Applied Mathematics & Optimization
https://doi.org/10.1007/s00245-018-9528-3

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

V. M. Veliov[1] · **P. T. Vuong**[1]

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. A numerical example is given, confirming the theoretical findings.

Keywords Optimal control · Mathematical programming · Numerical methods · Gradient methods · Affine control systems · Bang–bang control

Mathematics Subject Classification 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 typi-

This research is supported by the Austrian Science Foundation (FWF) under Grant No. P31400-N32.

✉ V. M. Veliov
vladimir.veliov@tuwien.ac.at

P. T. Vuong
vuong.phan@tuwien.ac.at

[1] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Published online: 06 October 2018

cal 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 [1,2,25,27] apply to problems with linear dynamics, while [3,11] 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 [21,24] 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 [12,23] 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:

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

subject to

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + d(t), \quad x(0) = x_0, \quad t \in [0,T], \qquad (2)$$

$$u(t) \in U := [-1,1]^m. \qquad (3)$$

Here, $[0,T]$ is a fixed time horizon, $Q \in \mathbb{R}^{n\times n}, q \in \mathbb{R}^n$, $A(t), W(t) \in \mathbb{R}^{n\times n}$, $B(t), S(t) \in \mathbb{R}^{n\times m}$, $d(t) \in \mathbb{R}^n$ 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 [21,24]. 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. [18,19] 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 [6,15], where strong convexity is assumed for the feasible set instead of the objective function. However, as clarified in the end of Sect. 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 Sect. 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 [6,15]. 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 Sect. 2.2.

In Sect. 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 [21,24] and the error estimate proved in [24]. 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 Sects. 3.3 and 3.4 we prove the applicability of the abstract convergence results, obtained in Sect. 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 Sect. 3.5.

The paper concludes with indication of some open problems for further research (Sect. 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. We remind that if $w^* \in K$ is a (local) solution of (4) and f is Fréchet-differentiable at w^* then

$$\left\langle \nabla f(w^*), y - w^* \right\rangle \geq 0 \quad \forall y \in K.$$

Convergence results for gradient projection methods for this problem in finite dimensional spaces and convex f are known (see e.g. [19]). 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 even need 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 [6,15], 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 [16, 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

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

An alternative definition is often used in the literature: a set is strongly convex (with respect to the number $R > 0$) if it coincides with the intersection of all balls of radius

R containing this set. The two definitions are equivalent (see e.g. [28, Theorem 1]) and the relation between γ and R is that $R = 1/\gamma$.[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 "paraconvexity" (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 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

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

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 K is γ-strongly convex, f is differentiable on K and $\hat{w} \in K$ is a solution of problem (4) such that $\|\nabla f(\hat{w})\| \geq \rho$ for some number $\rho > 0$.*

[1] The equivalence is proved in [28, Theorem 1] for finite-dimensional Hilbert spaces only, but the proof uses only two-dimensional geometric considerations that work in any Hilbert space.

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

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

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 completed. □

Property (7) will play an important role in the further analysis. In fact, the (ε, δ)-approximate convexity of f and the strong convexity of K were needed just to ensure existence of $\nu > 0$ and $\delta \geq 0$ for which condition (7) is fulfilled. We mention that (7) is always fulfilled if the set K is convex and the function f is strongly convex, which is not the case here.

Lemma 2.1 *Let f be differentiable on K and let condition (7) be fulfilled with some $\nu > 0$. 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. □

Lemma 2.2 *Let f be differentiable on K and let condition (7) be fulfilled with some $\nu > 0$. If for some $w \in K$ it 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. □

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 $w_k = P_K(w_k - \nabla f(w_k))$ 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 $\lambda_k = \lambda \in (0, \frac{1}{L}]$, see e.g. [7]. More precisely, if problem (4) has a solution and $\hat{f}$ is the minimal value of f on K, then

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

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 linearly to the unique solution of (4). If f is only convex (but not necessarily strongly convex), the sequence $\{w_k\}$ converges weakly [20]. When K is strongly convex, the linear convergence of $\{w_k\}$ is obtained under additional conditions (too strong for our main application) in [5,6,15].

In this subsection, we prove that if condition (7) is fulfilled with $\nu > 0$ then the sequence $\{w_k\}$ generated by the GPM linearly approaches $\hat{w}$ at least until entering a δ-neighborhood of $\hat{w}$. Proposition 2.1 gives conditions for existing of such ν in terms of strong convexity of the set K and (ε, δ)-approximate convexity of the function f. We mention that if the above algorithm of the GPM stops at Step 1 for some k then, according to Lemma 2.1, $\|w_k - \hat{w}\| \le \delta$, that is, a δ-approximate solution is attained (obviously this is meaningful only if δ is sufficiently small).

