TECHNISCHE UNIVERSITÄT WIEN

Dissertation

Metric regularity and approximations of generalized equations with applications to optimal control

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Univ.Prof. Dr.techn. Vladimir M. Veliov

E105 - Institut für Stochastik und Wirtschaftsmathematik
eingereicht an der Technischen Universität Wien
an der Fakultät für Mathematik und Geoinformation

von

Jakob Preininger, MSc
Matr.Nr.: 0704849

Wien, im Mai 2018

Abstract

The aim of this thesis is to study regularity properties and approximations of generalized equations and to apply them for optimal control problems. The thesis is cumulative and consists of four published or submitted for publication papers.

The first one investigates the convergence properties of Newton-type methods for solving generalized equations. Classical results use the properties of metric regularity or strong metric regularity of the generalized equation at the solution to show convergence of the Newton method when the initial point is in a neighborhood of the solution. In contrast theorems of Kantorovich-type impose regularity conditions on the initial point rather than the solutions and therefore allow an a priori convergence analysis which is more useful for practical purposes. A known result of Kantorovich-type for generalized equations requires the single-valued part of the generalized equations to be differentiable with Lipschitz continuous derivative. A new nonsmooth version of this result showing linear convergence of the Newton method is proved.

The second paper introduces uniform versions of metric regularity and strong metric regularity on compact sets and uses them to analyze two path-following schemes for tracking a solution trajectory of a differential generalized equation, which use ideas of the Euler/Heun method and the Newton method simultaneously.

The third paper studies the necessary optimality condition for solutions of general optimal control problems in Bolza form obtained by the Pontryagin maximum principle. This condition can be rewritten as a generalized equation in suitable Sobolev spaces. Hence results about Newton-type methods from the first part can be used for solving these problems. Known results showing regularity mostly assume continuity of the optimal control which is not fulfilled for some of the most basic Bolza problems, namely those that are linear in control. Usually these problems have optimal controls of bang-bang type, i.e. they contain a finite number of switching points where the control is discontinuous. Under weak convexity assumptions metric subregularity as well as strong bimetric regularity of the generalized equations associated with these problems are proved and used to show a convergence result about the Newton method applied to such problems.

The final paper deals with the gradient projection method which, among other things, can be used to solve the linearized problems which appear when using the Newton method on the generalized equations obtained in the previous part. A new result about the convergence speed of the gradient projection method in case of bang-bang controls is proved and some analytical and numerical examples are given.

Kurzfassung

Ziel dieser Arbeit ist es, Regularitätseigenschaften und Approximationen von verallgemeinerten Gleichungen zu untersuchen und auf optimale Steuerungsprobleme anzuwenden. Die Arbeit ist kumulativ und besteht aus vier veröffentlichten oder zur Veröffentlichung eingereichten Artikeln.

Der erste untersucht die Konvergenzeigenschaften von Newton- und Newton-ähnlichen Verfahren zur Lösung verallgemeinerter Gleichungen. Klassische Ergebnisse verwenden die Eigenschaften der metrischen Regularität oder der starken metrischen Regularität der verallgemeinerten Gleichung im Lösungspunkt, um die Konvergenz der Newton-Methode zu zeigen, wenn der Anfangspunkt in einer Umgebung der Lösung liegt. Im Gegensatz dazu verwenden Theoreme vom Kantorovich-Typ Regularitätsbedingungen im Anfangspunkt und nicht im Lösungspunkt und erlauben daher eine a priori Konvergenzanalyse, die für praktische Zwecke nützlicher ist. Ein bekanntes Kantorovich-Theorem für verallgemeinerte Gleichungen erfordert, dass der einwertige Teil der verallgemeinerten Gleichungen differenzierbar mit Lipschitz-stetiger Ableitung ist. Hier wird eine neue, nicht glatte Version dieses Ergebnisses, mit linearer Konvergenzgeschwindigkeit bewiesen.

Der zweite Artikel führt uniforme Versionen von metrischer Regularität und starker metrischer Regularität auf kompakten Mengen ein und verwendet diese, um zwei path-following schemes zum Auffinden einer Lösungstrajektorie einer differenziellen verallgemeinerten Gleichung (differential generalized equation), die gleichzeitig die Ideen des Euler/Heun-Verfahrens und des Newton-Verfahrens benutzen, zu analysieren.

Der dritte Artikel untersucht die notwendige Optimalitätsbedingung für Lösungen von allgemeinen optimalen Steuerungsproblemen in Bolza-Form, die durch das Pontryagin-Maximum-Prinzip erhalten werden. Diese Bedingung kann als verallgemeinerte Gleichung in geeigneten Sobolev-Räumen umgeschrieben werden. Daher können Ergebnisse über das Newton-Verfahren aus dem ersten Teil zur Lösung dieser Probleme verwendet werden. Bekannte Ergebnisse, die Regularität zeigen, nehmen meist die Stetigkeit der optimalen Steuerung an, die für einige der grundlegendsten Bolza-Probleme nicht erfüllt ist, nämlich jenen, die in der Steuerung linear sind. Üblicherweise haben diese Probleme optimale Steuerungen vom Bang-Bang-Typ, d.h. sie enthalten eine endliche Anzahl von switching points, an denen die Steuerung unstetig ist. Unter schwachen Konvexitätsannahmen werden die metrische Subregularität sowie die starke bimetrische Regularität der verallgemeinerten Gleichungen dieser Probleme bewiesen und verwendet, um ein Ergebnis zur Konvergenz der Newton-Methode für diese Probleme zu zeigen.

Der letzte Artikel beschäftigt sich mit der Gradientenprojektionsmethode, die unter anderem dazu verwendet werden kann, die linearisierten Probleme zu lösen, die auftreten, wenn das Newton-Verfahren auf die im letzten Teil erhaltenen verallgemeinerten Gleichungen angewendet wird. Ein neues Ergebnis zur Konvergenzgeschwindigkeit der Gradientenprojektionsmethode im Falle von Bang-Bang-Steuerungen wurde nachgewiesen und einige analytische und numerische Beispiele angegeben.

