TECHNISCHE UNIVERSITÄT WIEN

Operations Research and Control Systems
SWM ORCOS

Gradient Methods on Strongly Convex Feasible Sets and Optimal Control of Affine Systems

Vladimir Veliov, Phan Tu Vuong

Research Report 2017-11

November 2017

ISSN 2521-313X

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

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

Gradient Methods on Strongly Convex Feasible Sets and Optimal Control of Affine Systems*

V.M. Veliov† P.T. Vuong‡

Abstract

The paper presents new results about convergence of the gradient projection and the conditional gradient methods for abstract minimization problems on strongly convex sets. In particular, linear convergence is proved, although the objective functional does not need to be convex. Such problems arise, in particular, when a recently developed discretization technique is applied to optimal control problems which are affine with respect to the control. This discretization technique has the advantage to provide higher accuracy of discretization (compared with the known discretization schemes) and involves strongly convex constraints and possibly non-convex objective functional. The applicability of the abstract results is proved in the case of linear-quadratic affine optimal control problems, and error estimates are obtained. A numerical example is given, confirming the theoretical findings.

Key words: optimal control, mathematical programming, numerical methods, gradient methods, affine control systems, bang-bang control

AMS subject classifications: 49M25, 90C25, 90C48, 49M37.

1 Introduction

Solving numerically optimal control problems in which the control function appears linearly, and performing error analysis, are still challenging issues due to the typical discontinuity of the optimal control. Considerable progress was made in the past decade in the analysis of discretization schemes in combination with various methods of solving the resulting discrete-time optimization problems. The papers [26, 1, 24, 2] apply to problems with linear dynamics, while [10, 3] address nonlinear affine (in the control) dynamics. Usually the discretization is performed by Runge-Kutta schemes (mainly the Euler scheme) and the accuracy is at most of first order due to the discontinuity of the optimal control. Discretization schemes of higher accuracy were recently proposed in [19, 23]

*This research is supported by the Austrian Science Foundation (FWF) under grant No P26640-N25.

†Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Austria, vladimir.veliov@tuwien.ac.at, tel. +43 1 58 801 10540, Fax: +43-1-58801-10599.

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

for systems with linear dynamics and Mayer or Bolza problems. In both cases the error analysis is based on the assumption that the optimal control is of purely bang-bang type.

On the other hand, the papers [11, 20] present convergence results for a version of the (abstract) Newton method for nonlinear problems, affine with respect to the control. Every step of the Newton method requires solving a linear-quadratic (affine in the control) optimal control problem for a linear system, namely a problem of the following type:

$$\underset{x,u}{\text{minimize}}\ J(x,u) := \frac{1}{2}x(T)^\top Q x(T) + q^\top x(T) + \int_0^T \left(\frac{1}{2}x(t)^\top W(t)x(t) + x(t)^\top S(t)u(t)\right) \mathrm{d}t. \quad (1)$$

subject to

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) + d(t), \quad x(0) = x_0, \quad t \in [0,T], & (2)\\ & u(t) \in U := [-1,1]^m. & (3) \end{aligned}$$

Here, $[0,T]$ is a fixed time horizon, $A(t), W(t) \in \mathbb{R}^{n\times n}$, $B(t), S(t) \in \mathbb{R}^{n\times m}$ for every $t \in [0,T]$, the superscript $\top$ means transposition. Admissible controls are all measurable functions $u : [0,T] \to U$. The state of the system at time t is $x(t) \in \mathbb{R}^n$, where $x(\cdot)$ is the (absolutely continuous) solution of (2), given an admissible control $u(\cdot)$. Linear terms are not included in the integrand in (1), since they can be shifted in a standard way into the differential equation (2).

For solving the above problem one can apply the high-order discretization scheme developed in [19, 23]. It results in a discrete-time optimal control problem (a mathematical programming problem), where the gradient of the objective function can be calculated following a standard procedure involving the solution of the associated adjoint system, so that gradient-type methods are conveniently applicable. And here we encounter a remarkable fact: although neither the objective functional (1) of the continuous-time problem (1)–(3) nor the control constraints (3) are strongly convex, it turns out that the feasible set of the discretized problem is strongly convex. This brings into consideration the issue of convergence of gradient methods for problems with strongly convex feasible sets and possibly non-convex objective functions (even if the functional J in (1) is convex on the set of admissible control-trajectory pairs, the discretized problem may fail to be convex!).

Versions of the Gradient Projection Method (GPM) and the Conditional Gradient Method (CGM) are widely studied (see e.g. [17, 18] and the references therein), but results about linear convergence of the generated sequence of iterates seem to be available only for problems with strongly convex objective functions. Exceptions are the papers [5, 14], where strong convexity is assumed for the feasible set instead of the objective function. However, as clarified in the end of Subsection 2.1 below, the additional assumptions in these two papers are rather strong and are not fulfilled for the problem arising in the optimal control context as described above.

In this paper we present convergence results for the gradient projection and the conditional gradient methods for minimization problems in a Hilbert space, where the feasible set is strongly convex but the objective functional is not necessarily convex. These results are new even for convex or strongly convex objective functional, but we relax the convexity assumption due to the needs of our main

goal – to cover the problems arising in optimal control of affine systems, as described above. For that we consider objective functionals that we called, for shortness, (ε, δ)-approximately convex. These functions constitute a larger class than that of the weakly convex functions (see e.g. [4]). In Subsection 2.1 we prove linear convergence of the sequence of approximate solutions generated by the GPM, provided that the step sizes are appropriately chosen. Apart from the applicability for non-convex objective functionals, this result does not require the additional conditions in [5, 14]. As usual, the "appropriate" choice of the step sizes is expressed by some constants related to the data of the problem, which are often not available (or very roughly estimated). Therefore, we present an additional convergence result involving a rather general and constructive condition for the step sizes (well-known in the literature).

The conditional gradient method may have some advantages (compared with the GPM) in our optimal control application. For this reason we also prove a linear convergence result for the CGM. This is done in Subsection 2.2.

In Section 3 we turn back to the optimal control problem (1)–(3). The first two subsections are preliminary, where we introduce notations, formulate assumptions and present the discrete approximation introduced in [19, 23] and the error estimate proved in [23]. All this is needed for understanding of the implementation of the GPM and the CGM and of the proofs of the error estimations. Then, in subsections 3.3 and 3.4 we prove the applicability of the abstract convergence results, obtained in Section 2, to our discretized optimal control problem and present details about the implementation of the GPM and the CGM. A numerical example that confirms the theoretical findings is given in Subsection 3.5.

The paper concludes with indication of some open problems for further research (Section 4).

2 Gradient methods for problems with strongly convex feasible set

In this section we investigate the convergence of certain gradient methods for an abstract minimization problem of the form

$$\min_{w \in K} f(w), \tag{4}$$

where K is a convex subset of a real Hilbert space H and $f : H \to \mathbb{R}$ is a function for which certain conditions weaker than convexity will be posed. Convergence results for gradient projection methods for this problem in finite dimensional spaces and convex f are known (see e.g. [18]). It has been proved that the iterative sequence generated by versions of the gradient projection method converges linearly to a solution, provided that the objective function f is strongly convex and its gradient is Lipschitz continuous. Extensions to infinite dimensional Hilbert spaces are straightforward. In contrast, in our results below the function f does not need even to be convex, while the set K is assumed strongly convex. Some convergence results for smooth convex functions f and strongly convex sets K are obtained in [5, 14], but under suppositions that (apart from the convexity of f) are not satisfied in our main motivation as described in the introduction (see

Remark 2.3 below). The convergence results presented in this section are substantially stronger.

As usual, $\langle \cdot, \cdot \rangle$ denotes the inner product in H and $\| \cdot \|$ – the induced norm.

Let K be a nonempty closed convex subset of H. For each $u \in H$, there exists a unique point in K (see [15, p. 8]), denoted by $P_K(u)$, such that

$$\|u - P_K(u)\| \leq \|u - v\| \quad \forall v \in K.$$

It is well-known that the metric projection P_K is a nonexpansive mapping, i.e., for all $u, v \in H$

$$\|P_K(u) - P_K(v)\| \leq \|u - v\|.$$

Moreover for any $u \in H$ and $v \in K$, it holds that

$$\langle u - P_K(u), v - P_K(u) \rangle \leq 0. \tag{5}$$

Conversely, if $w \in K$ and $\langle u - w, v - w \rangle \leq 0$ for all $v \in K$, then $w = P_K(u)$.

Below we remind the following notions.

Definition 2.1 The set $K \subset H$ is called *strongly convex* or γ-strongly convex if there exists a number $\gamma > 0$ (called modulus of strong convexity) such that for any $u, v \in K$ and any $\lambda \in [0, 1]$ it holds that

$$\lambda u + (1 - \lambda)v + \lambda(1 - \lambda)\frac{\gamma}{2}\|u - v\|^2 z \in K \qquad \forall\, z \text{ with } \|z\| \leq 1.$$

Definition 2.2 A function $f : H \to \mathbb{R}$ is called *L-smooth* on K if f is Fréchet differentiable and its derivative, ∇f, is L-Lipschitz continuous on K, i.e.,

$$\|\nabla f(u) - \nabla f(v)\| \leq L\|u - v\| \qquad \forall\, u, v \in K.$$

The following definition introduces a property that is usually called "weak convexity" or "para-convexity" (see e.g. [4]).

Definition 2.3 A function $f : H \to \mathbb{R}$ is called *ε-convex* (with $\varepsilon \geq 0$) on a convex subset $K \subset H$ at $\hat{w} \in K$ if the function $f_\varepsilon(w) := f(w) + \frac{1}{2}\varepsilon\|w - \hat{w}\|^2$ is convex on K at $\hat{w}$, i.e.

$$f_\varepsilon(\alpha w + (1 - \alpha)\hat{w}) \leq \alpha f_\varepsilon(w) + (1 - \alpha) f_\varepsilon(\hat{w})$$

for every $w \in K$ and $\alpha \in (0, 1)$.

If $f : H \to \mathbb{R}$ is ε-convex at $\hat{w}$ and differentiable, then

$$\langle \nabla f_\varepsilon(w) - \nabla f_\varepsilon(\hat{w}), w - \hat{w} \rangle \geq 0 \qquad \forall\, w \in K.$$

This implies that

$$\langle \nabla f(w) - \nabla f(\hat{w}), w - \hat{w} \rangle \geq -\varepsilon\|w - \hat{w}\|^2 \qquad \forall\, w \in K.$$

In the our main application, the function f does not need to be even ε-convex with ε reasonably small. Therefore we further weaken the convexity as in the following definition.

Definition 2.4 A Fréchet-differentiable function $f : H \to \mathbb{R}$ is called (ε, δ)-*approximately convex* (with $\varepsilon, \delta \geq 0$) on a convex subset $K \subset H$ at $\hat{w} \in K$ if

$$\langle \nabla f(w) - \nabla f(\hat{w}), w - \hat{w} \rangle \geq -\varepsilon \|w - \hat{w}\|^2 \qquad \forall\, w \in K \text{ with } \|w - \hat{w}\| \geq \delta. \tag{6}$$

Notice that δ can be taken equal to zero in the above definition, in which case the (ε, δ)-approximate convexity reduces to ε-convexity.

The following three results provide the ground for the error analysis of the GPM and the CGM.

Proposition 2.1 *Assume that f is L-smooth, K is γ-strongly convex and $\hat{w} \in K$ is a solution of problem (4) such that $\|\nabla f(\hat{w})\| \geq \rho$ for some number $\rho > 0$. Assume also that f is (ε, δ)-approximately convex on K at $\hat{w}$ and that the number $\nu := \frac{\gamma\rho}{4} - \varepsilon$ is positive. Then*

$$\langle \nabla f(w), w - \hat{w} \rangle \geq \nu \|w - \hat{w}\|^2 \qquad \forall\, w \in K \text{ with } \|w - \hat{w}\| \geq \delta. \tag{7}$$

Moreover, any solution of problem (4) is at distance at most δ from $\hat{w}$.