Proposition 2.2 *Let f be L-smooth on K, let condition (7) be fulfilled with some $\nu > 0$, and let $\|w_0 - \hat{w}\| \ge \delta$. Then the sequence $\{w_k\}$ generated by the GPM satisfies the inequality*

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

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

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

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

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

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

or equivalently

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

Since $w_{k+1} \in K$ and $\lambda_k > 0$, if $\|w_{k+1} - \hat{w}\| \geq \delta$ then due to (7)

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

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,

$$2\langle w_k - w_{k+1}, \hat{w} - w_{k+1} \rangle = \|w_k - w_{k+1}\|^2 + \|w_{k+1} - \hat{w}\|^2 - \|w_k - \hat{w}\|^2. \tag{15}$$

Combining (14) and (15) we obtain that

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

hence (10) is satisfied. □

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*

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

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 $[1 + \lambda_k(2\nu - \lambda_k L^2)] \geq [1 + a(2\nu - bL^2)] > 1$ 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),

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

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

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

□

Remark 2.1 If the constants L and ν 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 L and ν 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 positive number $\delta' \geq \delta$ 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 $4\lambda_k L^2 < \gamma\rho - 4\varepsilon$ for every $k \geq k_0$. Hence,

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

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

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

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 $\|w_{i+1} - \hat{w}\| \geq \delta$ for $i = k_0, \ldots, k$, we obtain that either $\|w_k - \hat{w}\| \longrightarrow 0$ or $\|w_k - \hat{w}\| < \delta$ for some $k \geq k_0$. If $\|w_k - \hat{w}\| \longrightarrow 0$ then the claim is true since $\delta' > 0$. If $\|w_k - \hat{w}\| < \delta$ for some $k \geq k_0$, then $\|w_{k+1} - \hat{w}\| < \delta$. Indeed, if $\|w_{k+1} - \hat{w}\| \geq \delta$ then we have from (10)

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

which is a contradiction. Thus w_k remains in the δ'-neighborhood of $\hat{w}$ for all $k \geq k_0$. The proof is completed. □

Remark 2.3 Using the contractivity of the projection onto strongly convex sets, Balashov and Golubev [6] and Golubev [15] 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

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

where $N_K(w_k)$ is the normal cone to K at w_k defined as

$$N_K(w_k) := \begin{cases} \emptyset & \text{if } \ w_k \notin K, \\ \{l \in H : \langle l, v - w_k\rangle \leq 0 \ \forall v \in K\} & \text{if } \ w_k \in K. \end{cases}$$

(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 [13] 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 [8,9,14].

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 $\eta_k \in (0, \min\{1, \frac{\gamma \|\nabla f(w_k)\|}{4L}\}]$, 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, under the assumptions of 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 not a polytope, 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 on K 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 $\nu := \frac{\gamma\rho}{4} - \varepsilon$ 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*

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

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 on K we have (see, for example, [20, Lemma 1.30])

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

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

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

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

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

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

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

Combining (33) with (32) we have

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

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

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

where the last inequality follows from (34). Combining (30) with (35), we obtain that

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

□

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 number* $\underline{\eta}$ *and the sequence* $\{\eta_k\}$ *be chosen such that*

$$0 < \underline{\eta} \leq \eta_k \leq \min\left\{1, \frac{2\nu+\varepsilon}{\nu}, \frac{\gamma\|\nabla f(w_k)\|}{4L}\right\} \quad \forall k. \tag{36}$$

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

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

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 Take k with $\|w_k - \hat{w}\| \geq \delta$. From (36) we have

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

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

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

which implies

$$f(w_{k+1}) - \hat{f} \leq \theta\left(f(w_k) - \hat{f}\right). \tag{37}$$

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

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

This and (33) imply

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

□

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 projection methods considered in the previous section are applicable to the (high order) discretization of the problem recently developed in [21,24]. (This also applies to the conditional gradient method, where the analysis is similar). 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 [24] that are necessary for understanding the implementation of the GPM 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

It will be convenient to introduce the space $H := ((\mathbb{R}^2)^m)^N$ consisting of vectors $w = (w_0, \dots, w_{N-1})$ with $w_i = (w_i^1, \dots, w_i^m)$ and $w_i^j = (u_i^j, v_i^j) \in \mathbb{R}^2$. We regard this space as a Hilbert space with the scalar product

$$\langle w, \tilde{w}\rangle := \frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=1}^{m}\langle w_i^j, \tilde{w}_i^j\rangle, \quad \langle w_i^j, \tilde{w}_i^j\rangle := u_i^j\tilde{u}_i^j + v_i^j\tilde{v}_i^j$$

The scalar product is normalized by division by N since below N will be a "large" number and the sum will be a proxy for integration on a fixed interval $[0, T]$ by using values on a mesh with size $h = T/N$. We also denote $|w_i| := \sqrt{\sum_{j=1}^m |w_i^j|^2}$, $|w_i^j|^2 := (|u_i^j|^2 + |v_i^j|^2)$. The l_1, l_2, and l_∞ norms in H will be respectively

$$\|w\|_1 := \frac{1}{N}\sum_{i=0}^{N-1}|w_i|, \quad \|w\|_2 := \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}|w_i|^2}, \quad \|w\|_\infty = \max_i |w_i|. \tag{38}$$

Clearly, the inequality $\|w\|_1 \le \|w\|_2 \le \|w\|_\infty$ holds for every $w \in H$.

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 $\|\cdot\|_2$.

We begin with some assumptions concerning the problem (1)–(3).