Contents

Acknowledgement

First I would like to thank my supervisor Vladimir Veliov who introduced me to the topics treated in this thesis. He was very patient with me, always gave me guidance when I needed it and gave me the opportunity to make this thesis a reality. Next I thank all my coauthors Radek Cibulka, Asen Dontchev, Tomas Roubal, Teresa Scarinci, Vladimir Veliov and Phan Vuong, who made writing this thesis much more enjoyable and helped me immensely to improve my knowledge on optimal control and mathematics in general. Further I thank the Austrian Science Foundation (FWF) for their support under grant P26640-N25. Last I would like to thank my parents Elisabeth and Helmut Preininger for their support and guidance for this thesis and life in general. Without their help this thesis would not have been possible.

Introduction

Outline

This cumulative thesis consists of four papers: [1] and [4] are published, [3] is conditionally accepted and [2] is submitted for publication.

- [1] Cibulka R., Dontchev A.L., Preininger J., Roubal T., Veliov V.M.: Kantorovich-type Theorems for Generalized Equations. Journal Convex Analysis, 25(2), 459–486 (2018)
- [2] Cibulka R., Preininger J., Roubal T.: On uniform regularity and strong regulartiy. Submitted in Journal of Optimization (2018)
- [3] Preininger J., Scarinci T., Veliov V.M.: Metric regularity properties in bang-bang type linear quadratic optimal control problems. To appear in Journal of Set-Valued and Variational Analysis: Theory and Applications (2018)
- [4] Preininger J., Vuong, P.T.: On the Convergence of the Gradient Projection Method of Optimal Control Problems with Bang-bang Solutions. Computational Optimization and Applications, 70(1), 221–238 (2018)

The contributions of the author of the thesis to the papers is clarified in the following. In [1] the author mainly contributed in the formulation and proof of the main theorem (Theorem 2.2), its Corollary 2.5, the examples in Section 3 and the numerical treatments in Section 5. In [2] the author was involved in making the formulations and proofs in chapter 2 and 4. In [3] the author contributed mainly to the chapters 1-4 and some ideas in chapter 5. In [4] the author was again involved in chapters 1-3, mainly in the ideas and formulations of the main theorems (Theorem 3.2 and Theorem 3.6).

In the following introduction we will give some preliminaries and a summary of the results of these articles. The subsequent chapters consist of the above-mentioned articles.

Preliminaries

Notation

In the following if not stated otherwise X and Y are Banach spaces, $f : X \to Y$ is a Fréchet differentiable single-valued mapping and $F : X \rightrightarrows Y$ is a set valued mapping.

The Newton method

One of the most fundamental algorithms to numerically solve nonlinear equations of the form

$$f(x) = 0, \tag{1}$$

is the Newton method defined as follows. Given an initial point $x_0 \in X$ define a sequence $\{x_k\}$ via the following iteration

$$f(x_k) + Df(x_k)(x_{k+1} - x_k) = 0, \quad k = 0, 1, \dots. \tag{2}$$

When f is "nice enough" (e.g. if f has Lipschitz continuous derivative in a neighborhood of the solution and the derivative at the solution point is invertible) one can show a quadratic convergence rate of this method near the solution. Because of its simplicity and this fast convergence rate the Newton method is one of the most effective methods for solving nonlinear equations. For an in depth discussion on the classical Newton method see, for instance, [24].

Here we focus on the Newton method for the relatively modern notion of generalized equations i.e. inclusions of the form

$$0 \in f(x) + F(x). \tag{3}$$

Then the Newton method looks like the following

$$0 \in f(x_k) + Df(x_k)(x_{k+1} - x_k) + F(x_{k+1}). \tag{4}$$

This general version of the Newton method covers a huge territory of iterative methods in variational analysis, optimization and control. For instance if $F \equiv 0$ then the Newton method reduces to the classical Newton method. If $X = \mathbb{R}^n$, $Y = \mathbb{R}^{p+q}$ and $F \equiv \mathbb{R}^p_- \times 0$ is the product of the non-positive orthant in $\mathbb{R}^p$ with the origin in $\mathbb{R}^q$ then (3) describes a system of p inequalities and q equalities. Further if $Y = X^*$ is the dual of X and $F \equiv N_C$ is the normal cone mapping

$$N_C(x) = \begin{cases} \emptyset & \text{if } x \notin C \\ \{l \in X^* : \langle l, y - x\rangle \leq 0 \ \forall y \in C\} & \text{if } x \in C, \end{cases} \tag{5}$$

where $C \subset X^*$ is a nonempty convex set then (3) represents the variational inequality

$$\text{Find } x \in C \text{ such that } \langle f(x), y - x\rangle \geq 0 \ \forall y \in C. \tag{6}$$

In particular this includes the Karush-Kuhn-Tucker (KKT) optimality conditions of nonlinear programming (NLP) problems. Consider the NLP

$$\text{minimize } f(x) \quad \text{subject to } g(x) \leq 0,\ h(x) = 0, \tag{7}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^p$ and $h : \mathbb{R}^n \to \mathbb{R}^q$. Then the KKT conditions are given as follows

$$\nabla L_x(x, \lambda, \mu) = 0, \quad g(x) \leq 0, \quad h(x) = 0, \quad \mu \geq 0, \quad \langle \mu, g(x)\rangle = 0, \tag{8}$$

where $L : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is the Lagrangian of (7) given by

$$L(x, \lambda, \mu) = f(x) + \langle \lambda, h(x)\rangle + \langle \mu, g(x)\rangle. \tag{9}$$

As the KKT-conditions are a set of equalities and inequalities the Newton-method is applicable. In this context the Newton method is strongly connected to sequential quadratic programming (SQP).

Strong metric regularity

Studying the proof of convergence of the classical Newton algorithm (2) and the exact conditions on f (in particular the invertibility of the derivative $Df(\bar{x})$ at the solution $\bar{x}$) it becomes clear that at its core the Newton algorithm relies on the implicit function theorem, which in turn uses the contraction mapping principle, i.e. a fixed point theorem. To obtain a result about convergence of the Newton method for generalized equations one would expect to need analogous

theorems for generalized equations. This leads to the notion of strong metric regularity first introduced by Robinson in 1980 (see [29]) which guarantees that these theorems work.[1]

Definition 1. A set-valued mapping $F : X \rightrightarrows Y$ is called strongly metrically regular at $\bar{x} \in X$ for $\bar{y} \in Y$ with constant κ if $(\bar{x}, \bar{y}) \in gphF$ and its inverse F^{-1} has a Lipschitz continuous single-valued localization around $\bar{y}$ for $\bar{x}$ with Lipschitz constant κ.

