathrm{D}_{x_k}\mathcal{L}_k| + \sum_{k=1}^{N} |\mathrm{D}_{y_k}\mathcal{L}_k| \\ &+ \sum_{k=0}^{N-1} |x_{k+1} - x(t_{k+1}; t_k, x_k, y_k, u_k)| \\ &+ \sum_{k=0}^{N-1} |y_{k+1} - y(t_{k+1}; t_k, x_k, y_k, u_k)|\,. \end{aligned}$$

Here, $\mathcal{L}_k$ has to be evaluated at the current iterates, i. e.

$$\mathcal{L}_k = \mathcal{L}_k(u_k^{(\ell)}, x_k^{(\ell)}, y_k^{(\ell)}, x_{k+1}^{(\ell)}, y_{k+1}^{(\ell)}, p_k^{(\ell)}, q_k^{(\ell)})\,,$$

where p_k, q_k are the multipliers. Again, for the approximate calculation of the integrals appearing in the definition of the residual, a finite time T has been used. Let us indicate this by denoting $R_\ell(T)$. The observation $R_\ell(T) \to 0$ then indicates convergence of the iterates. Furthermore, for

$T' \gg T$ the additional condition $R_\ell(T') \approx R_\ell(T)$ indicates that $u^{\ell,T}$ is a suitable approximation for $u^{*,T'}$ and $u^{*,\infty}$, respectively.

For problem (P_I), the equation determining u_k can be solved explicitly whereas for problem (P_J) we have employed Newtons' method in its standard form (with excellent convergence properties).

After all, the following has been observed:

(1) Problem (P_I) is solved fast and stable even for large T. For instance, with $N = 50$, $T = 2800$, $\tau_k = const$, $\zeta = 0$, Algorithm 1.2 takes 17 iterations to stop with a residual $R(T') < 0.001$ $(T' = 2T)$. As one expects, Algorithm 1.1 is less fast and needs 20 iterations to reach the same accuracy. Increasing the number of unknowns up to $N = 500$, Algorithm 1.2 needs *ceteris paribus* 6 iterations more. The values $I(u^{\ell,T})$ and $R_\ell(2T)$ decrease strictly.

(2) In accordance with the theoretical results, the rate of convergence is higher in the case $\lambda > 0$. Keeping the values from the previous item, we only have 13 iterations for a run with $\zeta = 0.025$ instead of $\zeta = 0$.

(3) The discrete solution approximates the continuous one quite well with respect to the states $x(t)$ and $y(t)$. Figure 5.1 shows the trajectories for different discretisations. Choosing large values for N, e. g. $N \geq 100$, there is no obvious difference to the continuous solution.

(4) The error with respect to the objective value of the continuous problem can be reduced significantly when using a time partition with constant distances $\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}}$. This decreases also the number of necessary iterations. This observation might be useful to find an optimal time partition when discretising time-continuous control problems. Also the estimates in the proofs of Theorem 3.1 and 3.2 as well as in Theorem 4.1 show that the term $\mathrm{e}^{-rt_k} - \mathrm{e}^{-rt_{k+1}}$ influences the convergence.

(5) Problem (P_J) cannot be solved for large values of T. Taking $N = 10$, $\zeta = 0$, and an equidistant time partition, Algorithm 1.2 does not converge for $T \geq 275$; the iterates are alternating. The residual evaluation of the short time solution shows that it is a bad approximation of large time problems. This behaviour does not change when increasing the number of unknowns since this does not resolve the trade-off between N and $\tau_{\max}$ stated in Theorem 3.2. However, it is possible to solve the problem with a larger discount rate r, but this does not lead to a solution of the original problem.