Proof. Setting $z = \frac{-\nabla f(\hat{w})}{\|\nabla f(\hat{w})\|}$, we have $\|z\| = 1$. By the strong convexity of K we obtain that for any $w \in K$

$$y := \frac{1}{2}(w + \hat{w}) + \frac{\gamma}{8}\|w - \hat{w}\|^2 z \in K.$$

Due to (6), for all $w \in K$ with $\|w - \hat{w}\| \geq \delta$ we have

$$\langle \nabla f(w) - \nabla f(\hat{w}), w - \hat{w} \rangle \geq -\varepsilon \|w - \hat{w}\|^2.$$

Hence,

$$\begin{aligned} \langle \nabla f(w), w - \hat{w} \rangle &\geq \langle \nabla f(\hat{w}), w - \hat{w} \rangle - \varepsilon \|w - \hat{w}\|^2 \\ &= 2\left\langle \nabla f(\hat{w}), \frac{w + \hat{w}}{2} - y \right\rangle + 2\langle \nabla f(\hat{w}), y - \hat{w} \rangle - \varepsilon \|w - \hat{w}\|^2. \end{aligned} \tag{8}$$

The optimality of $\hat{w}$ implies that

$$\langle \nabla f(\hat{w}), y - \hat{w} \rangle \geq 0.$$

Then from (8) we obtain that

$$\begin{aligned} \langle \nabla f(w), w - \hat{w} \rangle &\geq 2\left\langle \nabla f(\hat{w}), \frac{\gamma}{8}\|w - \hat{w}\|^2 \frac{\nabla f(\hat{w})}{\|\nabla f(\hat{w})\|} \right\rangle - \varepsilon \|w - \hat{w}\|^2 \\ &= \frac{\gamma}{4}\|\nabla f(\hat{w})\| \|w - \hat{w}\|^2 - \varepsilon \|w - \hat{w}\|^2 \geq \nu \|w - \hat{w}\|^2, \end{aligned}$$

that is, (7).

Now assume that $\bar{w}$ is another solution of (4). The optimality of $\bar{w}$ implies, in particular, that

$$\langle \nabla f(\bar{w}), \hat{w} - \bar{w} \rangle \geq 0.$$

Assuming that $\|\bar{w} - \hat{w}\| > \delta$ we may substitute $w = \bar{w} \in K$ in (7), which gives

$$\langle \nabla f(\bar{w}), \bar{w} - \hat{w} \rangle \geq \nu \|\bar{w} - \hat{w}\|^2.$$

Adding the last two inequalities we obtain that

$$0 \geq \nu \|\bar{w} - \hat{w}\|^2.$$

which contradicts the assumption $\|\bar{w} - \hat{w}\| > \delta$. The proof is complete. Q.E.D.

Lemma 2.1 *Let the assumptions of Proposition 2.1 be satisfied. If for some $w \in K$ and $\lambda > 0$ it holds that $P_K(w - \lambda \nabla f(w)) = w$, then $\|w - \hat{w}\| \leq \delta$.*

Proof. Contrary to the claim of the lemma, assume that $\|w - \hat{w}\| > \delta$. Then from Proposition 2.1 we have that the first inequality in (7) is fulfilled by w. From the condition $P_K(w - \lambda \nabla f(w)) = w$ we have that

$$\langle \nabla f(w), u - w \rangle \geq 0 \qquad \forall u \in K.$$

Applying this inequality for $u = \hat{w}$ and adding it to the first inequality in (7) we obtain that

$$0 \geq \nu \|w - \hat{w}\|^2,$$

which is a contradiction. Q.E.D.

Lemma 2.2 *Let the assumptions of Proposition 2.1 be satisfied. If for some $w \in K$ if holds that $\nabla f(w) = 0$, then $\|w - \hat{w}\| \leq \delta$.*

Proof. If we assume $\|w - \hat{w}\| > \delta$, then from the first inequality in (7) we have

$$0 \geq \nu \|w - \hat{w}\|^2,$$

which is a contradiction. Q.E.D.

2.1 The gradient projection method

For solving the minimization problem (4), we consider first the most classical algorithm, the Gradient Projection Method (GPM) stated below. In the formulation of the algorithm we only assume that f is L-smooth.

Algorithm GPM.

Step 0: Choose $w_0 \in K$. Set $k = 0$.

Step 1: If $w_k = P_K\left(w_k - \nabla f(w_k)\right)$ then Stop. Otherwise, go to Step 2.

Step 2: Choose $\lambda_k > 0$ and calculate

$$w_{k+1} = P_K\left(w_k - \lambda_k \nabla f(w_k)\right). \tag{9}$$

Replace k by $k+1$; go to Step 1.

It is well-known that for convex f and K the GPM has the error estimate $O(\frac{1}{k})$ in term of the objective function when $\lambda_k = \lambda \in \left(0, \frac{1}{L}\right]$, see e.g. [6]. More precisely, if problem (4) has a solution and $\hat{f}$ is the minimal value of f on K, then

$$f(w_k) - \hat{f} \leq \frac{Lm_0}{2k} \qquad \forall k,$$

where m_0 is the distance from w_0 to the solution set of (4). If in addition, f is strongly convex, then the sequence $\{w_k\}$ converges to the unique solution of (4). If f is only convex (but not necessarily strongly convex), there are no convergence results, in the known to us literature, concerning the iterative sequence $\{w_k\}$.

In this subsection, we prove that if the set K is strongly convex and the function f is (ε, δ)-approximately convex then the sequence $\{w_k\}$ generated by the GPM linearly approaches $\hat{w}$ at least until entering a δ-neighborhood of $\hat{w}$. We mention that if the above algorithm of the GPM stops at Step 1 for some k then, according to Lemma 2.1, $\|w_k - \hat{w}\| \leq \delta$, that is, an approximate solution is attained.

Using Proposition 2.1, we obtain the following main estimation which will be repeatedly used in the sequel.

Proposition 2.2 *Let all the assumptions in Proposition 2.1 be satisfied, and let $\|w_0 - \hat{w}\| \geq \delta$. Then the sequence $\{w_k\}$ generated by the GPM satisfies the inequality*

$$\left[1 + \lambda_k\left(2\nu - \lambda_k L^2\right)\right]\|w_{k+1} - \hat{w}\|^2 \leq \|w_k - \hat{w}\|^2 \quad \forall k \tag{10}$$

at least as long as $\|w_{k+1} - \hat{w}\| \geq \delta$.

Proof. Since $w_{k+1} = P_K(w_k - \lambda_k \nabla f(w_k))$, due inequality (5) we have

$$\langle w_k - \lambda_k \nabla f(w_k) - w_{k+1}, w - w_{k+1}\rangle \leq 0 \quad \forall w \in K.$$

Substitution of $w = \hat{w} \in K$ in this inequality yields

$$\langle w_k - \lambda_k \nabla f(w_k) - w_{k+1}, \hat{w} - w_{k+1}\rangle \leq 0,$$

or equivalently

$$\begin{aligned} &2\langle w_k - w_{k+1}, \hat{w} - w_{k+1}\rangle \leq 2\lambda_k \langle \nabla f(w_k), \hat{w} - w_{k+1}\rangle \\ &= -2\lambda_k \langle \nabla f(w_{k+1}), w_{k+1} - \hat{w}\rangle + 2\lambda_k \langle \nabla f(w_k) - \nabla f(w_{k+1}), \hat{w} - w_{k+1}\rangle. \end{aligned} \tag{11}$$

Since $w_{k+1} \in K$ and $\lambda_k > 0$, if $\|w_{k+1} - \hat{w}\| \geq \delta$ then due to Proposition 2.1

$$-2\lambda_k \langle \nabla f(w_{k+1}), w_{k+1} - \hat{w} \rangle \leq -2\lambda_k \nu \|w_{k+1} - \hat{w}\|^2. \tag{12}$$

By the Cauchy-Schwarz inequality and the Lipschitz continuity of ∇f, we obtain that

$$\begin{aligned} 2\lambda_k \langle \nabla f(w_k) - \nabla f(w_{k+1}), \hat{w} - w_{k+1} \rangle &\leq 2\lambda_k \|\nabla f(w_k) - \nabla f(w_{k+1})\| \|w_{k+1} - \hat{w}\| \\ &\leq 2\lambda_k L \|w_k - w_{k+1}\| \|w_{k+1} - \hat{w}\| \\ &\leq \|w_k - w_{k+1}\|^2 + (\lambda_k L)^2 \|w_{k+1} - \hat{w}\|^2. \end{aligned} \tag{13}$$

Inequalities (11), (12) and (13) imply that

$$2\langle w_k - w_{k+1}, \hat{w} - w_{k+1} \rangle \leq -2\lambda_k \nu \|w_{k+1} - \hat{w}\|^2 + \|w_k - w_{k+1}\|^2 + (\lambda_k L)^2 \|w_{k+1} - \hat{w}\|^2. \tag{14}$$

On the other hand,

$$\begin{aligned} 2\langle w_k - w_{k+1}, \hat{w} - w_{k+1} \rangle &= \|w_k - w_{k+1}\|^2 + \|\hat{w} - w_{k+1}\|^2 - \|(w_k - w_{k+1}) - (\hat{w} - w_{k+1})\|^2 \\ &= \|w_k - w_{k+1}\|^2 + \|w_{k+1} - \hat{w}\|^2 - \|w_k - \hat{w}\|^2. \end{aligned} \tag{15}$$

Combining (14) and (15) we obtain that

$$\|w_k - w_{k+1}\|^2 + \|w_{k+1} - \hat{w}\|^2 - \|w_k - \hat{w}\|^2 \leq -2\lambda_k \nu \|w_{k+1} - \hat{w}\|^2 + \|w_k - w_{k+1}\|^2 + (\lambda_k L)^2 \|w_{k+1} - \hat{w}\|^2,$$

hence (10) is satisfied. Q.E.D.

Now we can state and prove the main convergence result for the GPM.

Theorem 2.1 *Let all the assumptions in Proposition 2.2 be satisfied. Let the sequence $\{\lambda_k\}$ be chosen such that*

$$0 < a \leq \lambda_k \leq b < \frac{2\nu}{L^2} \quad \forall k, \tag{16}$$

where a, b are some positive constants. Define

$$\mu = \frac{1}{\sqrt{1 + a\,(2\nu - bL^2)}} \in (0, 1). \tag{17}$$

Let $\{w_k\}$ be the sequence generated by the GPM. Then for every k, if $\|w_{k+1} - \hat{w}\| \geq \delta$ then

$$\|w_{k+1} - \hat{w}\| \leq \mu \, \|w_k - \hat{w}\|. \tag{18}$$

Moreover, for every k, if $\|w_{i+1} - \hat{w}\| \geq \delta$, $i = 0, \ldots, k$, then the following a priori and a posteriori error estimates hold:

$$\|w_{k+1} - \hat{w}\| \leq \frac{\mu^{k+1}}{1 - \mu} \|w_1 - w_0\|, \tag{19}$$

and

$$\|w_{k+1} - \hat{w}\| \leq \frac{\mu}{1 - \mu} \|w_{k+1} - w_k\|. \tag{20}$$

Before proving the theorem we mention that in the case of an ε-convex function f (that is, if $\delta = 0$) the first claim of the theorem means that the sequence generated by the GPM converges linearly to the (unique) solution $\hat{w}$. In the case $\delta > 0$ we also have linear convergence at least until the generated sequence enters the δ-neighborhood of $\hat{w}$. Thus in this case the theorem is meaningful only if δ is reasonably small.

Proof. It follows from (16) that $\left[1 + \lambda_k \left(2\nu - \lambda_k L^2\right)\right] \geq \left[1 + a \left(2\nu - bL^2\right)\right] > 1$ for all k. By (10) and the above inequalities,

$$\left[1 + a \left(2\nu - bL^2\right)\right] \|w_{k+1} - \hat{w}\|^2 \leq \|w_k - \hat{w}\|^2,$$

provided that $\|w_{k+1} - \hat{w}\| \geq \delta$. Hence

$$\|w_{k+1} - \hat{w}\| \leq \mu \|w_k - \hat{w}\| \tag{21}$$

with $\mu \in (0, 1)$ being defined by (17).