Of fundamental importance is the stability of the property of strong metric regularity under single-valued Lipschitz perturbations. This fact is often called the Lyusternik-Graves theorem. In its easiest form it states the following.

Theorem 2. *Consider a set-valued mapping $F : X \rightrightarrows Y$, a point $(\bar{x}, \bar{y}) \in gph(F)$ and a Lipschitz continuous function $g : X \to Y$ with Lipschitz constant μ. Assume that F is strongly metrically regular at $\bar{x}$ for $\bar{y}$ with constant κ such that $\kappa\mu < 1$. Then the mapping $g+F$ is strongly metrically regular at $\bar{x}$ for $\bar{y} + g(\bar{x})$ with constant $\frac{\kappa}{1-\kappa\mu}$.*

In fact this is the theorem used to show that the contraction mapping principle and in turn the Newton theorem works for generalized equations. A full in depth analysis of this fact and of (strong) metric regularity and its importance for iterative methods in general can be found in [13]. Here we give the simplest version of Newton's theorem for generalized equations.

Theorem 3. *Let $f : X \to Y$ a Fréchet differentiable function with Lipschitz continuous derivative and $F : X \rightrightarrows Y$ a set-valued mapping with closed graph. Assume that $\bar{x}$ is a solution of* (3) *and that $f + F$ is strongly metrically regular at $\bar{x}$ for* 0*. Then there exists a neighborhood O of $\bar{x}$ such that for any starting point $x_0 \in O$ the Newton algorithm* (4) *generates a unique sequence $\{x_k\}$ that stays in O and converges quadratically to $\bar{x}$.*

There are a variety of different versions of this theorem spread in the literature (see e.g. [7] [10], [19], [20]) with slightly different assumptions and corresponding convergence results.

We mention two of those which use the notions similar to strong metric regularity namely (non-strong) metric regularity and strong metric subregularity.

Definition 4. A set-valued mapping $F : X \rightrightarrows Y$ is called metrically regular at $\bar{x} \in X$ for $\bar{y} \in Y$ with constant κ if $(\bar{x}, \bar{y}) \in gphF$ and there are neighborhoods U and V of $\bar{x}$ and $\bar{y}$ respectively such that

$$d(x, F^{-1}(y)) \leq \kappa d(y, F(x)) \quad \forall (x, y) \in U \times V. \tag{10}$$

Theorem 5. *Let $f : X \to Y$ a Fréchet differentiable function with Lipschitz continuous derivative and $F : X \rightrightarrows Y$ a set-valued mapping with closed graph. Assume that $\bar{x}$ is a solution of* (3) *and that $f + F$ is metrically regular at $\bar{x}$ for* 0*. Then there exists a neighborhood O of $\bar{x}$ such that for any starting point $x_0 \in O$ there exists a sequence $\{x_k\}$ fulfilling the Newton algorithm* (4) *which stays in O and converges quadratically to $\bar{x}$.*

The main difference to Theorem 3 is that one no longer obtains a unique sequence but in every iteration multiple choices of the next iterate may be possible and not every choice necessarily leads to a convergent sequence. Therefore this result is more of theoretic interest and strong metric regularity is usually desired for practical purposes.

Definition 6. A set-valued mapping $F : X \rightrightarrows Y$ is called strongly metrically subregular at $\bar{x} \in X$ for $\bar{y} \in Y$ with constant κ if $(\bar{x}, \bar{y}) \in gphF$ and there are neighborhoods U and V of $\bar{x}$ and $\bar{y}$ respectively such that

$$\|x - \bar{x}\| \leq \kappa d(\bar{y}, F(x) \cap V) \quad \forall x \in U. \tag{11}$$

[1] Note however that there are even weaker notions like hemi- and semistability guaranteeing convergence of the Newton method for generalized equations (see e.g. [21]).

Theorem 7. *Let $f : X \to Y$ a Fréchet differentiable function with Lipschitz continuous derivative and $F : X \rightrightarrows Y$ a set-valued mapping with closed graph. Assume that $\bar{x}$ is a solution of* (3) *and that $f + F$ is strongly metrically subregular at $\bar{x}$ for* 0*. Then there exists a neighborhood O of $\bar{x}$ such that if a sequence $\{x_k\}$ is generated by the Newton method* (4) *and has a tail $\{x_k\}_{k\geq k_0}$ with $x_k \in O$ for all $k \geq k_0$ then $\{x_k\}$ is quadratically convergent to $\bar{x}$.*

This theorem has been shown very recently in [10]. In contrast to Theorem 3, here it is not guaranteed that a convergent sequence exists at all. Hence if one only has strong subregularity at the solution one has to prove existence of such a sequence by other means. Then this theorem gives information about the speed of convergence.

The theorems above are all local results. I.e. they impose conditions on the solution and therefore only get convergence in a small neighborhood (of undetermined size) around the solution.

Newton-Kantorovich

Kantorovich [22] was the first to obtain a Newton-type theorem which imposes conditions on the starting point rather than the solution, which makes the theorem far more useful as the conditions can be checked before computing a solution. This idea was expanded to generalized equations (e.g. in [11]) but as Kantorovich's original theorem all the known results focus on the smooth case.

Kantorovich [23] himself noted that to achieve linear convergence to the solution it is not necessary to use the derivative $Df(x_k)$ at the current iteration but using $Df(x_0)$ in every iteration is also sufficient. He called this method the modified Newton process, which today is predominantly known as the chord method. Bartle [8] extended this idea and showed that it is not necessary to choose a derivative $Df(x_k)$ of an iterate x_k at all but any "arbitrary selected point ... sufficiently close to the solution desired" is feasible. In fact it is not important to use a derivative at all. In a nonsmooth setting (i.e. f is continuous but not necessarily differentiable) Qi and Sun [27, Theorem 3.3] proved linear convergence in a Kantorovich-type theorem using suitable linear mappings $A_k : X \to Y$. Note however that the assumption they impose on the mappings A_k are quite strong and restrict the functions f for which the theorem can be used since there are nonsmooth functions f for which the assumptions fail to be satisfied for any linear map.