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 $B(t)^\top S(t)$ 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

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

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. [26]. 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 the triple $(\hat{x}, \hat{u}, \hat{p})$ 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. Following [10], 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 linear growth in a neighborhood of any zero.

Assumption A3 (strict bang–bang property)

There exist real numbers $\alpha, \tau > 0$ such that for all $j \in \{1, \ldots, m\}$ and $s \in [0, T]$ with $\hat{\sigma}^j(s) = 0$ (the j-th component of $\hat{\sigma}$) we have

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

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

3.2 High-Order Time-Discretization

In this subsection we recall the discretization scheme for problem (1)–(3) presented in [24], 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

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

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

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

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

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 [21], 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 $\varphi_1(\alpha) := \frac{1}{4}(-1 + 2\alpha + \alpha^2)$ and $\varphi_2(\alpha) := \frac{1}{4}(1 + 2\alpha - \alpha^2)$.

For the subsequent analysis it will be important that the set $Z \subset \mathbb{R}^2$ is strongly convex. This is evident from Fig. 1, but the calculation of a modulus γ is cumbersome and we skip the details. In this calculation we use Theorem 1 in [28] (expressing γ 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 [21, Sect. 4]. The number $\gamma = 1/\sqrt{32}$ turns out to be a modulus of strong convexity of Z.

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

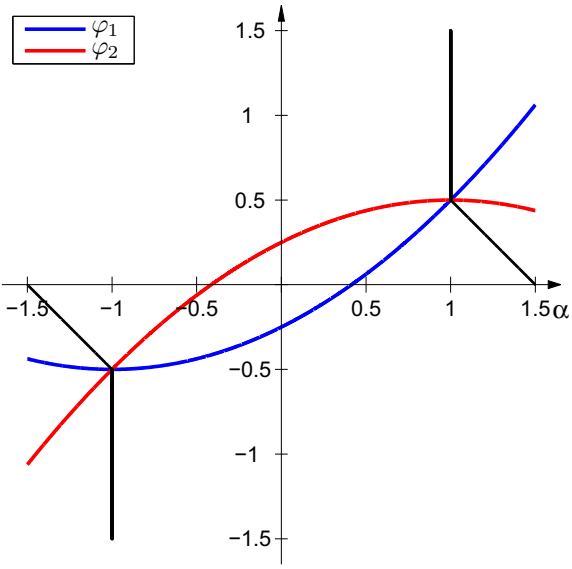

Fig. 1 The set Z as the area between the two parabolas φ_1 (lower) and φ_2 (upper)

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

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,

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

We also introduce the discrete adjoint equation (see formula (3.11) in [24])

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

$i = N-1, \ldots, 0$, with the end condition

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

Section 3.3 in [24] presents a construction which for every sequence $w = (w_0, \ldots w_{N-1}) \in K$ 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 any component $j = 1, \ldots, m$ of $u(t) = \Phi^h(w)(t)$ to $[t_i, t_{i+1}]$ depends only on w_i^j. Moreover, the following equalities hold (see (3.14) in [24]): for every $w = ((u_0, v_0), \ldots, (u_{N-1}, v_{N-1}))$

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

In addition, the function Φ^h has the important property that there exists a constant $\tilde{c}$ independent of N such that for every i, j and $w_i^j, \tilde{w}_i^j \in Z$

$$\int_{t_i}^{t_{i+1}} \left| [\Phi^h(w) - \Phi^h(\tilde{w})]^j \right| \mathrm{d}t \le \frac{\tilde{c}}{N} |w_i^j - \tilde{w}_i^j|.$$

Clearly, this implies

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

Below we will use the metric

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

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

The following theorem is extracted from Theorem 3.1 in [24].

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, 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 $(\hat{x}^N, \hat{w}^N) = \{(\hat{x}_i^N, \hat{w}_i^N)\}$ and for every such solution and the corresponding discrete adjoint sequence $(\hat{p}_0^N, \ldots, \hat{p}_N^N)$ solving (47), (48), the following error estimate holds:*

$$\max_{i=0,\ldots,N} \left(|\hat{x}_i^N - \hat{x}(t_i)| + |\hat{p}_i^N - \hat{p}(t_i)| \right) + d^\# \left(\Phi^h(\hat{w}^N), \hat{u} \right) \le c\,h^2, \tag{51}$$

where c is independent of N.

We mention that the above discretization scheme is meaningful even without assuming (A2) and (A3). These assumptions 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 $w = (w_0, \dots, w_{N-1}) \in K$, $w_i = ((u_i^1, v_i^1), \dots (u_i^m, v_i^m)) \in Z^m$, with the given initial condition x_0.

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

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 the second term in the right hand side of (3.12) in [24])

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

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.

Next, we present five technical lemmas which are needed in the proof of the main result in this section—Proposition 3.1. In the proofs, $c_1, c_2, \dots$ 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.