The proof of (19) and (20) is standard, but we present it for completeness. By (21),

$$\|w_{k+1} - \hat{w}\| \leq \mu \|w_k - \hat{w}\| \leq \mu^2 \|w_{k-1} - \hat{w}\| \leq \ldots \leq \mu^{k+1} \|w_0 - \hat{w}\|.$$

Observe that

$$\|w_k - \hat{w}\| \leq \|w_k - w_{k+1}\| + \|w_{k+1} - \hat{w}\| \leq \|w_k - w_{k+1}\| + \mu \|w_k - \hat{w}\|,$$

and so $\|w_k - \hat{w}\| \leq \frac{1}{1-\mu} \|w_k - w_{k+1}\|$ for all k. Hence

$$\|w_{k+1} - \hat{w}\| \leq \mu^{k+1} \|w_0 - \hat{w}\| \leq \frac{\mu^{k+1}}{1-\mu} \|w_0 - w_1\|,$$

$$\|w_{k+1} - \hat{w}\| \leq \mu \|w_k - \hat{w}\| \leq \frac{\mu}{1-\mu} \|w_k - w_{k+1}\|.$$

Q.E.D.

Remark 2.1 If the constants L, γ and ρ can be reasonably estimated, then inequalities (19) and (20) can be used to estimate the number of iterations of the GPM needed to achieve a given accuracy.

Remark 2.2 The value μ in (17) can be regarded as a function $\mu = \mu(a, b)$ of the variable (a, b) belonging to the domain

$$\left\{(a, b) \in \mathbb{R}^2 : 0 < a \leq b < \frac{2\nu}{L^2}\right\}.$$

It is a routine task to obtain that the minimum of $\mu(a, b)$ under the above constraints is achieved at $(a_*, b_*) := (\frac{\nu}{L^2}, \frac{\nu}{L^2})$ and the minimal value is $\mu_* := \frac{L}{\sqrt{L^2+\nu^2}}$. Hence, $\lambda_k = \frac{\nu}{L^2}$ would be an optimal choice of λ_k.

Since the parameters γ, ρ and L are usually not known in advance, we can consider the step size sequence $\{\lambda_k\}$ as any non-summable converging to zero sequence of positive real numbers as it follows in the next theorem.

Theorem 2.2 *Let the assumptions in Proposition 2.2 be satisfied. Let $\{\lambda_k\}$ be a sequence of positive scalars such that*

$$\sum_{k=0}^{\infty} \lambda_k = +\infty, \quad \lim_{k\to\infty} \lambda_k = 0. \tag{22}$$

Then for every $\delta' > \delta$ all elements of the sequence $\{w_k\}$ with sufficiently large k are contained in the δ'-neighborhood of $\hat{w}$. Moreover, there exists a natural number k_0 such that for each $k \geq k_0$ for which $\|w_{i+1} - \hat{w}\| \geq \delta$ is fulfilled for $i = k_0, \ldots, k$, it holds that $\lambda_k(2\nu - \lambda_k L^2) > 0$, and

$$\|w_{k+1} - \hat{w}\| \leq \frac{1}{\sqrt{\prod_{i=k_0}^{k}[1 + \lambda_i(2\nu - \lambda_i L^2)]}} \|w_{k_0} - \hat{w}\|. \tag{23}$$

Clearly, in the case $\delta = 0$ the first claim of the theorem implies strong convergence of the sequence $\{w_k\}$.

Proof. Since $\lambda_k \to 0$, there exists k_0 such that $4\lambda_k L^2 < \gamma\rho$ for every $k \geq k_0$. Hence,

$$\lambda_k \left(2\nu - \lambda_k L^2\right) > \lambda_k \left(2\nu - \nu\right) = \nu\lambda_k > 0,$$

for all $k \geq k_0$. If k is such that $\|w^{i+1} - \hat{w}\| \geq \delta$, $i = k_0, \ldots, k$. Then from (10) it follows that

$$\begin{aligned}
\|w_{k+1} - \hat{w}\|^2 &\leq \frac{1}{1 + \lambda_k \left(2\nu - \lambda_k L^2\right)} \|w_k - \hat{w}\|^2 \\
&\leq \frac{1}{[1 + \lambda_k \left(2\nu - \lambda_k L^2\right)]} \frac{1}{[1 + \lambda_{k-1}(2\nu - \lambda_{k-1} L^2)]} \|w_{k-1} - \hat{w}\|^2 \\
&\vdots \\
&\leq \frac{1}{\prod_{i=k_0}^{k}[1 + \lambda_i(2\nu - \lambda_i L^2)]} \|w_{k_0} - \hat{w}\|^2,
\end{aligned}$$

which proves (23).

Let us now prove the first claim of the theorem. For each k set

$$\alpha_k = \lambda_k \left(2\nu - \lambda_k L^2\right)$$

and rewrite (23) (if it holds for k) as

$$\|w_{k+1} - \hat{w}\| \leq \frac{1}{\sqrt{\prod_{i=k_0}^{k}(1 + \alpha_i)}} \|w_{k_0} - \hat{w}\|. \tag{24}$$

Since $\alpha_k = \lambda_k(2\nu - \lambda_k L^2) > \nu\lambda_k$ for each $k \geq k_0$, it follows from (22) that $\sum_{k=k_0}^{\infty} \alpha_k = +\infty$. Hence

$$\prod_{i=k_0}^{k} (1 + \alpha_i) \geq 1 + \sum_{i=k_0}^{k} \alpha_i \longrightarrow +\infty$$

as $k \to \infty$. Since (24) holds as long as $\|w_{k+1} - \hat{w}\| \geq \delta$, we obtain that either $\|w_k - \hat{w}\| \longrightarrow 0$ or $\|w_k - \hat{w}\| < \delta$ for some $k \geq k_0$. In the second case we either have $\|w_{k+1} - \hat{w}\| < \delta$, or $\|w_{k+1} - \hat{w}\| \geq \delta$. In the second case, again, we have from (10)

$$\|w_{k+1} - \hat{w}\|^2 \leq \frac{1}{1+\alpha_k}\|w_k - \hat{w}\|^2 \leq \delta^2.$$

Thus w_k remains in the δ-neighborhood of $\hat{w}$ for all k. The proof is complete. Q.E.D.

Remark 2.3 Using the contractivity of the projection onto strongly convex sets, Balashov and Golubev [5] and Golubev [14] obtained the linear convergence of the GPM for smooth, convex optimization problem with the following additional conditions:

(i) For any k, there exists a unit vector $n(w_k) \in N_K(w_k)$ such that

$$\langle n(w_k), \nabla f(w_k) \rangle \leq 0,$$

i.e., $w_k - \lambda_k \nabla f(w_k) \notin K$ for any $\lambda_k > 0$.

(ii) The problem (4) has a unique solution and it belongs to the boundary of K.

In our convergence analysis in Theorem 2.1, the assumptions (i), (ii) are eliminated, which is important for our main motivation (see the next section). Also important is that our result applies under the (ε, δ)-approximate convexity instead of convexity.

2.2 The conditional gradient method

In this subsection, we consider the Conditional Gradient Method (CGM) for solving problem (4) with a γ-strongly convex set K and an (ε, δ)-approximate convex and L-smooth function f. This method dates back to the original work of Frank and Wolfe [12] which presented an algorithm for minimizing a quadratic function over a polytope using only linear optimization steps over the feasible set. The CGM for solving (strongly) convex problem was investigated in [7, 8, 13].

Algorithm CGM.

Step 0: Choose $w_0 \in K$. Set $k = 0$.

Step 1: If $\nabla f(w_k) = 0$, then Stop. Otherwise, find a solution x_k of the problem

$$\min_{y \in K} \langle \nabla f(w_k), y \rangle . \tag{25}$$

Step 2: If $x_k = w_k$, then Stop. Otherwise, go to Step 3.

Step 3: If $\nabla f(w_k) \neq 0$, choose $\eta_k \in \left(0, \min\left\{1, \frac{\gamma \|\nabla f(w_k)\|}{4L}\right\}\right]$, calculate

$$w_{k+1} = (1 - \eta_k) w_k + \eta_k x_k, \tag{26}$$

replace k by $k+1$, and go to Step 1. Else the iteration process terminates.

Notice that if the above algorithm stops at Step 1 or Step 3 for some k then, according to Lemma 2.2, $\|w_k - \hat{w}\| \leq \delta$, that is, an approximate solution is attained.

In general, problem (25) may fail to have a solution, in which case the CGM is not executable.

Remark 2.4 The objective function in the subproblem (25) in the CGM is linear, thus if K is a polytope, we encounter a linear programming problem which should be easier to solve than the quadratic programming subproblem (9) in the GPM. In the case considered in this paper the set K is strongly convex, thus (25) is not a linear programming problem. However, in our main application (see the next section) the set K is a product of (possibly large number of) simple two-dimensional strongly convex sets, so that (25) decomposes into two-dimensional subproblems that are easy to solve.

We will use the following global version of (ε, δ)-approximate convexity.

Definition 2.5 A Fréchet-differentiable function $f : H \to \mathbb{R}$ is called (ε, δ)-*approximately convex* on a convex subset $K \subset H$ if

$$f(w) - f(v) \geq \langle \nabla f(v), w - v \rangle - \frac{\varepsilon}{2}\|w - v\|^2 \qquad \forall\, w, v \in K \text{ with } \|w - v\| \geq \delta. \tag{27}$$

Clearly, (27) implies (6).

We begin the convergence analysis of the CGM with an inequality which will play a key role for obtaining convergence results. For convenience we assume that if the CGM terminates at some finite iteration $k = i$, (due to $\nabla f(w_i) = 0$) then the sequence $\{w_k\}$ is extended as $w_k = w_i$ for $k > i$.

Proposition 2.3 *Assume that K is γ-strongly convex, f is L-smooth and $\hat{w}$ is a solution of problem (4) such that $\|\nabla f(\hat{w})\| \geq \rho$ for some number $\rho > 0$. Assume also that f is (ε, δ)-approximately convex on K and that the number $\nu := \frac{\gamma\rho}{4} - \epsilon$ is positive. Further, assume that at any iteration k a solution of the subproblem (25) does exist, and let $\{w_k\}$ be the sequence generated by the CGM. Denote $\hat{f} := f(\hat{w})$ and $\Delta_k := f(w_k) - \hat{f}$. Then*

$$\Delta_{k+1} \leq \left(1 - \frac{\nu\eta_k}{2\nu + \epsilon}\right)\Delta_k - \frac{\eta_k}{2}\left(\frac{\gamma\|\nabla f(w_k)\|}{4} - L\eta_k\right)\|x_k - w_k\|^2, \tag{28}$$

at least as long as $\|w_k - \hat{w}\| \geq \delta$.

Proof. If $\nabla f(w_i) = 0$ for some i, we have $x_k = w_k$ and $\Delta_k = 0$ for all $k \geq i$, hence (28). Thus we may assume that $\nabla f(w_k) \neq 0$ for the arbitrarily fixed k in the consideration below.