Technically speaking all these ideas are special cases of what today are known as quasi Newton methods where instead of the exact derivative of the current iteration an approximation of that value is used. More details on Kantorovich's theorem and quasi-Newton methods can be found in various textbooks about the Newton method (e.g. [24]).

The Bolza problem

One of the most important classes of problems in optimal control are problems of Bolza type, i.e. of the form

$$\text{minimize } \psi(x,u) := g(x(T)) + \int_0^T h(t,x(t),u(t))dt \tag{12}$$

subject to

$$\dot{x}(t) = f(t,x(t),u(t)) \text{ for a.e. } t \in [0,T], \quad x(0) = x_0, \tag{13}$$

and

$$u(t) \in U := [-1,1]^m \text{ for a.e. } t \in [0,T]. \tag{14}$$

Here $[0,T]$ is a fixed time horizon any measurable function $u:[0,T]\to\mathbb{R}^m$ is called admissible control, $x:[0,T]\to\mathbb{R}^n$ differentiable a.e. is the state function while the functions $f:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^n$, $g:\mathbb{R}^n\to\mathbb{R}$ and $h:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ are given.

To solve these problems one usually uses a set of necessary conditions known as the Pontryagin maximum principle.

The Pontryagin maximum principle

Similar to the KKT-condition the Pontryagin maximum principle (PMP) gives necessary conditions for a solution (x^*,u^*) of the problem (12)-(14). Using the Hamiltonian

$$H(t,x,p,u)=\langle p,f(t,x,u)\rangle+h(t,x,u) \tag{15}$$

the PMP says that for a given solution (x^*,u^*) of (12)-(14) there exists an absolutely continuous function p^*, called dual function, such that (x^*,p^*,u^*) solves the adjoint equation

$$\begin{array}{rcl}\dot{p}(t) &=& -H_x(t,x(t),p(t),u(t))=-f_x(t,x(t),u(t))^\top p(t)-h_x(t,x(t),u(t))^\top \text{ for a.e. } t\in[0,T]\\ p(T) &=& \nabla g(x(T)),\end{array} \tag{16}$$

and for every $u\in U$

$$\langle H_u(t,x^*(t),p^*(t),u^*(t)),u-u^*(t)\rangle\geq 0 \text{ for a.e. } t\in[0,T]. \tag{17}$$

Rewriting these necessary conditions in the form (3) one can now apply Newton's method (4) and gets an infinite dimensional analogue to the SQP algorithm. To guarantee convergence however some regularity conditions on $f+F$ have to be assumed and in particular an appropriate metric for the spaces that x, p and u lie in has to be found. E.g. in [12] it is proved that if one uses appropriate Sobolev spaces and assumes coercivity, a strong form of second-order sufficient condition, one gets strong metric regularity. Other necessary and/or sufficient condition for metric regularity in optimal control are very rare. An overview on known results and open problems can be found in [15].

Bang-bang type optimal control

In this thesis our focus lies on Bolza problems which are linear in control and usually do not satisfy the coercivity condition. The solutions of these problems are usually of bang-bang type i.e. there is a finite number of points where the optimal control switches from one extremal to another and is constant otherwise. This is due to the fact that the so called switching function

$$\sigma^*(t):=H_u(t,x^*(t),p^*(t),u^*(t))$$

does not depend on $u(t)$ directly and therefore is usually nonzero at all but a finite number of points.

The study of regularity properties of optimal control problems with bang-bang solutions has recently gained some popularity and some progress has been made (e.g. in [5], [6], [17], [28], [30]).

All of these papers use some version the following two assumptions. First some convex or convex-like assumption has to be made on the cost functional. Second a growth condition of the switching function around its zeros has to be assumed. In particular in [3] and [4] we use the following

Assumption. There exist real numbers $\theta, \alpha, \tau > 0$ such that for all $j \in \{1, \dots, m\}$ and $s \in [0, T]$ with $\sigma_j^*(s) = 0$ we have

$$|\sigma_j^*(t)| \geq \alpha |t - s|^\theta \quad \forall t \in [s - \tau, s + \tau] \cap [0, T]. \tag{18}$$

This assumption ensures that under certain perturbations (made precise in [3, Proposition 4.3]) the bang-bang property remains stable and so called singular arcs do not occur.

Gradient projection for bang-bang type optimal control

Another way to use iterative methods in optimal control is to not apply the method on the necessary conditions, but rather directly on the optimal control problem. To do so we view the optimal control problem as an infinite dimensional optimization problem in the control i.e. a problem of the form

$$\text{minimize } J(u) \tag{19}$$

subject to

$$u \in \mathcal{U}, \tag{20}$$

where $J(u) = \psi(x(u), u)$ is the cost function as a function of u, $x(u)$ is the unique trajectory given by the dynamics of the control and $\mathcal{U} = \{u \in L^1([0, T], \mathbb{R}^m) : u(t) \in [-1, 1]^m \text{ for a.e. } t \in [0, T]\}$.

For problems of this type one can use the gradient projection method which works as follows. For a given starting point $u_0 \in \mathcal{U}$ generate a sequence $\{u_k\}$ by iteratively computing

$$u_{k+1} = P_{\mathcal{U}}(u_k - \lambda_k DJ(u_k)), \tag{21}$$

where $P_{\mathcal{U}} : L^1([0, T], \mathbb{R}^m) \to \mathcal{U}$ is the operator projecting onto $\mathcal{U}$ and λ_k are predetermined positive parameters. For strongly convex objective functions J it is known that the iterative sequence $\{u_k\}$ converges linearly to the unique solution. More details about the classical gradient descent and gradient projection method can be found e.g. in [25].

In the bang-bang case however the cost functional J is usually not strongly convex. In [4] we address this problem.