Lemma 3.1 *There exist constants c' and c'' independent of h, such that for every $w', w'' \in K$ and $\Delta w \in K - K$*

$$|\langle \nabla f^h(w') - \nabla f^h(w''), \Delta w\rangle| \le c'\|w' - w''\|_1 \, \|\Delta w\|_1 + c''h^2 \sum_{i=1}^{N-1} |u_i' - u_i''| \, |\Delta u_i|$$
$$\le c'\|w' - w''\|_1 \, \|\Delta w\|_1 + c''h\|w' - w''\|_2 \, \|\Delta w\|_2,$$

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'':

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

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

$$|\langle \nabla f^h(w') - \nabla f^h(w''), \Delta w\rangle| \le c_1 \left(\|x' - x''\|_\infty + \|p' - p''\|_\infty\right) \|\Delta w\|_1$$
$$+ c_2h^2 \sum_{i=1}^{N-1} |u_i' - u_i''| \, |\Delta u_i|,$$

which together with (57) and (58) implies the firts inequality in the lemma. The second one follows by application of the Cauchy–Schwarz inequality and the definition of the norms. □

Lemma 3.2 *There exists a number c^* such that for every natural number N, for every $\bar{w} \in K$ and for every $\Delta \in L_2([0, T]; \mathbb{R}^m)$*

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

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

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

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 in [24, Sect. 4] (see (4.5) there) imply that for $t \in [t_i, t_{i+1}]$

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

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:

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

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

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

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

Lemma 3.3 *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 [24]), 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 3.1). □

We remind that $\hat{w}^N \in K$ denoted in Theorem 3.1 an optimal control sequence in the discrete problem (53). Further it will be convenient to skip the superscript N in this notation.

Lemma 3.4 *There exist numbers N_0 and δ_1 such that for every $N \ge N_0$*

$$\langle \nabla f^h(\hat{w}), w - \hat{w}\rangle \ge \frac{\alpha\gamma}{64} h\|w - \hat{w}\|_2^2 \quad \textit{for every } w \in K \textit{ with } \|w - \hat{w}\|_2 \ge \delta_1\sqrt{h}$$

(α *is the number from Assumption* (A3) *and* $\gamma \ge 1/\sqrt{32}$ *is a modulus of strong convexity of* Z).

Proof The following expression is obtained in [24] (see formula (4.5) there, applied for $t = t_{i+1}$):

$$B(t_{i+1})^\top \hat{p}_{i+1}^N + S(t_{i+1})^\top \hat{x}_{i+1}^N = (B_i + hC_i)^\top \hat{p}_{i+1}^N + B(t_i)^\top S(t_i) \int_{t_i}^{t_{i+1}} u(s)\,\mathrm{d}s + S(t_i)^\top (I + hA(t_i))\hat{x}_i^N + h\dot{S}(t_i)^\top \hat{x}_i^N + O(h^2),$$

where $u = \Phi^h(\hat{w})$. Comparing this with the expression (55) we see that

$$B(t_{i+1})^\top \hat{p}_{i+1}^N + S(t_{i+1})^\top \hat{x}_{i+1}^N = \nabla_{u_i} f^h(\hat{w}) + \nabla_{v_i} f^h(\hat{w}) + O(h^2).$$

Then using Theorem 3.1 we obtain that

$$\begin{aligned} &|\hat{\sigma}(t_{i+1}) - \nabla_{u_i} f^h(\hat{w}) - \nabla_{v_i} f^h(\hat{w})| \\ &\quad \le \left| B(t_{i+1})^\top \hat{p}(t_{i+1}) + S(t_{i+1})^\top \hat{x}(t_{i+1}) - B(t_{i+1})^\top \hat{p}_{i+1}^N - S(t_{i+1})^\top \hat{x}_{i+1}^N \right| \le \bar{c}h^2, \end{aligned}$$

where $\bar{c}$ is an appropriate constant. Written for the jth components of the vectors in the left-hand side, the inequality becomes

$$|\hat{\sigma}^j(t_{i+1}) - \nabla_{u_i^j} f^h(\hat{w}) - \nabla_{v_i^j} f^h(\hat{w})| \le \bar{c}h^2, \quad j = 1, \dots, m. \tag{59}$$

Assumption (A3) implies that there exist a natural number r and a real number $\tau_0 \in (0, \tau)$ such that every component $\hat{\sigma}^j$ of $\hat{\sigma}$ has at most r zeros in $[0, T]$, and

$|\hat{\sigma}^j(t)| \geq \alpha\tau'$ every $\tau' \in (0, \tau_0]$ and for every t which does not belong to a τ'-neighborhood of a zero of $\hat{\sigma}^j$.

Now, let us define $\delta_1 := M\sqrt{2mr/T}$, where M is the diameter of the set Z (which is $\sqrt{5}$). Moreover, define the natural number N_0 as bigger than $4\bar{c}T/\alpha$, so that $\bar{c}h \leq \alpha/4$.