Since f is L-smooth we have

$$\begin{aligned} f(w_{k+1}) &\leq f(w_k) + \langle \nabla f(w_k), w_{k+1} - w_k \rangle + \frac{L}{2}\|w_{k+1} - w_k\|^2 \\ &= f(w_k) + \eta_k \langle \nabla f(w_k), x_k - w_k \rangle + \frac{L}{2}\eta_k^2 \|x_k - w_k\|^2. \end{aligned} \tag{29}$$

Subtracting $\hat{f}$ from both sizes of (29), we obtain

$$\Delta_{k+1} \leq \Delta_k + \eta_k \langle \nabla f(w_k), x_k - w_k \rangle + \frac{L}{2}\eta_k^2 \|x_k - w_k\|^2. \tag{30}$$

By the optimality of x_k in (25), we have

$$\langle \nabla f(w_k), x_k \rangle \leq \langle \nabla f(w_k), \hat{w} \rangle . \tag{31}$$

Assume from now on that $\|w_k - \hat{w}\| \geq \delta$. From (31) and the (ε, δ)-approximate convexity of f it follows that

$$\begin{aligned} \langle \nabla f(w_k), x_k - w_k \rangle &\leq \langle \nabla f(w_k), \hat{w} - w_k \rangle \\ &\leq f(\hat{w}) - f(w_k) + \frac{\epsilon}{2}\|w_k - \hat{w}\|^2 = -\Delta_k + \frac{\epsilon}{2}\|w_k - \hat{w}\|^2. \end{aligned} \tag{32}$$

Setting $z = \frac{-\nabla f(\hat{w})}{\|\nabla f(\hat{w})\|}$, we have $\|z\| = 1$. By the strong convexity of K we obtain that

$$y_k := \frac{1}{2}(w_k + \hat{w}) + \frac{\gamma}{8}\|w_k - \hat{w}\|^2 z \in K.$$

Therefore, from the (ε, δ)-approximate convexity of f and the optimality of $\hat{w}$, we obtain

$$\begin{aligned} \Delta_k = f(w_k) - f(\hat{w}) &\geq \langle \nabla f(\hat{w}), w_k - \hat{w} \rangle - \frac{\epsilon}{2}\|w_k - \hat{w}\|^2 \\ &= 2\left\langle \nabla f(\hat{w}), \frac{w_k + \hat{w}}{2} - y_k \right\rangle + 2\langle \nabla f(\hat{w}), y_k - \hat{w} \rangle - \frac{\epsilon}{2}\|w_k - \hat{w}\|^2 \\ &\geq 2\left\langle \nabla f(\hat{w}), \frac{w_k + \hat{w}}{2} - y_k \right\rangle - \frac{\epsilon}{2}\|w_k - \hat{w}\|^2 \\ &= 2\left\langle \nabla f(\hat{w}), \frac{\gamma}{8}\|w_k - \hat{w}\|^2 \frac{\nabla f(\hat{w})}{\|\nabla f(\hat{w})\|} \right\rangle - \frac{\epsilon}{2}\|w_k - \hat{w}\|^2 \\ &= \frac{\gamma}{4}\|\nabla f(\hat{w})\|\|w_k - \hat{w}\|^2 - \frac{\epsilon}{2}\|w_k - \hat{w}\|^2 \\ &\geq \left(\frac{\gamma\rho}{4} - \frac{\epsilon}{2}\right)\|w_k - \hat{w}\|^2 = \left(\nu + \frac{\epsilon}{2}\right)\|w_k - \hat{w}\|^2. \end{aligned} \tag{33}$$

Combining (33) with (32) we have

$$\langle \nabla f(w_k), x_k - w_k \rangle \leq -\Delta_k + \frac{\epsilon/2}{\nu + \epsilon/2}\Delta_k = -\frac{\nu}{\nu + \epsilon/2}\Delta_k. \tag{34}$$

Setting $z_k = \frac{-\nabla f(w_k)}{\|\nabla f(w_k)\|}$, we have $\|z_k\| = 1$. By the strong convexity of K we have that

$$y_k := \frac{1}{2}(w_k + x_k) + \frac{\gamma}{8}\|w_k - x_k\|^2 z_k \in K.$$

The optimality of x_k in (25) yields that

$$\begin{aligned}
\langle \nabla f(w_k), x_k - w_k \rangle &\leq \langle \nabla f(w_k), y_k - w_k \rangle \\
&= \left\langle \nabla f(w_k), \frac{1}{2}(x_k - w_k) + \frac{\gamma}{8}\|w_k - x_k\|^2 z_k \right\rangle \\
&= \frac{1}{2}\langle \nabla f(w_k), x_k - w_k \rangle + \frac{\gamma}{8}\|w_k - x_k\|^2 \left\langle \nabla f(w_k), \frac{-\nabla f(w_k)}{\|\nabla f(w_k)\|} \right\rangle \\
&= \frac{1}{2}\langle \nabla f(w_k), x_k - w_k \rangle - \frac{\gamma}{8}\|w_k - x_k\|^2 \|\nabla f(w_k)\| \\
&\leq -\frac{1}{2}\frac{\nu}{\nu + \epsilon/2}\Delta_k - \frac{\gamma}{8}\|w_k - x_k\|^2 \|\nabla f(w_k)\|, \qquad (35)
\end{aligned}$$

where the last inequality follows from (34). Combining (29) with (35), we obtain that

$$\Delta_{k+1} \leq \left(1 - \frac{\nu\eta_k}{2\nu + \epsilon}\right)\Delta_k - \frac{\eta_k}{2}\left(\frac{\gamma\|\nabla f(w_k)\|}{4} - L\eta_k\right)\|x_k - w_k\|^2.$$

Q.E.D.

We are now in a position to establish the convergence results for the CGM.

Theorem 2.3 *Let all the assumptions in Proposition 2.3 be satisfied. Assume also that* $\|w_0 - \hat{w}\| \geq \delta$ *and the sequence* $\{w_k\}$ *generated by the CGM satisfies* $\|\nabla f(w_k)\| \geq \rho$ *for all* k. *Let the sequence* $\{\eta_k\}$ *be chosen such that*

$$0 < \underline{\eta} \leq \eta_k \leq \min\left\{\frac{2\nu + \epsilon}{\nu}, \frac{\gamma\|\nabla f(w_k)\|}{4L}\right\} \quad \forall k. \qquad (36)$$

Then for every $k \in \mathbb{N}$, *if* $\|w_k - \hat{w}\| \geq \delta$ *then*

$$f(w_{k+1}) - \hat{f} \leq \theta\big(f(w_k) - \hat{f}\big),$$

where $\theta = 1 - \frac{\nu\underline{\eta}}{2\nu+\epsilon} \in (0, 1)$. *Moreover, for every* k, *if* $\|w_i - \hat{w}\| \geq \delta$, $i = 0, ..., k$, *then*

$$\|w_k - \hat{w}\|^2 \leq \frac{\Delta_0}{\nu + \epsilon/2}\theta^k,$$

Clearly, in the case $\delta = 0$, the first and the second claims of the theorem mean that the sequences $\{f(w_k)\}$ and $\{w_k\}$ converge linearly to $\hat{f}$ and $\hat{w}$, respectively. In the case $\delta > 0$ we also have linear convergence at least until the generated sequence enters the δ-neighborhood of $\hat{w}$.

Proof. From (36) we have

$$\frac{\gamma\|\nabla f(w_k)\|}{4} - L\eta_k \geq 0, \quad \text{and} \quad 1 \geq \frac{\nu\eta_k}{2\nu + \epsilon} \geq \frac{\nu\underline{\eta}}{2\nu + \epsilon} \quad \forall k.$$

Therefore, it follows from (28) that, for all k, it holds

$$\Delta_{k+1} \leq \left(1 - \frac{\nu\underline{\eta}}{2\nu + \epsilon}\right)\Delta_k,$$

which implies

$$f(w_{k+1}) - f^* \le \theta\,(f(w_k) - f^*)\,. \tag{37}$$

In addition, if $\|w_i - \hat{w}\| \ge \delta$, $i = 0, ..., k$, then we have

$$\Delta_k \le \theta^k \Delta_0.$$

This and (33) imply

$$\|w_k - \hat{w}\|^2 \le \frac{1}{\nu + \epsilon/2}\Delta_k \le \frac{\Delta_0}{\nu + \epsilon/2}\theta^k.$$

Q.E.D.

3 The affine optimal control problem

In this section we turn back to the control-affine linear-quadratic problem (1)–(3) and prove that the gradient methods considered in the previous section are applicable to the (high order) discretization of the problem recently developed in [19, 23]. We also provide error estimates regarding both the errors due to discretization and those due to truncation of the gradient projection iterations.

The first two subsections reproduce assumptions and results from [23] that are necessary for understanding the implementation of the GPM and the CGM to the discretized version of problem (1)–(3). The next subsections prove the applicability of the abstract results obtained above, present details about the implementation of the gradient methods, and provide results of computational experiments.

3.1 Notations and assumptions

Below $\mathbb{R}^n$ denotes the n dimensional Euclidean space (with its elements considered as vector-columns), $|\cdot|$ and $\langle\cdot,\cdot\rangle$ denote the norm and the scalar product, respectively, the superscript $\top$ denotes transposition of vectors and matrices. When dealing with "long" sequences of vectors $w = (w_0, \dots, w_{N-1})$, where $w_i \in \mathbb{R}^p$, so that $w \in (\mathbb{R}^p)^N$, it is convenient to introduce the norms

$$\|w\|_1 := h\sum_{i=0}^{N-1} |w_i|, \qquad \|w\|_2 := \sqrt{h\sum_{i=0}^{N-1} |w_i|^2}, \tag{38}$$

where $h := T/N$. This ensures, in particular, that $\|w\|_1 \le \sqrt{T}\,\|w\|_2$. We also define $\|w\|_\infty = \max_i |w_i|$. As usual, $L_2([0,T];\mathbb{R}^m)$ denotes the Hilbert space of all measurable square-integrable functions $[0,T] \to \mathbb{R}^m$ with scalar product $\langle u_1, u_2\rangle = \int_0^T \langle u_1(t), u_2(t)\rangle\,\mathrm{d}t$ and the corresponding norm is denoted again by $\|\cdot\|_2$.

Let H be the Hilbert space $\left(\mathbb{R}^{2m}\right)^N$ with the scalar product $\langle w', w''\rangle = h\sum_{1=0}^{N-1}\langle w_i', w_i''\rangle$, where each component, w_i, of any $w \in H$ is a pair (u_i, v_i) with $u_i, v_i \in \mathbb{R}^m$.

Assumption (A1) The matrix functions $A(t), B(t), W(t)$ and $S(t)$, $t \in [0,T]$, have Lipschitz continuous first derivatives, Q and $W(t)$ are symmetric. Moreover, the matrix $B^\top(t)S(t)$ is symmetric for all $t \in [0,T]$.

Denote by $\mathcal{F}$ the set of all admissible control-trajectory pairs (u,x), that is, all pairs of an admissible control u and the corresponding (absolutely continuous) solution x of (2). By a standard argument, problem (1)–(3) has a solution, $(\hat{x},\hat{u}) \in \mathcal{F}$, which from now on will be considered as fixed.

Assumption (A2)

$$\frac{1}{2}z(T)^\top Q z(T) + q^\top x(T) + \int_0^T \left(\frac{1}{2}z(t)^\top W(t) z(t) + z(t)^\top S(t) v(t)\right) \mathrm{d}t \geq 0 \qquad \forall\, (z,v) \in \mathcal{F} - (\hat{x},\hat{u}).$$

The first part of Assumption (A1) is standard, while the last requirement is demanding but known from the literature, usually expressed in terms of the Lie brackets of the involved controlled vector fields see e.g. [25]. It is certainly fulfilled in the case of single-input systems, $m = 1$. Assumption (A2) is a directional convexity assumption at $(\hat{x},\hat{u})$, which is somewhat weaker than the usual convexity assumption for the functional J in (1) regarded as a functional on the set of admissible controls (viewing x as a function of u).

The Pontryagin principle implies that there exists an absolutely continuous function $\hat{p} : [0,T] \to \mathbb{R}^n$ such that $(\hat{x},\hat{u},\hat{p})$ satisfies the following system of generalized equations: for a.e. $t \in [0,T]$,

$$0 = \dot{x}(t) - A(t)x(t) + B(t)u(t), x(0) = x_0, \tag{39}$$

$$0 = \dot{p}(t) + A(t)^\top p(t) + W(t)x(t) + S(t)u(t), \tag{40}$$

$$0 \in B(t)^\top p(t) + S(t)^\top x(t) + N_U(u(t)), \tag{41}$$

$$0 = p(T) - Qx(t) - q, \tag{42}$$

where $N_U(u)$ is the normal cone to U at u:

$$N_U(u) := \begin{cases} \emptyset & \text{if } u \notin U, \\ \{l \in \mathbb{R}^m : \langle l, v-u\rangle \leq 0 \ \forall v \in U\} & \text{if } u \in U. \end{cases}$$

Following [23], we assume that the optimal control $\hat{u}$ is *strictly bang-bang*, with a finite number of *switching times* on $[0,T]$, and that the so-called *switching function*,

$$\hat{\sigma}(t) := B(t)^\top \hat{p}(t) + S(t)^\top \hat{x}(t),$$

exhibits a certain growth in a neighborhood of any zero[1].