Differential generalized equations

Another way to look at the necessary conditions in optimal control is to seperate the differential equations from the (set-vauled) algebraic conditions. This motivates the notion of differential generalized equations (DGE), i.e. a differential equation coupled with a generalized equation.

$$\begin{aligned} \dot{x}(t) &= g(x(t), u(t)) \\ 0 &\in f(x(t), x(0), x(T), u(t)) + F(u(t)) \end{aligned}$$

This notion has been introduced very recently in [9], wherein regularity properties of this notion are studied. It allows for a general comparison of pointwise versions of metric regularity of the set-valued algebraic part where the spaces involved are finite dimensional and its infinite-dimensional counterparts.

Summary of the results

In [1] we review the Kantorovich-type theorems discussed above and extend them for generalized equations. We obtain a result [1, Theorem 2.2] extending the theorem of Qi and Sun[27, Theorem

3.3] to generalized equations with nonsmooth single valued part f and linear mappings A_k replacing the derivatives $Df(x_k)$. We then use this general result to prove various other known Kantorovich-type theorems dealing with smoother cases including one which is very similar to Kantorovich's original statement but applied to generalized equations.[2] Additionally we include some elementary examples illustrating the difference between the Newton and the chord method regarding radius of convergence and convergence speed. Finally we apply the Newton and the chord method for some examples of generalized equations, namely for nonsmooth inequalities and for a model of economic equilibrium introduced in [14] given by a specific variational inequality.

Further we want to apply the Newton method for optimal control problems. More precisely we apply the Newton method onto the necessary optimality conditions where the Newton method becomes an infinite dimensional analogue to the SQP algorithm in nonlinear programming.

Specifically in [3] we restrict ourselves to optimal control problems that are linear in control i.e. we look at the following problem:

$$\begin{array}{ll} \text{minimize} & g(x(T)) + \int_0^T \left[w(x(t),t) + \langle s(x(t),t), u(t)\rangle\right] dt \\ \text{subject to} & \dot{x}(t) = a(x(t),t) + B(x(t),t)u(t), \quad t \in [0,T], \\ & u(t) \in U := [-1,1]^m, \\ & x(0) = x_0. \end{array} \tag{22}$$

Then a Newton method as described above reduces solving these problems to solving a series of optimal control problems which are linear in control and quadratic in the state variable, i.e. problems of the form

$$\begin{array}{ll} \text{minimize} & g(x(T)) + \int_0^T \left(\frac{1}{2}x(t)^\top W(t)x(t) + x(t)^\top S(t)u(t)\right) \\ \text{subject to} & \dot{x}(t) = A(t)x(t) + B(t)u(t) + d(t), \quad t \in [0,T], \\ & u(t) \in U := [-1,1]^m, \\ & x(0) = x_0, \end{array} \tag{23}$$

which in the following we call LQ-bang-bang-type problems. It is therefore natural to focus first on the regularity properties of these problems. Writing the associated generalized equation

$$0 \in F(z), \tag{24}$$

where $z = (x, p, u)$ and $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is the mapping associated with the PMP-system (see [3, page 2-3] for details) in appropriate spaces $\mathcal{X}$ and $\mathcal{Y}$ (see [3, page 5-6]) and assuming smoothness, a convexity-type assumption, and the growth condition (18) ((A1)-(A3) in [3]) one can show that a unique solution of bang-bang-type exists. Unfortunately at this solution strong regularity is usually not satisfied. But we could show ([3, Theorem 3.3]) a stronger version of strong metric subregularity of F at the solution $\hat{z} \in \mathcal{X}$, namely that for every $b > 0$ there exists $c > 0$ such that for any $y \in \mathcal{Y}$ with $\|y\| \leq b$ there exists $z \in \mathcal{X}$ such that $y \in F(z)$ and for any such z we have

$$\|z - \hat{z}\| \leq c\|y\|^{\frac{1}{\theta}}, \tag{25}$$

where $\theta \geq 1$ is the constant given in (18). Some slightly different version with stronger assumptions of this result was shown in [6].

Next we present a theorem ([3, Theorem 5.1]) which implies quadratic convergence of Newton's method assuming (25) for the linearized problem at the solution and existence of any Newton sequence. This theorem is quite similar to Theorem 7, where the main difference lies in the fact that we only need to assume that the starting point is close enough to the solution rather than a

[2] Note that there exists a version of this Kantorovich-type theorem in [11] but with slightly different assumptions.

whole tail of the sequence. In our situation a compactness argument shows that such a Newton sequence indeed exists. In summary we showed (see [3, Theorem 5.4]) that for the set-valued mapping F corresponding to the problem (22) the Newton method converges quadratically if there is a (then unique) solution $\hat{z}$ such that the linearization $LP(\hat{z})$ at $\hat{z}$ fulfills the assumptions for [3, Theorem 3.3].

Additionally in the paper [3] we extend the notion of strong bi-metric regularity introduced in [28] for Mayer problems to include LQ-bang-bang-type problems. This notion allows a more precise treatment of perturbation analysis for bang-bang-type problems by using two seperate metrics for measuring the perturbation of the point $(\bar{x}, \bar{y})$ and the Lipschitz continuity.

Definition 8. The map $\Phi : X \rightrightarrows Y$ is strongly bi-metrically regular (relative to $\widetilde{Y} \subset Y$) at $\bar{x} \in X$ for $\bar{y} \in \widetilde{Y}$ with constants $\varsigma \geq 0$, $a > 0$ and $b > 0$ if $(\bar{x}, \bar{y}) \in gph(\Phi)$ and the following properties are fulfilled:

1. the mapping $B_{\widetilde{Y}}(\bar{y}; b) \ni y \mapsto \Phi^{-1}(y) \cap B_X(\bar{x}; a)$ is single-valued, and
2. for all $y, y' \in B_{\widetilde{Y}}(\bar{y}; b)$,

$$d_X(\Phi^{-1}(y) \cap B_X(\bar{x}; a), \Phi^{-1}(y') \cap B_X(\bar{x}; a)) \leq \varsigma d_Y(y, y'). \tag{26}$$

Note that this notion generalizes strong metric regularity insofar as if we choose $\widetilde{Y} = Y$ and $d_{\widetilde{Y}} = d_Y$ then bi-metric regularity reduces to strong metric regularity. This notion is needed since in the L^∞ norm for two nonidentical bang-bang controls there is a positive lower bound for the distance between those two. I.e. a sequence of bang-bang controls which is not eventually constant is never convergent in L^∞. In contrast this is certainly possible for example in L^1. On the other hand the growth condition (18) is only stable in $W^{1,\infty}$, which we also proved ([3, Proposition 4.3]). As we show in [3, Theorem 4.5] if the condition (18) is fulfilled with $\theta = 1$ then under slightly stronger conditions ((A1')-(A2') in [3]) the mapping F is strongly bi-metrically regular with constant $\varsigma = 1$ for appropriately chosen spaces.