Let $w = (w_0, \dots, w_{N-1})$ with $w_i = (w_i^1, \dots, w_i^m)$ and $w_i^j = (u_i^j, v_i^j) \in Z$ be arbitrarily chosen. Due to the γ-strong convexity of Z we have that

$$y_i^j := \frac{1}{2}(w_i^j - \hat{w}_i^j) + \frac{\gamma}{8}|w_i^j - \hat{w}_i^j|^2\zeta_i^j \in Z$$

for every $\zeta_i^j \in \mathbb{R}^2$ with $|\zeta_i^j| \leq 1$. With the choice $\zeta_i^j = -\nabla_{w_i^j} f^h(\hat{w})/|\nabla_{w_i^j} f^h(\hat{w})|$ (whenever the denominator is non-zero) we obtain exactly in the same way as in the proof of Proposition 2.1 that

$$\begin{aligned}\langle \nabla f^h(\hat{w}), w - \hat{w}\rangle &= \frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=1}^{m}\left\langle \nabla_{w_i^j} f^h(\hat{w}), w_i^j - \hat{w}_i^j\right\rangle \\ &= 2\langle \nabla f^h(\hat{w}), y - \hat{w}\rangle - \frac{\gamma}{4N}\sum_{i=0}^{N-1}\sum_{j=1}^{m}|w_i^j - \hat{w}_i^j|^2\left\langle \nabla_{w_i^j} f^h(\hat{w}), \zeta_i^j\right\rangle \\ &\geq \frac{\gamma}{4N}\sum_{i=0}^{N-1}\sum_{j=1}^{m}|w_i^j - \hat{w}_i^j|^2\left|\nabla_{w_i^j} f^h(\hat{w})\right|.\end{aligned}$$

Denote by Ω^j the set of all indexes i such that $|\nabla_{w_i^j} f^h(\hat{w})| < \alpha h/8$. Then

$$\begin{aligned}\langle \nabla f^h(\hat{w}), w - \hat{w}\rangle &\geq \frac{\gamma}{4N}\sum_{j=1}^{m}\sum_{i\notin\Omega^j}\frac{\alpha h}{8}|w_i^j - \hat{w}_i^j|^2 \\ &= \frac{\gamma\alpha h}{32N}\left[\sum_{j=1}^{m}\sum_{i=0}^{N-1}|w_i^j - \hat{w}_i^j|^2 - \sum_{j=1}^{m}\sum_{i\in\Omega^j}|w_i^j - \hat{w}_i^j|^2\right] \\ &= \frac{\gamma\alpha h}{32}\left[\|w - \hat{w}\|_2^2 - \frac{1}{N}\sum_{j=1}^{m}\sum_{i\in\Omega^j}|w_i^j - \hat{w}_i^j|^2\right]. \qquad (60)\end{aligned}$$

Consider an arbitrary $i \in \Omega^j$. Since $|\nabla_{w_i^j} f^h(\hat{w})| < \alpha h/8$, according to (59) we have

$$|\hat{\sigma}^j(t_{i+1})| < \frac{\alpha h}{4} + \bar{c}h^2 \leq \frac{\alpha h}{2},$$

where we also use that $\bar{c}h \leq \alpha/4$. Then t_{i+1} belongs to the $h/2$-neighborhood of some zero of $\hat{\sigma}^j$ (see the paragraph after (59)). Then no other point $t_k \neq t_{i+1}$ belongs to this neighborhood. Since $\hat{\sigma}^j$ has at most r zeros, the set Ω^j consists of at most r points.

Then continuing (60) we obtain

$$\langle \nabla f^h(\hat{w}), w - \hat{w} \rangle \geq \frac{\gamma \alpha h}{32} \left[\|w - \hat{w}\|_2^2 - \frac{1}{N} rmM^2 \right]$$

Now assume that $\|w - \hat{w}\|_2 \geq \delta_1 \sqrt{h} = M\sqrt{2mr/T}\sqrt{h}$. Then

$$\langle \nabla f^h(\hat{w}), w - \hat{w} \rangle \geq \frac{\gamma \alpha h}{32} \|w - \hat{w}\|_2^2 \left[1 - \frac{1}{(\delta_1)^2 h} \frac{rmM^2}{N} \right] = \frac{\gamma \alpha h}{64} \|w - \hat{w}\|_2^2.$$

The proof is complete. □

Lemma 3.5 *There exists a constant $\nu_1 > 0$ such that*

$$\langle \nabla f^h(w) - \nabla f^h(\hat{w}), w - \hat{w} \rangle \geq -\nu_1 h^2 (\|w - \hat{w}\|_2 + h) \quad \textit{for every } w \in K. \tag{61}$$

Proof As before, let $\hat{u}$ be the optimal control in the continuous-time problem (1)–(3). Denote

$$\tilde{w}_i = \left(\frac{1}{h} \int_{t_i}^{t_{i+1}} \hat{u}(t)\, \mathrm{d}t, \frac{1}{h^2} \int_{t_i}^{t_{i+1}} (t - t_i) \hat{u}(t)\, \mathrm{d}t \right).$$

Denote $\mu_i := \mathrm{meas}\{t \in [t_i, t_{i+1}] : \Phi^h(\hat{w})(t) \neq \hat{u}(t)\}$. According to Theorem 3.1, $\sum_{i=0}^{N-1} \mu_i \leq ch^2$. Due to (49), we have

$$|\tilde{w}_i - \hat{w}_i| \leq \frac{c_1}{h} \mu_i,$$

hence

$$\|\tilde{w} - \hat{w}\|_1 = \frac{1}{N} \sum_{i=0}^{N-1} |\tilde{w}_i - \hat{w}_i| \leq \frac{h}{T} \sum_{i=0}^{N-1} \frac{c_1}{h} \mu_i \leq c_2 h^2. \tag{62}$$