Assumption (A3) (strict bang-bang property)
There exist real numbers $\kappa \geq 1$ and $\alpha, \tau > 0$ such that for all $j \in \{1,\dots,m\}$ and $s \in [0,T]$ with $\hat{\sigma}_j(s) = 0$ (the j-th component of $\hat{\sigma}$) we have

$$|\hat{\sigma}_j(t)| \geq \alpha |t-s|^\kappa \quad \forall t \in [s-\tau, s+\tau] \cap [0,T].$$

Assumptions (A1)–(A3) will be standing in this section.

[1] A similar assumption is introduced in [9] in the case $\kappa = 1$ and in [22, 24] for $\kappa \geq 1$.

3.2 High-order time-discretization

In this subsection we recall the discretization scheme for problem (1)-(3) presented in [23], which has a higher accuracy than the Euler scheme without a substantial increase of the numerical complexity of the discretized problem. The approach uses second order truncated Volterra-Fliess series. The discretization scheme is described as follows.

For any natural number N denote $h = T/N$ and define the mesh $\{t_i\}_0^N$ with $t_i = ih$. Introducing the notations (where a dot above the symbol of a function denotes the time-derivative)

$$\begin{aligned} A_i &:= A(t_i) + \frac{h}{2}\left(A(t_i)^2 + \dot{A}(t_i)\right), \\ B_i &:= B(t_i) + hA(t_i)B(t_i), \\ C_i &:= -A(t_i)B(t_i) + \dot{B}(t_i), \end{aligned}$$

we replace the differential equation (2) with the discrete-time controlled dynamics

$$x_{i+1} = x_i + h(A_i x_i + B_i u_i + hC_i v_i), \quad i = 0, ..., N-1, \quad x_0 \text{ given}, \tag{43}$$

$$w_i := (u_i, v_i) \in Z^m, \quad i = 0, ..., N-1, \tag{44}$$

where Z^m is the Cartesian product $\Pi_1^m Z$ and Z is the Aumann integral

$$Z := \int_0^1 \begin{pmatrix} 1 \\ s \end{pmatrix} [-1, 1]\, ds.$$

As pointed out in [19], the set Z can be easily represented in the more convenient way as

$$Z = \{(\alpha, \beta) : \alpha \in [-1, 1], \beta \in [\varphi_1(\alpha), \varphi_2(\alpha)]\}, \tag{45}$$

where $\varphi_1(\alpha) := \frac{1}{4}\left(-1 + 2\alpha + \alpha^2\right)$ and $\varphi_2(\alpha) := \frac{1}{4}\left(1 + 2\alpha - \alpha^2\right)$.

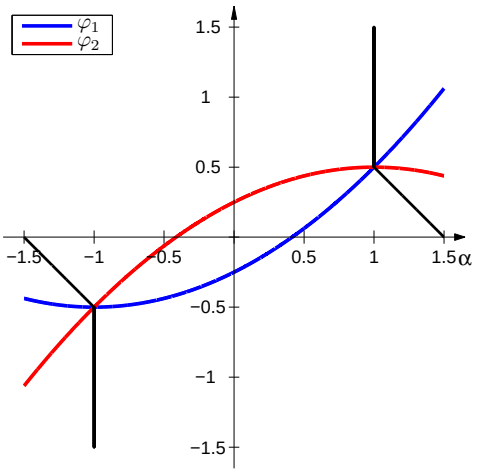

Figure 1: The set Z as the area between the two parabolas φ_1 (lower) and φ_2 (upper).

We introduce the discrete-time counterpart of the objective functional J in (1): for $x = (x_0, ..., x_N)$, $w = (w_0, \dots, w_{N-1}) = ((u_0, v_0), \dots, (u_{N-1}, v_{N-1}))$,

$$J^h(x,w) := \frac{1}{2} x_N^\top (Qx_N + q) + \frac{h}{2} \sum_{i=0}^{N-1} \left(x_i^\top W(t_i) (x_i + hA(t_i)x_i) + \frac{h}{2} x_i^\top \dot{W}(t_i) x_i \right) \tag{46}$$

$$+h \sum_{i=0}^{N-1} \Big(hB(t_i)(u_i - v_i) + x_i^\top \left(S(t_i)u_i + h\dot{S}(t_i)v_i \right) + h \left(A(t_i)x_i \right)^\top S(t_i)v_i + \frac{h}{2} \langle B(t_i)^\top S(t_i)u_i, u_i \rangle \Big).$$

Then we consider the problem of minimization of the functional J^h defined in (46) subject to the constraints (43)–(44). The set of admissible discrete controls in this problem is denoted by $K \subset H$, that is,

$$K := \{(w_0, \dots, w_{N-1}) \in \mathbb{R}^{2m \times N} : \, w_i = (u_i, v_i) \in Z^m\}.$$

We also introduce the discrete adjoint equation (see formula (3.11) in [23])

$$\begin{aligned} p_i \;\; = \;\; & \left(I + hA_i^\top \right) p_{i+1} + h \left(S(t_i)u_i + h\dot{S}(t_i)v_i + hA(t_i)^\top S(t_i)v_i \right) \\ & + h \left(W(t_i) + \frac{h}{2} W(t_i)A(t_i) + \frac{h}{2} A(t_i)^\top W(t_i) + \frac{h}{2} \dot{W}(t_i) \right) x_i + h^2 W(t_i)B(t_i)(u_i - v_i) \end{aligned} \tag{47}$$

with the end condition

$$p_N = Q^\top x_N + q. \tag{48}$$

Section 3.3 in [23] presents a construction which for every sequence $w = (w_0, \dots w_{N-1}) \in K$ defines an admissible control $u = \Phi^h(w)$ in problem (1)–(3), with values ± 1 and with at most two switches in every interval $[t_i, t_{i+1}]$ of each of its components. We do not reproduce this construction here, only mentioning that it requires only a few calculations (to define the switching points), and the restriction of $u(t) = \Phi^h(w)(t)$ to $[t_i, t_{i+1}]$ depends only on w_i. Moreover, the following equalities hold (see (3.14) in [23]): for every $w = ((u_0, v_0), \dots, (u_{N-1}, v_{N-1}))$

$$\int_{t_i}^{t_{i+1}} \Phi^h(w)(s) \, \mathrm{d}s = hu_i, \quad \int_{t_i}^{t_{i+1}} (s - t_i) \Phi^h(w)(s) \, \mathrm{d}s = h^2 v_i, \quad i = 0, \dots, N-1. \tag{49}$$

In addition, the function Φ^h has the important property that there exists a constant $\bar{c}$ independent of N such that

$$\|\Phi^h(w') - \Phi^h(w'')\|_1 \le \bar{c} \|w' - w''\|_1 \qquad \forall w', w'' \in K, \tag{50}$$

Below we will use the metric

$$d^\#(u_1, u_2) = \mathrm{meas} \, \{t \in [0,1] : u_1(t) \neq u_2(t)\}$$

in the set of admissible controls in problem (1)–(3).

The following theorem is extracted from Theorem 3.1 in [23].

Theorem 3.1 *Let Assumption (A1) be fulfilled. Let $(\hat{x}, \hat{u})$ be a solution of problem (1)–(3) for which assumptions (A2) and (A3) are fulfilled with some $\kappa \ge 1$, and let $\hat{p}$ the corresponding solution of the adjoint equation (40) with end-condition (42). Then for every natural number N the*

problem of minimization of (46) under constrains (43)–(44) has a solution $\{(x_i, w_i)\}$ *and for every such solution and the corresponding discrete adjoint sequence* $(p_0, \dots, p_N)$ *solving (47), (48), the following error estimate holds:*

$$\max_{i=0,\dots,N} \left(|x_i - \hat{x}(t_i)| + |p_i - \hat{p}(t_i)|\right) \,+\, d^{\#}\left(\Phi^h(w), \hat{u}\right) \,\leq\, c\, h^{2/\kappa}, \tag{51}$$

where c *is independent of* N*.*

We mention that the above discretization scheme is meaningful only under Assumption (A1). Assumptions (A2) and (A3) are only needed for the error estimate in Theorem 3.1.

3.3 Applicability of the results about gradient-type methods

First of all, we reformulate the problem of minimization of (46) under the constraints (43)–(44) as a minimization problem on the set

$$K := \prod_0^{N-1} Z^m \subset H, \tag{52}$$

namely,

$$\underset{w\in K}{\text{minimize}} \left\{ f^h(w) := J^h(x^h[w], w) \right\}, \tag{53}$$

where $x^h[w]$ is the solution of the discrete-time equation (43) for $w = \{(u_i, v_i)\}_{i=0}^{N-1} \in K$, with the given initial condition x_0.

In this subsection we prove that the assumptions needed for applicability of the results in Section 2 to the above problem are fulfilled.

Lemma 3.1 *The set* K *defined in (52) is strongly convex with modulus* $\gamma \geq \frac{\sqrt{h}}{\sqrt{32}}$*.*

Proof. First of all, the set $Z \subset \mathbb{R}^2$ is strongly convex. This is evident from Figure 1, but the calculation of a modulus γ_0 is cumbersome and we skip the details. In this calculation we use Theorem 1 in [27] (expressing γ_0 by the Lipschitz constant of the mapping that maps a unit vector to that point on the boundary of Z at which this vector is normal to Z) and the explicit formula for the normal cone to Z given in [19, Section 4]. The number $\gamma_0 = 1/\sqrt{32}$ turns out to be a modulus of strong convexity of Z.

Since the norm of $y = (z_1, \dots, z_m) \in \mathbb{R}^{2m} = (\mathbb{R}^2)^m$ with $z_i \in \mathbb{R}^2$ is defined as $|y| = \sqrt{\sum_{i=1}^m |z_i|^2}$, we easily obtain that Z^m is strongly convex with the same modulus γ_0.

For estimating the modulus of strong convexity of $K = (Z^m)^N$ we should take into account that the norm in H is $\|\cdot\|_2$ as defined in (38) with $p = 2m$. Then if we denote by $\mathbb{B}_{\mathrm{H}}$ and by $\mathbb{B}_{\mathbb{R}^{2m}}$ the unit balls in H and $\mathbb{R}^{2m}$, respectively, we have the straightforward inclusion

$$\mathbb{B}_{\mathrm{H}} \subset \frac{1}{\sqrt{\mathrm{h}}} \prod_0^{\mathrm{N}-1} \mathbb{B}_{\mathbb{R}^{2\mathrm{m}}}.$$

The claim of the lemma directly follows from this inclusion, the definition of strong convexity, the product structure of K, and the (upper) estimation of the modulus of strong convexity of Z^m. Q.E.D.

Let us denote by f the objective functional in problem (1)-(3), regarded as a function of the control, namely, $f(u) := J(x[u], u)$, where $x[u]$ is the solution of (2) corresponding to $u \in L_2([0,T];\mathbb{R}^m)$. It is well known that the functional $f : L_2([0,T];\mathbb{R}^m) \to \mathbb{R}$ is Fréchet differentiable at any u and its derivative has the functional representation

$$\nabla f(u)(t) = B(t)^\top p(t) + S(t)^\top x(t), \tag{54}$$

where x and p are the solutions of (39), (40), (42) corresponding to u. Similarly, the function $f^h : H \to \mathbb{R}$ is Fréchet differentiable, and its derivative has the representation (see (3.12) in [23])

$$\begin{aligned} \nabla_{w_i} f^h(w) &= \begin{pmatrix} \nabla_{u_i} f^h(w) \\ \nabla_{v_i} f^h(w) \end{pmatrix} \\ &= \begin{pmatrix} B_i^\top p_{i+1} + S(t_i)^\top x_i + hB(t_i)^\top W(t_i) x_i + hB(t_i)^\top S(t_i) u_i \\ h\left(C_i^\top p_{i+1} - B(t_i)^\top W(t_i) x_i + \left(S(t_i)^\top A(t_i) + S'(t_i)^\top\right) x_i\right) \end{pmatrix}. \end{aligned} \tag{55}$$

We mention that Assumption (A2) implies that f is convex at $\hat{u}$, hence

$$\langle \nabla f(u) - \nabla f(\hat{u}), u - \hat{u} \rangle \geq 0 \qquad \text{for all admissible controls } u. \tag{56}$$

In contrast, f^h does not need to be convex (cf. Lemma 3.6 below).