To solve optimal control problems directly without using the necessary optimality conditions in practice one often uses the gradient projection method (GPM). In [4] we investigate the GPM (21) for optimal control problems linear in control. As the cost function $J(u)$ in this case is usually not strongly convex the classical theory about the GPM fails. However, using assumptions ((A1)-(A5) in [4]) similar to those in [3] including convexity of J and a growth condition for J around the solution u^*, which is fulfilled if (18) is satisfied, we show sublinear convergence in [4, Thoerem 3.2]. More precisely we showed that for any chosen sequence $\{\lambda_k\}$ with

$$0 < \lambda_{\min} \leq \lambda_k \leq \frac{1}{L} \quad \forall k \in \mathbb{N}, \tag{27}$$

we have the following sublinear estimate for u_k

$$\|u_k - u^*\|^2 \leq \eta k^{-\frac{1}{\theta}} \quad \forall k \in \mathbb{N}, \tag{28}$$

where η is a constant. Additionally we show that the sequence $J(u_k)$ is monotone decreasing. Further we give a very simple example ([4, Example 3.4]) which shows that the estimation (28) is sharp and illustrate by two practical examples ([4, Example 4.1-2]) taken from other papers about bang-bang controls that the results are plausible.

In a very recent development [9] the authors introduced the notion of differential generalized equations. This notion covers a large territory of problems in control and optimization, such as control systems with constraints, necessary optimality conditions as well as differential variational

inequalities. Further [9] studies (strong) metric regularity and especially the interplay between the pointwise versions of these properties and their infinite-dimensional counterparts. In [2] we extend these ideas. In particular we formalize the concept of uniform (strong) regularity, which is used in [9] in a rather informal way. We prove that (strong) metric regularity at each point of a compact set implies uniform (strong) metric regularity i.e. that it is possible to choose a common regularity constant κ and neighborhood sizes for all these points. Further we extend the error estimates for the predictor-corrector path-following scheme treated in [9] to a path following scheme using a Heun-scheme-type predictor step. Additionally for constant set-valued parts we prove ([2, Theorem 4.3]) that that pointwise regularity at a solution of a DGE is equivalent to regularity in the infinite dimensional setting. In particular we use this along continuous paths which allows us to show error estimates for predictor-corrector path-following schemes.

Further research

As with every research, further open questions remain. Firstly, we did not deal with singular arc solutions in [3]. Some progress in this field has been made recently by Felgenhauer ([16],[18]). In addition we only considered optimal control problems in finite horizon. Regularity properties in infinite horizon optimal control are a widely open field where almost no research has been done so far. Further the study of regularity properties for DGEs is everything but finished.

Bibliography

[1] Cibulka R., Dontchev A.L., Preininger J., Roubal T., Veliov V.M.: Kantorovich-type Theorems for Generalized Equations. Journal Convex Analysis, 25(2), 459–486 (2018)

[2] Cibulka R., Preininger J., Roubal T.: On uniform regularity and strong regulartiy. Submitted in Journal of Optimization (2018)

[3] Preininger J., Scarinci T., Veliov V.M.: Metric regularity properties in bang-bang type linear quadratic optimal control problems. To appear in Journal of Set-Valued and Variational Analysis: Theory and Applications (2018)

[4] Preininger J., Vuong, P.T.: On the Convergence of the Gradient Projection Method of Optimal Control Problems with Bang-bang Solutions. Computational Optimization and Applications, 70(1), 221–238 (2018)

[5] Alt W., Baier R., Gerdts M., Lempio F.: Approximation of Linear Control Problems with Bang-Bang Solutions. Optimization, 62(1), 9–32 (2013)

[6] Alt W., Schneider C, Seydenschwanz M.: Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang-bang solutions. Appl. Math. and Comp., 287–288, 104–124 (2016)

[7] Aragón Artacho F.J., Belyakov A., Dontchev A.L., López M.: Local convergence of quasi-Newton methods under metric regularity. Comput. Optim. Appl., 58(1), 225–247 (2014)

[8] Bartle R.G.: Newton's method in Banach spaces. Proc. Amer. Math. Soc. 6 827–831 (1955)

[9] Cibulka R., Dontchev A.L., Krastanov M., Veliov V.M.: Metrically Regular Differential Generalized Equations. SIAM J. Control Optim., 56(1), 316–342 (2018)

[10] Cibulka R., Dontchev A.L., Kruger A.Y.: Strong metric subregularity of mappings in variational analysis and optimization. J. Math. Anal. Appl. 457, 1247–1282 (2017)

[11] Dontchev A.L.: Local analysis of a Newton-type method based on partial linearization. in: The Mathematics of Numerical Analysis, (Park City, 1995), Lectures Appl. Math. 32, Amer. Math. Soc., Providence, (1996) 295–306.