Moreover,

$$\|\tilde{w} - \hat{w}\|_2 \leq \sqrt{\|\tilde{w} - \hat{w}\|_\infty \|\tilde{w} - \hat{w}\|_1} \leq c_3 h. \tag{63}$$

Now we denote the left-hand side of (61) by D and represent $D = D_1 + D_2 + D_3$, where

$$\begin{aligned} D_1 &:= \langle \nabla f^h(\tilde{w}) - \nabla f^h(\hat{w}), w - \hat{w} \rangle, \\ D_2 &:= \langle \nabla f^h(w) - \nabla f^h(\tilde{w}), w - \tilde{w} \rangle, \\ D_3 &:= \langle \nabla f^h(w) - \nabla f^h(\tilde{w}), \tilde{w} - \hat{w} \rangle. \end{aligned}$$

We shall estimate each of these terms separately. From Lemma 3.1 we obtain

$$\begin{aligned} D_1 &\geq -c' \|\tilde{w} - \hat{w}\|_1 \|w - \hat{w}\|_1 - c'' h \|\tilde{w} - \hat{w}\|_2 \|w - \hat{w}\|_2 \\ &\geq -c' c_2 h^2 \|w - \hat{w}\|_1 - c'' c_3 h^2 \|w - \hat{w}\|_2 \geq -c_4 h^2 \|w - \hat{w}\|_2. \end{aligned}$$

In order to estimate D_2 we use Lemma 3.2 and the definition of $\tilde{w}$:

$$D_2 \geq \frac{1}{T}\langle \nabla f(\Phi^h(w)) - \nabla f(\hat{u}),\ \Phi^h(w) - \hat{u}\rangle - \frac{2}{T}c^* h^2 \|\Phi^h(w) - \hat{u}\|_1.$$

The first term in the right-hand side is non-negative due to (56). Hence, using also (50), we obtain that

$$\begin{aligned} D_2 &\geq -c_5 h^2 \|\Phi^h(w) - \hat{u}\|_1 \geq -c_6 h^2 \|w - \tilde{w}\|_1 \geq -c_6(\|w - \hat{w}\|_1 + \|\hat{w} - \tilde{w}\|_1) \\ &\geq -c_6(\|w - \hat{w}\|_1 + c_2 h^2). \end{aligned}$$

For estimating D_3 we use again Lemma 3.1, (62) and (63) :

$$\begin{aligned} D_3 &\geq -c' \|w - \tilde{w}\|_1 \, \|\tilde{w} - \hat{w}\|_1 - c'' h \|w - \tilde{w}\|_2 \, \|\tilde{w} - \hat{w}\|_2 \\ &\geq -c_7 h^2 \|w - \tilde{w}\|_1 - c_8 h^2 \|w - \tilde{w}\|_2 \geq -c_9 h^2 (\|w - \hat{w}\|_2 + \|\hat{w} - \tilde{w}\|_2) \\ &\geq -c_{10} h^2 (\|w - \hat{w}\|_2 + h). \end{aligned}$$

Combining the estimations for D_1, D_2 and D_3 we obtain (61). □

Proposition 3.1 *On the assumptions* (A1)–(A3)*, the function* f^h *is* L*-smooth on* K *and there exist numbers* N_0, $\nu_0 > 0$ *and* δ_0 *such that for every* $N \geq N_0$ *condition* (7) *in Proposition* 2.1 (*hence, also the assumptions in Proposition* 2.2 *and Theorems* 2.1 *and* 2.2) *is fulfilled for problem* (53) *with* $\nu = \nu_0 h$ *and* $\delta = \delta_0 \sqrt{h}$.

Proof The L-smoothness of f^h on K was proved in Lemma 3.3. Now, take an arbitrary $w \in K$ and consider

$$\langle \nabla f(w), w - \hat{w}\rangle = \langle \nabla f^h(\hat{w}), w - \hat{w}\rangle + \langle \nabla f^h(w) - \nabla f^h(\hat{w}), w - \hat{w}\rangle.$$

Using Lemmas 3.4 and 3.5 we estimate

$$\langle \nabla f(w), w - \hat{w}\rangle \geq \frac{\alpha\gamma}{64} h \|w - \hat{w}\|_2^2 - \nu_1 h^2 (\|w - \hat{w}\|_2 + h)$$

for every $w \in K$ with $\|w - \hat{w}\|_2 \geq \delta_1 \sqrt{h}$. Then for such w it holds that

$$\begin{aligned} \langle \nabla f(w), w - \hat{w}\rangle &\geq \frac{\alpha\gamma}{64} h \|w - \hat{w}\|_2^2 - \frac{\nu_1 h^2}{\delta_1 \sqrt{h}} \|w - \hat{w}\|_2^2 - \frac{\nu_1 h^3}{(\delta_1)^2 h} \|w - \hat{w}\|_2^2 \\ &= h \|w - \hat{w}\|_2^2 \left(\frac{\alpha\gamma}{64} - \frac{\nu_1 h^{1/2}}{\delta_1} - \frac{\nu_1 h}{(\delta_1)^2} \right). \end{aligned}$$