In the proofs of the next lemmas c_1, c_2, ... denote non-negative constants that may depend on the data of the problem (1)–(3) (and their derivatives) but are independent of N. These constants may have different values in different proofs.

The following two lemmas are technical and needed only to prove the last two lemmas in this section.

Lemma 3.2 *There exist constants c' and c'' independent of h, such that for every $w', w'' \in K$ and $\Delta w \in K - K$*

$$|\langle \nabla f^h(w') - \nabla f^h(w''), \Delta w \rangle| \leq c' \|w' - w''\|_1 \, \|\Delta w\|_1 + c'' h^2 \sum_{i=1}^{N-1} |u_i' - u_i''| \, |\Delta u_i|,$$

where u_i', u_i'', Δu_i are the first coordinates of the components w_i', w_i'', Δw_i of the elements w', w'' and Δw, respectively.

Proof. Considering the discrete equation (43), it is a standard procedure to obtain the following estimate for the solutions x' and x'' corresponding to w' and w'':

$$\|x' - x''\|_\infty \leq c_1 \|w' - w''\|_1. \tag{57}$$

Similarly, also using the last estimation, we obtain from (47), (48) that

$$\|p' - p''\|_\infty \le c_2 \|w' - w''\|_1. \tag{58}$$

Then using the explicit representation (55) we obtain that

$$|\langle \nabla f^h(w') - \nabla f^h(w''), \Delta w\rangle| \le c_1 \left(\|x' - x''\|_\infty + \|p' - p''\|_\infty\right) \|\Delta w\|_1 + c_2 h^2 \sum_{i=1}^{N-1} |u_i' - u_i''|\,|\Delta u_i|,$$

which together with (57) and (58) implies the claim of the lemma. Q.E.D.

Lemma 3.3 *There exists a number $\tilde{c}$ such that for every natural number N, for every $\bar{w} \in K$ and for every $\Delta \in L_2([0,T];\mathbb{R}^m)$*

$$\left|\langle \nabla f(\Phi^h(\bar{w})), \Delta\rangle - \langle \nabla f^h(\bar{w}), w(\Delta)\rangle\right| \le \tilde{c} h^2 \|\Delta\|_1,$$

where $w(\Delta) := \{(u_i, v_i)\}_0^{N-1}$ is defined as

$$u_i = \frac{1}{h}\int_{t_i}^{t_{i+1}} \Delta(t)\,\mathrm{d}t, \qquad v_i = \frac{1}{h^2}\int_{t_i}^{t_{i+1}} (t - t_i)\Delta(t)\,\mathrm{d}t.$$

Proof. Denote by $\bar{x}$ and $\bar{p}$ the solutions of (39) and (40), (42), corresponding to the control function $\bar{u} := \Phi^h(\bar{w})$. Similarly we denote by $\{\bar{x}_i\}$ and $\{\bar{p}_i\}$ the solutions of (43) and (47), (48), corresponding to $\bar{w}$. The results in points 2 and 3 (see (4.5)) in [23, Section 4] imply that for $t \in [t_i, t_{i+1}]$

$$\begin{aligned} B^\top(t)\bar{p}(t) + S(t)^\top \bar{x}(t) \;=\; & (B_i + (t - t_i)C_i)^\top \bar{p}_{i+1} + B(t_i)^\top \big((t_{i+1} - t)W(t_i)\bar{x}_i + S(t_i)\int_{t_i}^{t_{i+1}} \bar{u}(s)\,\mathrm{d}s\big) \\ & + S(t_i)^\top \left(I + (t - t_i)A(t_i)\right)\bar{x}_i + \dot{S}(t_i)^\top (t - t_i)\bar{x}_i + O(t; h^2), \end{aligned}$$

where $O(t;h^2)$ is measurable in t and $|O(t;h^2)| \le c_1 h^2$ for a.e. t. Using this expression and (54) we obtain the following equality:

$$\begin{aligned} \langle \nabla f(\Phi^h(\bar{w})), \Delta\rangle &= \int_0^T \langle B^\top(t)\bar{p}(t) + S(t)^\top \bar{x}(t), \Delta(t)\rangle\,\mathrm{d}t \\ &= \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \Big\langle (B_i + (t - t_i)C_i)^\top \bar{p}_{i+1} + B(t_i)^\top \big((t_{i+1} - t)W(t_i)\bar{x}_i + S(t_i)\int_{t_i}^{t_{i+1}} \bar{u}(s)\,\mathrm{d}s\big) \\ &\quad + S(t_i)^\top \left(I + (t - t_i)A(t_i)\right)\bar{x}_i + \dot{S}(t_i)^\top (t - t_i)\bar{x}_i + O(t; h^2), \Delta(t)\Big\rangle\,\mathrm{d}t. \end{aligned}$$

Using the expressions (55) we obtain, after a simple rearrangement of terms, that

$$\begin{aligned}
&\langle \nabla f(\Phi^h(\bar{w})), \Delta \rangle \\
&= \sum_{i=0}^{N-1} \left[\left\langle \nabla_{u_i} f^h(\bar{w}), \int_{t_i}^{t_{i+1}} \Delta(t)\,\mathrm{d}t \right\rangle + \left\langle \frac{1}{h} \nabla_{v_i} f^h(\bar{w}), \int_{t_i}^{t_{i+1}} (t-t_i)\Delta(t)\,\mathrm{d}t \right\rangle \right] + \int_0^T \langle O(t;h^2), \Delta(t) \rangle\,\mathrm{d}t \\
&= \sum_{i=0}^{N-1} \left[\left\langle \nabla_{u_i} f^h(\bar{w}), h u_i \right\rangle + \left\langle \nabla_{v_i} f^h(\bar{w}), h v_i \right\rangle \right] + \int_0^T \langle O(t;h^2), \Delta(t) \rangle\,\mathrm{d}t \\
&= h \sum_{i=0}^{N-1} \left\langle \nabla_{w_i} f^h(\bar{w}), w_i(\Delta) \right\rangle + \int_0^T \langle O(t;h^2), \Delta(t) \rangle\,\mathrm{d}t = \langle \nabla f^h(\bar{w}), w(\Delta) \rangle + \int_0^T \langle O(t;h^2), \Delta(t) \rangle\,\mathrm{d}t.
\end{aligned}$$

Then the estimation $|O(t;h^2)| \le c_1 h^2$ completes the proof. Q.E.D.

Lemma 3.4 *The function f^h defined in (53) is L-smooth on K with the Lipschitz constant of its derivative being independent of N:*

$$\|\nabla f^h(w') - \nabla f^h(w'')\|_2 \le L \|w' - w''\|_2.$$

Proof. The Fréchet differentiability of f^h was established in [23]), together with the representation (55) of its derivative. The Lipschitz continuity on K follows from this representation, together with (57) and (58) (the notations are as in the proof of Lemma 3.2). Q.E.D.

Lemma 3.5 *There is a constant $\rho > 0$ (independent of N) such that for every sufficiently large N and for every solution $\hat{w}^h$ of the discrete-time problem (53) it holds that $\|\nabla f^h(\hat{w}^h)\|_2 \ge \rho$.*

Proof. First of all, Assumption (A3) directly implies that there exists a number $\rho_0 > 0$ such that

$$\|\nabla f(\hat{u})\|_2 = \|\hat{\sigma}\|_2 \ge \rho_0.$$

We remind that the notations used in the equality above are introduced in Subsection 3.1 (see also (54)).

We utilize Lemma 3.3 with $\Delta = \nabla f(\hat{u})/\|\nabla f(\hat{u})\|_2$ and $\bar{w} = \hat{w}^h$, where $\hat{w}^h$ is an arbitrary solution of problem (53). The vector $w(\Delta)$ is defined as in Lemma 3.3. We have

$$\begin{aligned}
\langle \nabla f^h(\hat{w}^h), w(\Delta) \rangle &\ge \langle \nabla f(\Phi^h(\hat{w}^h)), \Delta \rangle - \tilde{c} h^2 \|\Delta\|_1 \\
&\ge \langle \nabla f(\hat{u}), \Delta \rangle - |\langle \nabla f(\Phi^h(\hat{w}^h)) - \nabla f(\hat{u}), \Delta \rangle| - \tilde{c}\sqrt{T} h^2 \|\Delta\|_2 \\
&\ge \rho_0 - L \|\Phi^h(\hat{w}^h) - \hat{u}\|_2 - \tilde{c}\sqrt{T} h^2 \ge \rho_0 - c_1 L h^{1/\kappa} - \tilde{c}\sqrt{T} h^2.
\end{aligned}$$

In the last row of the above inequalities we use the Lipschitz constant[2], L, of ∇f and the estimation in Theorem 3.1. We also use that $\|\Phi^h(\hat{w}^h) - \hat{u}\|_2 \le c_1\sqrt{d^{\#}(\Phi^h(\hat{w}^h), \hat{u})}$ and Theorem 3.1. This implies the claim of the lemma with $\rho = \rho_0/2$. Q.E.D.

Lemma 3.6 *There exists a constant $c_0 \ge 0$ such that for every natural number N, for any solution $\hat{w}^h$ of the discrete problem (53), and for any number $\varepsilon > 0$, the objective function f^h is (ε, δ)-approximately convex at $\hat{w}^h$ with any*

$$\delta \ge c_0 \max\left\{\frac{1}{\varepsilon}, \frac{1}{\sqrt{\varepsilon}}\right\} h^{2/\kappa}.$$

Proof. Define the vector $\tilde{w} = \{(\tilde{u}_i, \tilde{v}_i)\}$ with

$$\tilde{u}_i = \frac{1}{h}\int_{t_i}^{t_{i+1}} \hat{u}(t)\,\mathrm{d}t, \quad \tilde{v}_i = \frac{1}{h^2}\int_{t_i}^{t_{i+1}} (t - t_{i+1})\hat{u}(t)\,\mathrm{d}t, \quad i = 0, \dots, N-1,$$

where, as before, $\hat{u}$ is the optimal control in the problem (1)–(3). According to property (49), the functions Φ^h and $\hat{u}$ have the same zero-th and first integral moments on each interval $[t_i, t_{i+1}]$. Then formula (54) and the analysis in Section 3.1 in [23] imply the inequality

$$\|\nabla f(\Phi^h(\tilde{w})) - \nabla f(\hat{u})\|_\infty \le c_1 h^2.$$

Let us fix an arbitrary solution $\hat{w}^h$ of problem (53), an arbitrary $w \in K$. From the convexity of the functional f at $\hat{u}$ we have

$$\langle \nabla f(\Phi^h(w)) - \nabla f(\hat{u}), \Phi^h(w) - \hat{u}\rangle \ge 0,$$

hence

$$\langle \nabla f(\Phi^h(w)) - \nabla f(\Phi^h(\tilde{w})), \Phi^h(w) - \hat{u}\rangle \ge -c_1 h^2 \|\Phi^h(w) - \hat{u}\|_1.$$

Now we utilize Lemma 3.3 with $\Delta = \Phi^h(w) - \hat{u}$, first taking $\bar{w} = w$, then taking $\bar{w} = \tilde{w}$. This yields

$$\left|\langle \nabla f(\Phi^h(w)), \Delta\rangle - \langle \nabla f^h(w), w(\Delta)\rangle\right| \le \tilde{c} h^2 \|\Delta\|_1,$$

and

$$\left|\langle \nabla f(\Phi^h(\tilde{w})), \Delta\rangle - \langle \nabla f^h(\tilde{w}), w(\Delta)\rangle\right| \le \tilde{c} h^2 \|\Delta\|_1,$$

correspondingly. Then

$$\begin{aligned}\langle \nabla f^h(w), w(\Delta)\rangle - \langle \nabla f^h(\tilde{w}), w(\Delta)\rangle &\ge \langle \nabla f(\Phi^h(w)), \Delta\rangle - \langle \nabla f(\Phi^h(\tilde{w})), \Delta\rangle - 2\tilde{c} h^2 \|\Delta\|_1 \\ &\ge -c_1 h^2 \|\Phi^h(w) - \hat{u}\|_1 - 2\tilde{c} h^2 \|\Delta\|_1 = -c_2 h^2 \|\Phi^h(w) - \hat{u}\|_1.\end{aligned}$$

[2] The Lipschitz continuity of ∇f (with some constant L) follows from the representation (54), similarly as for ∇f^h in Lemma 3.4.