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

[13] Dontchev A.L., Rockafellar R.T.: Implicit Functions and Solution Mappings: A View from Variational Analysis. Second edition. Springer, New York (2014)

[14] Dontchev A.L., Rockafellar R.T.: Parametric stability of solutions in models of economic equilibrium. J. Convex Analysis 19(4), 975–997 (2012)

[15] Dontchev A.L., Veliov V.M.: Regularity Properties of Mappings in Optimal Control. Control Processes., 35–41, Proc. Internat. Conf. dedicated to the 90th anniversary of N.N. Krasovskii. Publisher: Institute of Mathematics and Mechanics, Ural Branch of the RAC (2015)

[16] Felgenhauer U.: A Newton-type method and optimality test for problems with bang-singular-bang optimal control. Pure Appl. Funct. Anal., 1 (2), 197–215 (2016)

[17] Felgenhauer U.:: On Stability of Bang-Bang Type Controls. SIAM J. Control Optim., 41(6), 1843–1867 (2003)

[18] Felgenhauer U.: Stability analysis of variational inequalities for bang-singular-bang controls. Control and Cybernetics, 42(3), 557–592 (2013)

[19] Izmailov A.F., Kurennoy A.S., Solodov M.V.: The Josephy-Newton Method for Semismooth Generalized Equations and Semismooth SQP for Optimization. Set-Valued Var. Anal, 21(1), 17–45 (2013)

[20] Izmailov A.F., Solodov M.V.: Inexact Josephy-Newton framework for generalized equations and its applications to local analysis of Newtonian methods for constrained optimization. Comp. Optim. Appl., 46(2), 347–368 (2010)

[21] Izmailov A.F., Solodov M.V.: Newton-type Methods for Optimization and Variational Problems. Springer, Berlin (2014)

[22] Kantorovich L.V.: On Newton's method for functional equations (Russian). Doklady Akad. Nauk SSSR (N.S.) 59, 1237–1240 (1948)

[23] Kantorovich L.V., Akilov G.P.: Functional Analysis (Russian), 2nd revised edition, Nauka, Moscow (1977)

[24] Kelley C.T.: Iterative Methods for Linear and Nonlinear Equations. Frontiers in Applied Mathematics, SIAM, Philadelphia (1995)

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

[26] Potra F.A., Pták V.: Nondiscrete induction and iterative processes. Research Notes in Mathematics 103, Pitman, Boston (1984)

[27] Qi L., Sun J.: A nonsmooth version of Newton's method. Math. Programming, 58(1-3), 353–367 (1993)

[28] Quincampoix M., Veliov V.M.: Metric Regularity and Stability of Optimal Control Problems for Linear Systems. SIAM J. Control Optim., 51(5), 4118–4137 (2013)

[29] Robinson S.M.: Strongly regular generalized equations. Math. Oper. Res. 5, 43—62 (1980)

[30] Seydenschwanz M.: Convergence results for the discrete regularization of linear-quadratic control problems with bang-bang solutions. Comput. Optim. Appl., 61(3), 731–760 (2015)

Journal of Convex Analysis
Volume 25 (2018), No. 2, 1–xx

Kantorovich-Type Theorems for Generalized Equations

Radek Cibulka*

NTIS – New Technologies for the Information Society and Dept. of Mathematics, Faculty of Applied Sciences, Univ. of West Bohemia, Univerzitní 22, 306 14 Pilsen, Czech Republic
cibi@kma.zcu.cz

Asen L. Dontchev*,†

Mathematical Reviews, 416 Fourth Street, Ann Arbor, MI 48107-8604, U.S.A.
and: Institute of Statistics and Mathematical Methods in Economics,
Vienna University of Technology, Wiedner Hauptstrasse 8, 1040 Vienna, Austria
ald@ams.org

Jakob Preininger†, Vladimir Veliov†

Institute of Statistics and Mathematical Methods in Economics,
Vienna University of Technology, Wiedner Hauptstrasse 8, 1040 Vienna, Austria
jakob.preininger@tuwien.ac.at, veliov@tuwien.ac.at

T. Roubal*

NTIS – New Technologies for the Information Society and Dept. of Mathematics, Faculty of Applied Sciences, Univ. of West Bohemia, Univerzitní 22, 306 14 Pilsen, Czech Republic
roubalt@students.zcu.cz

Dedicated to Antonino Maugeri on the occasion of his 70th birthday.

Received: November 20, 2015
Accepted: May 31, 2016

We study convergence of the Newton method for solving generalized equations of the form $f(x) + F(x) \ni 0$, where f is a continuous but not necessarily smooth function and F is a set-valued mapping with closed graph, both acting in Banach spaces. We present a Kantorovich-type theorem concerning r-linear convergence for a general algorithmic strategy covering both nonsmooth and smooth cases. Under various conditions we obtain higher-order convergence. Examples and computational experiments illustrate the theoretical results.

Keywords: Newton's method, generalized equation, variational inequality, metric regularity, Kantorovich theorem, linear/superlinear/quadratic convergence.

2010 Mathematics Subject Classification: 49J53, 49J40, 65J15, 90C30.

*Supported by the project GA15-00735S.
†Supported by Austrian Science Foundation (FWF) Grant P26640-N25.

ISSN 0944-6532 / $ 2.50

1. Introduction

While there is some disagreement among historians who actually invented the Newton method, see [34] for an excellent reading about early history of the method, it is well documented in the literature that L. V. Kantorovich [22] was the first to obtain convergence of the method on assumptions involving the point where iterations begin. Specifically, Kantorovich considered the Newton method for solving the equation $f(x) = 0$ and proved convergence by imposing conditions on the derivative $Df(x_0)$ of the function f and the residual $\|f(x_0)\|$ at the starting point x_0. These conditions can be actually checked, in contrast to the conventional approach utilizing the assumption that the derivative $Df(\bar{x})$ at a (unknown) root $\bar{x}$ of the equation is invertible and then claim that if the iteration starts close enough to $\bar{x}$ then it generates a convergent to $\bar{x}$ sequence. For this reason Kantorovich's theorem is usually called a global convergence theorem[1] whereas conventional convergence theorems are regarded as local theorems.

The following version of Kantorovich's theorem is close to that in [27]; for a proof see [27] or [23].