Then the claim of the proposition holds for all sufficiently small h. □

Let us interpret the above proposition in view of Theorem 2.1 for convergence of the gradient projection method (GPM) applied to the discrete problem (52) and (53). The linear rate of convergence, μ, as estimated in this theorem, may approach 1 when

ν approaches zero. In the same time, Proposition 3.1 estimates ν as proportional to h. Thus, although the convergence is linear, its rate, μ, may be close to one. Even more, this rate of convergence is valid only until an accuracy δ is achieved (see Theorem 2.1). The number δ in Proposition 3.1 is estimated as proportional to $\sqrt{h}$. Thus the convergence of the GPM does not seem to be consistent with the $O(h^2)$-approximation that the discretization method provides. On the other hand, the fact that the GPM is proved to converge (even linearly, in the sense of Theorem 2.1) is remarkable. Indeed, if the Euler discretization scheme is applied to the original problem (1)–(3) (as in most of the literature), the resulting discrete-time problem may fail to be convex, and no results about the rate of convergence of the GPM are available in the literature, to the authors' knowledge.

We do not present the convergence analysis of the CGM for problem (52) and (53), which is rather similar.

3.4 Implementation of the Gradient Methods

Now, we shall describe the implementation of the GPM and the CGM to the specific mathematical programming problem defined by (53) and (52).

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 [24, Sect. 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 two-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 [21, Sect. 4]:

$$N_Z(\alpha,\beta) = \begin{cases} \emptyset & \text{if} \quad (\alpha,\beta) \notin Z, \\ \left\{\alpha\,(\lambda, \mu-\lambda)^\top : \mu \geq 0, \lambda \geq 0\right\} & \text{if} \quad \alpha \in \{-1,1\}, \\ \left\{\mu\,(\zeta+\alpha, -2\zeta)^\top : \mu \geq 0\right\} & \text{if} \quad \alpha \in (-1,1) \wedge \beta \in \{\varphi_1(\alpha), \varphi_2(\alpha)\}, \\ \{0\} & \text{if} \quad \alpha \in (-1,1) \wedge \beta \in (\varphi_1(\alpha), \varphi_2(\alpha)), \end{cases} \tag{64}$$

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

$$P_Z(\xi) \in \xi - N_Z(P_Z(\xi)). \tag{65}$$

Therefore, using the formula (64), one can explicitly calculate $P_Z(\xi)$ as

$$P_Z(u, v) = \begin{cases} (u, v) & \text{if} \quad (u, v) \in Z, \\ (1, \frac{1}{2}) & \text{if} \quad u \geq 1 \ \text{ and } \ u + v \geq \frac{3}{2}, \\ (-1, -\frac{1}{2}) & \text{if} \quad u \leq -1 \ \text{ and } \ u + v \leq -\frac{3}{2}, \\ (\alpha_1, \varphi_1(\alpha_1)) & \text{if} \quad u > -1 \ \text{ and } \ u + v < \frac{3}{2} \ \text{ and } \ v < \varphi_1(u), \\ (\alpha_2, \varphi_2(\alpha_2)) & \text{if} \quad u < 1 \ \text{ and } \ u + v < -\frac{3}{2} \ \text{ and } \ v > \varphi_2(u), \end{cases} \tag{66}$$

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

$$\alpha^3 + 3\alpha^2 + (9 - 4v)\alpha - 8u - 4v - 1 = 0, \tag{67}$$

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

$$\alpha^3 - 3\alpha^2 + (9 + 4v)\alpha - 8u - 4v + 1 = 0. \tag{68}$$

Indeed, the first three cases in the representation (66) 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 Fig. 1). From (64), we have

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

Combining this with (65), 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 (67). 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 (64), can be explicitly represented (see [21]) by the following simple formula:

$$(\alpha, \beta) = \begin{cases} (-1, -1/2) & \text{if} \quad \xi_1 \le 0 \ \text{ and } \ \xi_1 + \xi_2 \le 0, \\ (1, 1/2) & \text{if} \quad \xi_1 > 0 \ \text{ and } \ \xi_1 + \xi_2 \ge 0, \\ (-1 - 2\xi_1/\xi_2, \varphi_1(\alpha)) & \text{if} \quad \xi_1 > 0 \ \text{ and } \ \xi_1 + \xi_2 < 0, \\ (1 + 2\xi_1/\xi_2, \varphi_2(\alpha)) & \text{if} \quad \xi_1 \le 0 \ \text{ and } \ \xi_1 + \xi_2 > 0. \end{cases} \tag{69}$$

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 subsection, we present some numerical experiments for the example of an affine linear-quadratic optimal control problem given in [24].

Example 3.1

$$\begin{array}{ll} \text{minimize} & -by(1) + \int_0^1 \frac{1}{2}\,(x(t))^2 \,\mathrm{d}t \\ \text{subject to} & \dot{x}(t) = y(t), \quad x_1(0) = a \\ & \dot{y}(t) = u(t), \quad y(0) = 1. \\ & u(t) \in [-1, 1]. \end{array} \tag{70}$$

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 [24], we choose $a = 1, b = 0.1$, then $\tau = 0.492487520$ is a simple zero of the switching function, thus Assumption (A3) is fulfilled. The exact optimal control is

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

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 $\|w_{k+1} - w_k\| \le 10^{-6}$ for the GPM and $\|x_k - w_k\| \le 10^{-6}$ for the CGM.