Observe that according to its definition in Lemma 3.3, $w(\Delta) = w - \tilde{w}$. Thus

$$\langle \nabla f^h(w) - \nabla f^h(\tilde{w}), w - \tilde{w} \rangle \geq -c_2 \|\Phi^h(w) - \hat{u}\|_1.$$

In the last inequality we replace $\tilde{w}$ with $\hat{w}^h$ obtaining that

$$\begin{aligned}
&\langle \nabla f^h(w) - \nabla f^h(\hat{w}^h), w - \hat{w}^h \rangle \\
&\geq -|\langle \nabla f^h(\hat{w}^h) - \nabla f^h(\tilde{w}), w - \hat{w}^h \rangle| - |\langle \nabla f^h(w) - \nabla f^h(\tilde{w}), \hat{w}^h - \tilde{w} \rangle| - c_2 h^2 \|\Phi^h(w) - \hat{u}\|_1.
\end{aligned}$$

Now we use Lemma 3.2 to estimate the two scalar products in the right-hand side as follows:

$$|\langle \nabla f^h(\hat{w}^h) - \nabla f^h(\tilde{w}), w - \hat{w}^h \rangle| \leq c' \|\hat{w}^h - \tilde{w}\|_1 \, \|w - \hat{w}^h\|_1 + c'' h^2 \sum_{i=1}^{N-1} |\hat{u}_i^h - \tilde{u}_i| \, |u_i - \hat{u}_i^h|,$$

$$|\langle \nabla f^h(w) - \nabla f^h(\tilde{w}), \hat{w}^h - \tilde{w} \rangle| \leq c' \|w - \tilde{w}\|_1 \, \|\hat{w}^h - \tilde{w}\|_1 + c'' h^2 \sum_{i=1}^{N-1} |u_i - \tilde{u}_i| \, |\hat{u}_i^h - \tilde{u}_i|.$$

Notice, that due to (49) and Theorem 3.1

$$|\hat{u}_i^h - \tilde{u}_i| = \frac{1}{h} \left| \int_{t_i}^{t_{i+1}} [\Phi^h(\hat{w}^h)(t) - \hat{u}(t)] \, \mathrm{d}t \right| \leq \frac{1}{h} c h^{2/\kappa},$$

hence

$$\begin{aligned}
&|\langle \nabla f^h(\hat{w}^h) - \nabla f^h(\tilde{w}), w - \hat{w}^h \rangle| + |\langle \nabla f^h(w) - \nabla f^h(\tilde{w}), \hat{w}^h - \tilde{w} \rangle| \\
&\quad \leq c_3 \left(\|\hat{w}^h - \tilde{w}\|_1 \, \|w - \hat{w}^h\|_1 + h^{2/\kappa} \|w - \hat{w}^h\|_1 + \|w - \tilde{w}\|_1 \, \|\hat{w}^h - \tilde{w}\|_1 + h^{2/\kappa} \|w - \tilde{w}\|_1 \right) \\
&\quad \leq c_3 \left(\|\hat{w}^h - w\|_1 + \|w - \tilde{w}\|_1 \right) \left(\|\hat{w}^h - \tilde{w}\|_1 + h^{2/\kappa} \right).
\end{aligned}$$

Thus we obtain that

$$\begin{aligned}
&\langle \nabla f^h(w) - \nabla f^h(\hat{w}^h), w - \hat{w}^h \rangle \\
&\qquad \geq -c_3 \left(\|\hat{w}^h - w\|_1 + \|w - \tilde{w}\|_1 \right) \left(\|\hat{w}^h - \tilde{w}\|_1 + h^{2/\kappa} \right) - c_2 h^2 \|\Phi^h(w) - \hat{u}\|_1.
\end{aligned} \tag{59}$$

According to (50) and Theorem 3.1 we have

$$\|\Phi^h(w) - \hat{u}\|_1 \leq \|\Phi^h(w) - \Phi^h(\hat{w}^h)\|_1 + \|\Phi^h(\hat{w}^h) - \hat{u}\|_1 \leq \bar{c} \|w - \hat{w}^h\|_1 + c h^{2/\kappa}.$$

Moreover,

$$\begin{aligned}
\|\hat{w}^h - \tilde{w}\|_1 &= h \sum_{i=0}^{N-1} |\hat{u}_i^h - \tilde{u}_i| + h \sum_{i=0}^{N-1} |\hat{v}_i^h - \tilde{v}_i| \\
&\leq h \left[\sum_{i=0}^{N-1} \frac{1}{h} \int_{t_i}^{t_{i+1}} |\Phi^h(\hat{w}^h)(t) - \hat{u}(t)| \, \mathrm{d}t + \sum_{i=0}^{N-1} \frac{1}{h^2} \int_{t_i}^{t_{i+1}} (t - t_i) |\Phi^h(\hat{w}^h)(t) - \hat{u}(t)| \, \mathrm{d}t \right] \\
&\leq c_4 \sum_{i=0}^{N-1} \mathrm{meas}\{t \in [t_i, t_{i+1}] : \, \Phi^h(\hat{w}^h)(t) \neq \hat{u}(t)\} \leq c_4 d^{\#}(\Phi^h(\hat{w}^h), \hat{u}) \leq c_5 h^{2/\kappa}.
\end{aligned}$$

In addition,

$$\|w - \tilde{w}\|_1 \le \|w - \hat{w}^h\|_1 + \|\hat{w}^h - \tilde{w}\|_1 \le \|\hat{w}^h - w\|_1 + c_5 h^{2/\kappa}.$$

Combining the last three (chains of) inequalities with (59) we obtain that

$$\langle \nabla f^h(w) - \nabla f^h(\hat{w}^h), w - \hat{w}^h \rangle \ge -c_6 h^{2/\kappa} \left(\|\hat{w}^h - w\|_1 + h^{2/\kappa} \right).$$

Now we fix an arbitrary $\varepsilon > 0$ and take δ so large that

$$2c_6\sqrt{T}\frac{h^{2/\kappa}}{\varepsilon} \le \delta \qquad \text{and} \qquad 2c_6\frac{h^{2/\kappa}}{\sqrt{\varepsilon}} \le \delta.$$

Then the inequality

$$\langle \nabla f^h(w) - \nabla f^h(\hat{w}^h), w - \hat{w}^h \rangle \ge -\varepsilon \|\hat{w}^h - w\|_2^2$$

holds for any $w \in K$ such that $\|\hat{w}^h - w\|_2 \ge \delta$. This completes the proof (see Definition 2.4). Q.E.D.

Let us interpret the convergence result in Theorem 2.1 in view of the above lemmas, focusing on the generic case $\kappa = 1$. From Lemma 3.1 we know that one can take $\gamma = \sqrt{\varepsilon}/\sqrt{32}$ and according to Lemma 3.4 and Lemma 3.5 the numbers L and ρ in Theorem 2.1 can be taken independent of N. Thus $\nu = \sqrt{h}\rho/16\sqrt{2} - \varepsilon$, and in order to ensure that $\nu > 0$ we have to choose

$$\varepsilon < \varepsilon_0 := \frac{\sqrt{h}\rho}{16\sqrt{2}}.$$

Choosing, for example, $\varepsilon = \varepsilon_0/2$ and taking into account that $\sqrt{h} > h$ for all sufficiently small h we take $\delta = 2c_2 h^2/\varepsilon_0 =: c^* h^{3/2}$. Thus the GPM converges linearly at least until the current iteration w_k enters into an $O(h^{3/2})$-neighborhood of a solution $\hat{w}^h$.

From (17) in Theorem 2.1 one can estimate from above the theoretical linear convergence rate as $\mu \le 1 - \beta\sqrt{h}$, which approaches 1 for small h. Fortunately, the second order approximation provided by our discretization scheme allows for using not too small h, and as shown in the next subsection by an example, the values of μ may be quite reasonable.

The analysis of the CGM is similar.

3.4 Implementation of the gradient methods

Now, we shall describe the implementation of the GPM and the CGM to the specific problem defined in (53) and (52). In the previous subsection we have shown that all assumptions required in the abstract results about the GPM in Section 2.1 are fulfilled. The same applies to the CGM, with the exception that the (ε, δ)-approximate convexity of f (in the optimal control context $f := f^h$) is required not only at a solution point $\hat{w}^h$, but on the whole set K. This property can easily be ensured modifying Assumption (A2) by replacing $\mathcal{F} - (\hat{u}, \hat{v})$ with $\mathcal{F} - \mathcal{F}$.

The two key points in the implementation of the gradient methods are: (i) calculation of the gradient $\nabla f^h(w)$; calculation of projections on K (for the GPM) or solving a linear optimization

problem on K (for the CGM). We do not discuss here the issue of the choice of the step sizes λ_k, for which numerous possibilities are known from the literature.

1. Calculation of $\nabla f^h(w)$. Since f^h represents the objective function of a discrete-time optimal control problem as a function of the control variables (the state being implicitly regarded as a function of the control), we employ the well known in control theory way for calculating its gradient: $\nabla f^h(w)$ is the derivative of the Hamiltonian with respect to the control, evaluated at the current control-trajectory pair, together with the corresponding solution of the adjoint equation. The explicit formula is given in (55), reproducing [23, Section 3.2].

2. Calculation of the projection on K.

The set K is a product of $m \times N$ copies of the strongly convex set Z, thus the projection of a vector $w \in H$ onto K is represented by projections onto Z of the 2-dimensional components of w. Thus we have to only calculate projections, $P_Z(u,v)$ on Z, where $(u,v)^\top \in \mathbb{R}^2$.

The following representation of the normal cone to the set Z is obtained in [19, Section 4]:

$$N_Z(\alpha,\beta) = \begin{cases} \emptyset & \text{if } (\alpha,\beta) \notin Z, \\ \left\{\alpha\,(\lambda, \mu-\lambda)^\top : \mu \geq 0, \lambda \geq 0\right\} & \text{if } \alpha \in \{-1,1\}, \\ \left\{\mu\,(\zeta+\alpha, -2\zeta)^\top : \mu \geq 0\right\} & \text{if } \alpha \in (-1,1) \wedge \beta \in \{\varphi_1(\alpha), \varphi_2(\alpha)\}, \\ \{0\} & \text{if } \alpha \in (-1,1) \wedge \beta \in (\varphi_1(\alpha), \varphi_2(\alpha)), \end{cases} \tag{60}$$

where $\zeta = \operatorname{sgn}(\alpha - 2\beta)$.