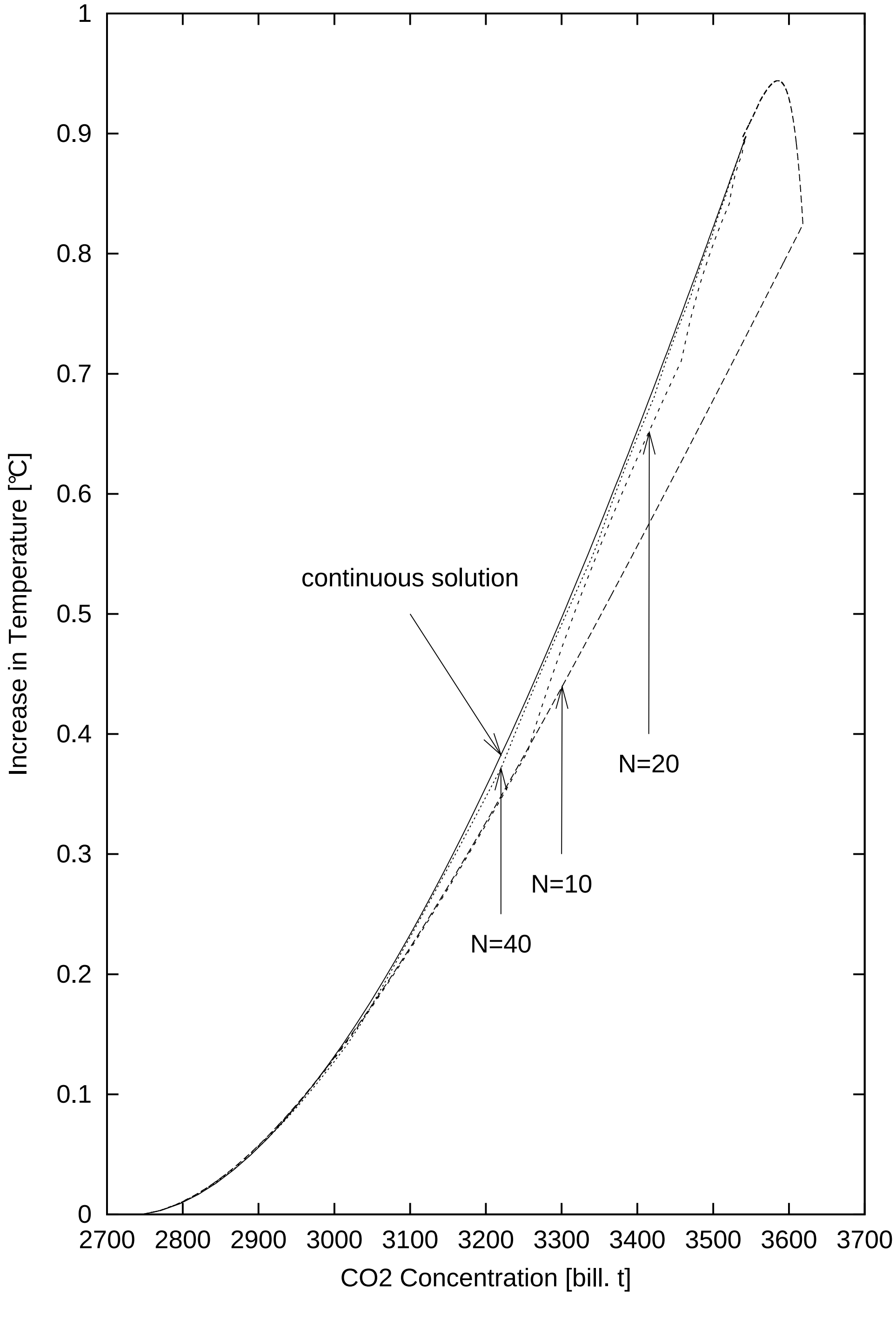

Figure 5.1: Continuous and discrete solutions for $\zeta > 0$ (equidistant time partition)

(6) Using the prox regularisation method (cf. [6]), i. e. in each iteration ℓ we deal with the regularised functionals

$$J_\ell := \sum_{k=0}^{N} \int_{t_k}^{t_{k+1}} \mathrm{e}^{-rt} \Big(\frac{\alpha}{2}\, u_k^2 + \frac{\sigma}{2}\, (u_k - u_k^{(\ell-1)})^2 + \mathrm{e}^{\gamma\, y(t;t_k,x_k,y_k,u_k)} \Big)\, dt$$

instead of J, where $\sigma > 0$ is some parameter chosen appropriately, leads to convergent sequences even for large T. The regularised functional has a curvature of at least $\alpha + \sigma$ that gives a better result in (4.1) and (4.2), where α can be replaced by $\alpha + \sigma$.

It should be noted that an additional condition for the control variables of the form $u_k \in \Omega_k \subset \mathbb{R}$, Ω_k closed, $(k = 0, \ldots, N)$, can easily be implemented by projecting the solution $u_k^{(\ell)}$ to (1.13) on Ω_k and using this projection instead of $u_k^{(\ell)}$.

6 Conclusions

In this work, we have presented two iterative algorithms for solving discrete discounted control problems as they typically arise in economic models. These problems may have an infinite time horizon but can be approximated with high accuracy by choosing large final time points. We have only considered the case of a control function that is piecewise constant on at most a finite number of time intervals. This situation might be given *a priori* or can arise from a time discretisation of a continuous model.

We have provided convergence results under suitable assumptions. These results show that the algorithms behave better for shorter time intervals, for larger discount rates, for particular partitions of the time interval, and for larger curvatures of the objective function. The convergence is at least linear.

Our numerical experiments are based upon a model of global warming that has been considered by Wirl [10]. The sequences generated by the iteration scheme converge very fast in the simple case of a quadratic objective function. We could also observe that the iterates do not converge in the case of a more general convex objective with a large final time, a problem that can be resolved by introducing a prox-regularisation.

We have, finally, to remark that the numerical experiments were successful although the convergence parameter ρ was not less than 1. This underlines that the theoretical results obtained here only present sufficient conditions for convergence and are restricted to situations as described in

Remark 3.1. Nevertheless, it is possible to find a sharper estimate for ρ in the case of our example when taking z's strong convexity into account for estimating the denominator in (3.4) using (2.5).

References

[1] H. Amann. *Gewöhnliche Differentialgleichungen.* W. de Gruyter, Berlin – New York, 1983.

[2] I. Capuzzo Dolcetta and H. Ishii. Approximate solutions of the Bellman equation of deterministic control theory. *Appl. Math. Optim.*, 11 (1984), pp. 161 – 181.

[3] M. Falcone. A numerical approach to the infinite horizon problem of deterministic control theory. *Appl. Math. Optim.*, 15 (1987), pp. 1 – 13.

[4] S. Felder, M. Meier, and H. Schmitt. Health care expenditure in the last months of life. *J. Health Econ.*, 19 (2000), pp. 679 – 695.

[5] L. Grüne. An adaptive grid scheme for the discrete Hamilton-Jacobi-Belman equation. *Numer. Math.*, 75 (1997), pp. 319 – 337.

[6] A. Kaplan and R. Tichatschke. *Stable Methods for Ill-Posed Variational Problems.* Akademie-Verlag, Berlin, 1994.

[7] J. Macki and A. Strauss. *Introduction to Optimal Control Theory.* Springer-Verlag, New York, 1982.

[8] J. Stoer. *Einführung in die Numerische Mathematik I.* Springer-Verlag, Berlin, 1983.

[9] P. Varaiya. *Notes on Optimization.* Van Nostrand Reinhold, New York, 1972.

[10] F. Wirl. Global warming and carbon taxes: Dynamic and strategic interactions between energy consumers and producers. *J. Policy Model.*, 16 (1994) 6, pp. 577 – 596.