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

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 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

1 when N increases. This is to be expected, since according to Theorem 2.1, the rate μ_N of linear convergence approaches 1 when ν goes to zero, and according to Proposition 3.1 ν estimated as proportional to h. 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 [22, Theorem 3.2]. We emphasize that due to the second order accuracy of discretization, the mesh size N does not 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 Sect. 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.

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 [21, 24] 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 [17], 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.

Acknowledgements Open access funding provided by Austrian Science Fund (FWF).

References

1. Alt, W., Baier, R., Lempio, F., Gerdts, M.: Approximations of linear control problems with bang-bang solutions. Optimization **62**, 9–32 (2013)
2. Alt, W., Schneider, C., Seydenschwanz, M.: Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang-bang solutions. Appl. Math. Comput. **287**, 104–124 (2016)
3. Alt, W., Felgenhauer, U., Seydenschwanz, M.: Euler discretization for a class of nonlinear optimal control problems with control appearing linearly. Comput. Optim. Appl. (2017). https://doi.org/10.1007/s10589-017-9969-7
4. Attouch, H., Aze, D.: Approximation and regularization of arbitrary functions in Hilbert spaces by the Lasry-Lions method. Ann. Inst. Henri Poincaré **3**, 289–312 (1993)
5. Balashov, M.V.: Maximization of a function with Lipschitz continuous gradient. J. Math. Sci. **209**, 12–18 (2015)
6. Balashov, M.V., Golubev, M.O.: About the Lipschitz property of the metric projection in the Hilbert space. J. Math. Anal. Appl. **394**, 545–551 (2012)
7. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
8. Demyanov, V.F., Rubinov, A.M.: Approximate Methods in Optimization Problems. Elsevier, New York (1970)
9. Dunn, J.C.: Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. SIAM J. Control Optim. **17**, 187–211 (1979)
10. Felgenhauer, U.: On stability of bang-bang type controls. SIAM J. Control Optim. **41**, 1843–1867 (2003)
11. Felgenhauer, U.: Discretization of semilinear bang-singular-bang control problems. Comput. Optim. Appl. **64**, 295–326 (2016)
12. Felgenhauer, U.: A Newton-type method and optimality test for problems with bang-singular-bang optimal control. Pure Appl. Funct. Anal. **1**, 197–215 (2016)
13. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Res. Logist. Q. **3**, 149–154 (1956)
14. Garber, D., Hazan, E.: Faster rates for the Frank-Wolfe method over strongly-convex sets. In: ICML'15, vol. 37, pp. 541–549 (2015)
15. Golubev, M.O.: Gradient projection method for convex function and strongly convex set. IFAC-PapersOnLine **48**, 202–205 (2015)
16. Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and Their Applications. Academic Press, New York (1980)
17. Lempio, F., Veliov, V.M.: Discrete approximations of differential inclusion. Bayreuth. Math. Schr. **54**, 149–232 (1998)
18. Luenberger, D.G., Ye, Y.: Linear and Nonlinear Programming. Springer, New York (2008)
19. Nesterov, Y.: Introductory Lectures on Convex Optimization. Springer, New York (2013)
20. Peypouquet, J.: Convex Optimization in Normed Spaces: Theory, Methods and Examples. Springer, Dordrecht (2015)
21. Pietrus, A., Scarinci, T., Veliov, V.M.: High order discrete approximations to Mayer's problems for linear systems. SIAM J. Control Optim. **56**, 102–119 (2018)
22. Preininger, J., Vuong, P.: On the convergence of the gradient projection method for optimal control problems with bang-bang solutions. Comput. Optim. Appl. **70**, 221–238 (2018)
23. Preininger, J., Scarinci, T., Veliov, V.M.: Metric regularity properties in bang-bang type linear-quadratic optimal control problems. Set-Valued Var. Anal. https://doi.org/10.1007/s11228-018-0488-1. Avail-

able as Research Report, 2017-07, ORCOS. TU Wien, Wien. https://orcos.tuwien.ac.at/fileadmin/t/orcos/Research_Reports/2017-07.pdf (2017)
24. Scarinci, T., Veliov, V.M.: Higher-order numerical schemes for linear quadratic problems with bang-bang controls. Comput. Optim. Appl. **69**, 403–422 (2018). https://doi.org/10.1007/s10589-017-9948-z
25. Seydenschwanz, M.: Convergence results for the discrete regularization of linear-quadratic control problems with bang-bang solutions. Comput. Optim. Appl. **61**, 731–760 (2015)
26. Veliov, V.M.: On the time-discretization of control systems. SIAM J. Control Optim. **35**, 1470–1486 (1997)
27. Veliov, V.M.: Error analysis of discrete approximation to bang-bang optimal control problems: the linear case. Control Cybern. **34**, 967–982 (2005)
28. Vial, J.-P.: Strong convexity of sets and functions. J. Math. Econ. **9**, 187–205 (1982)