Now, take arbitrarily a vector $\xi = (u,v)^\top \in \mathbb{R}^2$ and observe that $P_Z(\xi)$ is the unique solution of the inclusion

$$P_Z(\xi) \in \xi - N_Z(P_Z(\xi)). \tag{61}$$

Therefore, using the formula (60), one can explicitly calculate $P_Z(\xi)$ as

$$P_Z(u,v) = \begin{cases} (u,v) & \text{if } (u,v) \in Z, \\ (1, \frac{1}{2}) & \text{if } u \geq 1 \ \text{ and } \ u+v \geq \frac{3}{2}, \\ (-1, -\frac{1}{2}) & \text{if } u \leq -1 \ \text{ and } \ u+v \leq -\frac{3}{2}, \\ (\alpha_1, \varphi_1(\alpha_1)) & \text{if } u > -1 \ \text{ and } \ u+v < \frac{3}{2} \ \text{ and } \ v < \varphi_1(u), \\ (\alpha_2, \varphi_2(\alpha_2)) & \text{if } u < 1 \ \text{ and } \ u+v < -\frac{3}{2} \ \text{ and } \ v > \varphi_2(u), \end{cases} \tag{62}$$

where the functions φ_1 and φ_2 are defined after (45), α_1 is a solution in $[-1,1]$ of the third order equation

$$\alpha^3 + 3\alpha^2 + (9-4v)\alpha - 8u - 4v - 1 = 0, \tag{63}$$

and α_2 is a solution in $[-1,1]$ of the third order equation

$$\alpha^3 - 3\alpha^2 + (9+4v)\alpha - 8u - 4v + 1 = 0. \tag{64}$$

Indeed, the first three cases in the representation (62) are clear. In the fourth case

$$u > -1 \ \text{ and } \ u + v < \frac{3}{2} \ \text{ and } \ v < \varphi_1(u),$$

thus $P_Z(u, v)$ has the form $(\alpha, \varphi_1(\alpha))$ (see Figure 1). From (60), we have

$$N_Z((\alpha, \varphi_1(\alpha))) = \mu(1 + \alpha, -2)^\top.$$

Combining this with (61), one has

$$\begin{pmatrix} u - \alpha \\ v - \varphi_1(\alpha) \end{pmatrix} = \mu \begin{pmatrix} 1 + \alpha \\ -2, \end{pmatrix}$$

implying

$$\frac{u - \alpha}{v - \varphi_1(\alpha)} = \frac{1 + \alpha}{2},$$

which leads to (63). The last case is treated similarly.

3. Solving the auxiliary sub-problem in the CGM.

Now, we consider the subproblem $\min_{y \in K} \langle \nabla f^h(w), y \rangle$ which appears in the implementation of the CGM (see (25)).

Observe that, the necessary (and sufficient) optimality condition for this problem reads as

$$0 \in \nabla f^h(w) + N_K(y).$$

Each component of this inclusion has the form $(\xi_1, \xi_2) \in N_Z((\alpha, \beta))$, which, thanks to (60), can be explicitly represented (see [19]) by the following simple formula:

$$(\alpha, \beta) = \begin{cases} (-1, -1/2) & \text{if } \ \xi_1 \le 0 \ \text{ and } \ \xi_1 + \xi_2 \le 0, \\ (1, 1/2) & \text{if } \ \xi_1 > 0 \ \text{ and } \ \xi_1 + \xi_2 \ge 0, \\ (-1 - 2\xi_1/\xi_2, \varphi_1(\alpha)) & \text{if } \ \xi_1 > 0 \ \text{ and } \ \xi_1 + \xi_2 < 0, \\ (1 + 2\xi_1/\xi_2, \varphi_2(\alpha)) & \text{if } \ \xi_1 \le 0 \ \text{ and } \ \xi_1 + \xi_2 > 0. \end{cases} \tag{65}$$

Therefore, the subproblem (25) can be solved explicitly without solving any third order algebraic equation as in the GPM.

3.5 Numerical examples

In this section, we present some numerical experiments for the example of an affine linear-quadratic optimal control problem given in [23].

Example 3.1

$$\begin{array}{ll} \text{minimize} & -by(1) + \int_0^1 \frac{1}{2} \left(x(t)\right)^2 \, \mathrm{d}t \\ \text{subject to} & \dot{x}(t) = y(t), \quad x_1(0) = a \\ & \dot{y}(t) = u(t), \quad y(0) = 1. \\ & u(t) \in [-1, 1]. \end{array} \tag{66}$$

For appropriate values of a and b, there is a unique optimal solution $\hat{u}$ with a switch from -1 to 1 at time τ, which is a solution of the equation

$$-5\tau^4 + 24\tau^3 - (12a + 36)\tau^2 + (24a + 20)\tau + 24b - 12a - 3 = 0.$$

As in [23], we choose $a = 1, b = 0.1$, then $\tau = 0.492487520$ is a simple zero of the switching function. Here the number $\kappa = 1$ in Assumption (A3) and the exact optimal control is

$$\hat{u}(t) = \begin{cases} -1 & \text{if } t \in [0, \tau] \\ 1 & \text{if } t \in (\tau, 1]. \end{cases}$$

For each N, the iterates $\{w_k\}$ generated by GPM or CGM converge linearly to the unique (in this example) solution $\hat{w}^h$ with rates μ_N and θ_N, respectively. The starting control is chosen as $u_0(t) = 1$, $t \in [0, T]$, for both algorithms. In the following tables, we report these rates for some values of N. The stopping condition is $\|w_{k+1} - w_k\| \leq 10^{-6}$ for the GPM and $\|x_k - w_k\| \leq 10^{-6}$ for the CGM.

Table 1: Convergence rates for the GPM

N	10	20	30	40	50	60	70	80	90	100
μ_N	0.2744	0.4687	0.5742	0.6477	0.6874	0.7166	0.7327	0.8038	0.8736	0.8778

Table 1 indicates that the (numerically obtained) rate of linear convergence, μ_N, of the GPM depends on the mesh size N: it is monotone increasing and likely approaching 1 when N increases. This is to be expected, since according to Theorem 2.1, μ_N is inversely dependent on the index of strong convexity, γ, of the set K, which tends to zero when $N \to \infty$ (see Lemma 3.1). Actually, the convergence of μ_N to 1 is also consistent with the fact, that the GPM applied (theoretically) to the continuous-time problem (1)–(3) converges sub-linearly, as recently established in [21, Theorem 3.2]. We stress that due to the second order accuracy of discretization, the mesh size N does need to be taken large, therefore the rate of linear convergence may be reasonably good (see Table 1 for $N = 10 - 30$).

Table 2 presents the rate of linear convergence of the CGM applied to the same example. Although, as mentioned at the end of Subsection 3.4, the amount of computations at each step of the CGM is slightly lower than that for the GPM, the rate of linear convergence is worse.

Table 2: Convergence rates for the CGM

N	10	20	30	40	50	60	70	80	90	100
θ_N	0.8946	0.8999	0.9016	0.9023	0.9028	0.9030	0.9032	0.9034	0.9035	0.9036

4 Concluding remarks

In this paper we obtain a number of new results about the convergence of gradient methods for general optimization problems on strongly convex feasible sets. The main motivation is the application of a recently developed discretization scheme [19, 23] for linear-quadratic affine optimal control problems, which results in discrete-time problems of the same type, however, with *strongly convex* point-wise control constraints having rather simple representations by means of quadratic inequalities. This opens several directions of further research.

First, to develop more efficient (than gradient projection) methods using the specific linear-quadratic structure of the objective function and of the constraints.

Second, to investigate the applicability of gradient projection methods to discretized *nonlinear* optimal control problems with the control appearing linearly. As indicated in [16], our discretization approach is also applicable to such problems, and results in mathematical programming problems with strongly convex feasible sets. The general convergence results obtained in the present paper are also applicable, in principle. The main open problem here, is that the error analysis of the discretization is not developed for nonlinear problems, which also creates problems to justify the applicability and the convergence of gradient methods (cf. the analysis in Subsection 3.3).

References

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

[2] W. Alt, C. Schneider and M. Seydenschwanz, *Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang-bang solutions*, Appl. Math. Comput., 287 (2016), pp. 104–124.

[3] W. Alt, U. Felgenhauer and M. Seydenschwanz, *Euler discretization for a class of nonlinear optimal control problems with control appearing linearly*, Computational Optimization and Applications (2017), https://doi.org/10.1007/s10589-017-9969-7.

[4] H. Attouch and D. Aze, *Approximation and regularization of arbitrary functions in Hilbert spaces by the Lasry-Lions method*, Ann. Inst. Henri Poincaré, 3 (1993), pp. 289–312.

[5] M. V. Balashov and M. O. Golubev, *About the Lipschitz property of the metric projection in the Hilbert space*, J. Math. Anal. Appl., 394 (2012) 545–551.

[6] A. Beck and M. Teboulle, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[7] V. F., Demyanov and A. M. Rubinov, *Approximate methods in optimization problems*, Elsevier Publishing Company, 1970.

[8] J. C. Dunn, *Rates of Convergence for Conditional Gradient Algorithms Near Singular and Nonsingular Extremals,*, SIAM J. Control Optim. 17 (1979), pp. 187-211.

[9] U. Felgenhauer, *On stability of bang-bang type controls*, SIAM J. Control Optim., 41 (2003), pp.1843–1867.

[10] U. Felgenhauer, *Discretization of semilinear bang-singular-bang control problems*, Computational Optimization and Applications, 64 (2016), pp. 295–326.

[11] U. Felgenhauer, *A Newton-type method and optimality test for problems with bang-singular-bang optimal control.* Pure and Applied Functional Analysis, 1 (2016), pp. 197–215.

[12] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp.149–154.

[13] D. Garber and E. Hazan, *Faster Rates for the Frank-Wolfe Method over Strongly-Convex Sets*, ICML'15, 37 (2015), pp. 541–549.

[14] M. O. Golubev, *Gradient projection method for convex function and strongly convex set*, IFAC-PapersOnLine, 48 (2015), pp. 202–205.

[15] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York (1980).

[16] F. Lempio and V.M. Veliov. *Discrete approximations of differential inclusion.* Bayreuther Mathematische Schriften, 54 (1998), pp. 149–232.

[17] D.G. Luenberger and Y. Ye, *Linear and nonlinear programming*, Third Edition, Springer, 2008.

[18] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Springer Science and Business Media, 2013.

[19] A. Pietrus, T. Scarinci, and V.M. Veliov. *High order discrete approximations to Mayer's problems for linear systems.* To appear in SIAM J. Control Optim., 2017. Available as Research Report 2016-04, ORCOS, TU Wien, 2016, at
https://orcos.tuwien.ac.at/fileadmin/t/orcos/Research_Reports/2016-04.pdf.

[20] J. Preininger, T. Scarinci and V.M. Veliov, *Metric regularity properties in bang-bang type linear-quadratic optimal control problems*, submitted. Available as Research Report, 2017-07, ORCOS, TU Wien, 2017, at
https://orcos.tuwien.ac.at/fileadmin/t/orcos/Research_Reports/2017-07.pdf.

[21] J. Preininger and P. Vuong, *On the convergence of the gradient projection method for optimal control problems with bang-bang solutions*, submitted. Available as Research Report 2017-10, ORCOS, TU Wien, 2017, at
https://orcos.tuwien.ac.at/fileadmin/t/orcos/Research_Reports/2017-10.pdf.

[22] M. Quincampoix and V. Veliov, *Metric Regularity and Stability of Optimal Control Problems for Linear Systems*, SIAM J. Control Optim, 51 (2013), pp. 4118–4137.

[23] T. Scarinci and V.M. Veliov, *Higher-order numerical schemes for linear quadratic problems with bang-bang Controls*, Computational Optimization and Applications, (2017), DOI 10.1007/s10589-017-9948-z.

[24] M. Seydenschwanz, *Convergence results for the discrete regularization of linear-quadratic control problems with bang-bang solutions*, Comput. Optim. Appl., 61 (2015) pp. 731–760.

[25] V.M. Veliov. *On the time-discretization of control systems.*, SIAM J. Control Optim., 35 (1997), pp. 1470–1486.

[26] V.M. Veliov, *Error analysis of discrete Approximation to bang-bang optimal control problems: the linear case*, Control Cyberne., 34 (2005), pp. 967–982.

[27] J-P Vial, *Strong convexity of sets and functions*, Journal of Mathematical Economics, 9 (1982), 187–205.