Theorem 1.1 (Kantorovich). *Let X and Y be Banach spaces. Consider a function $f\colon X \to Y$, a point $x_0 \in X$ and a real $a > 0$, and suppose that f is continuously Fréchet differentiable in an open neighborhood of the ball $\mathbb{B}_a(x_0)$ and its Fréchet derivative Df is Lipschitz continuous in $\mathbb{B}_a(x_0)$ with a constant $L > 0$. Assume that there exist positive reals κ and η such that*

$$\|Df(x_0)^{-1}\| \le \kappa \quad \text{and} \quad \|Df(x_0)^{-1} f(x_0)\| < \eta.$$

If $\alpha := \kappa L \eta a < \frac{1}{2}$ and $a \ge a_0 := \frac{1-\sqrt{1-2\alpha}}{\kappa L}$, then there exists a unique sequence $\{x_k\}$ satisfying the iteration

$$f(x_k) + Df(x_k)(x_{k+1} - x_k) = 0, \quad k = 0, 1, \ldots, \tag{1}$$

with a starting point x_0; this sequence converges to a unique zero $\bar{x}$ of f in $\mathbb{B}_{a_0}(x_0)$ and the convergence rate is r-quadratic; specifically

$$\|x_k - \bar{x}\| \le \frac{\eta}{\alpha}(2\alpha)^{2^k}, \quad k = 0, 1, \ldots.$$

In his proof of convergence Kantorovich used a novel technique of *majorization* of the sequence of iterate increments by the increments of a sequence of scalars. Notice that the derivative Df is injective not only at x_0 but also at the solution $\bar{x}$; indeed, for any $y \in X$ with $\|y\| = 1$ we have

$$\|Df(\bar{x})y\| \ge \|Df(x_0)y\| - \|(Df(\bar{x}) - Df(x_0))y\| \ge \frac{1}{\kappa} - La_0 = \frac{\sqrt{1-2\alpha}}{\kappa} > 0.$$

[1] Some authors prefer to call such a result a semilocal convergence theorem.

In a related development, Kantorovich showed in [23, Chapter 18] that, under the same assumptions as in Theorem 1.1, to achieve linear convergence to a solution there is no need to calculate during iterations the derivative $Df(x_k)$ at the current point x_k— it is enough to use at each iteration the value of the derivative $Df(x_0)$ at the starting point, i.e., the iteration (1) becomes

$$f(x_k) + Df(x_0)(x_{k+1} - x_k) = 0, \quad k = 0, 1, \dots . \tag{2}$$

He called this method the *modified Newton process*. This method is also known as the *chord method*, see [24, Chapter 5].

The work of Kantorovich has been extended in a number of ways by, in particular, utilizing various extensions of the majorization technique, such as the method of nondiscrete induction, see e.g. [29]. We will not go into discussing these works here but rather focus on a version of Kantorovich's theorem due to R. G. Bartle [6], which has been largely forgotten if not ignored in the literature. A version of Bartle's theorem, without referring to [6], was given recently in [9, Theorem 5].

Specifically, Bartle [6] considered the equation $f(x) = 0$, for a function f acting between Banach spaces X and Y, which is solved by the iteration

$$f(x_k) + Df(z_k)(x_{k+1} - x_k) = 0, \quad k = 0, 1, \dots , \tag{3}$$

where z_k are, to quote [6], "arbitrarily selected points ... sufficiently close to the solution desired." For $z_k = x_k$ one obtains the usual Newton method, and for $z_k = x_0$ the modified Newton/chord method, but z_k may be chosen in other ways. For example as x_0 for the first s iterations and then the derivative could be calculated again every s iterations, obtaining in this way a *hybrid* version of the method. If computing the derivatives, in particular in the case they are obtained numerically, involves time consuming procedures, it is quite plausible to expect that for large scale problems the chord method or a hybrid version of it would possibly be faster than the usual method. We present here the following somewhat modified statement of Bartle's theorem which fits our purposes:

Theorem 1.2 (Bartle [6]). *Assume that the function $f\colon X \to Y$ is continuously Fréchet differentiable in an open set O. Let $x_0 \in O$ and let there exist positive reals a and κ such that for any three points $x_1, x_2, x_3 \in \mathbb{B}_a(x_0) \subset O$ we have*

$$\|Df(x_1)^{-1}\| < \kappa \quad \text{and} \quad \|f(x_1) - f(x_2) - Df(x_3)(x_1 - x_2)\| \le \frac{1}{2\kappa}\|x_1 - x_2\|, \tag{4}$$

and also

$$\|f(x_0)\| < \frac{a}{2\kappa}. \tag{5}$$

Then for every sequence $\{z_k\}$ with $z_k \in \mathbb{B}_a(x_0)$ there exists a unique sequence $\{x_k\}$ satisfying the iteration (3) *with initial point x_0; this sequence converges*

to a root $\bar{x}$ of f which is unique in $\mathbb{B}_a(x_0)$ and the convergence rate is r-linear; specifically

$$\|x_k - \bar{x}\| \le 2^{-k}a, \quad k = 0, 1, \dots.$$

In a path-breaking paper Qi and Sun [30] extended the Newton method to a nonsmooth equation by employing Clarke's generalized Jacobian $\bar{\partial}f$ of a function $f : \mathbb{R}^n \to \mathbb{R}^n$ instead of the derivative Df and proved convergence for a class of nonsmooth functions. Specifically, consider the following iteration: given x_k choose any matrix A_k from $\bar{\partial}f(x_k)$ and then find the next iterate by solving the linear equation

$$f(x_k) + A_k(x_{k+1} - x_k) = 0, \quad k = 0, 1, \dots. \tag{6}$$

The following convergence theorem was proved in [30, Theorem 3.2]:

Theorem 1.3. *Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous around a root $\bar{x}$ at which all matrices in $\bar{\partial}f(\bar{x})$ are nonsingular. Also assume that for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $x \in \mathbb{B}_\delta(\bar{x})$ and for every $A \in \bar{\partial}f(x)$ one has*

$$\|f(x) - f(\bar{x}) - A(x - \bar{x})\| \le \varepsilon\|x - \bar{x}\|. \tag{7}$$

Then there exists a neighborhood U of $\bar{x}$ such that for every starting point $x_0 \in U$ there exists a sequence satisfying the iteration (6) *and every such sequence is superlinearly convergent to $\bar{x}$.*

A function f which is Lipschitz continuous around a point $\bar{x}$ and satisfies (7) is said to be *semismooth*[2] at $\bar{x}$. Accordingly, the method (6) is a *semismooth Newton method* for solving equations. For more advanced versions of Theorem 1.3, see e.g. [15, Theorem 7.5.3], [21, Theorem 2.42] and [14, Theorem 6F.1].

In the same paper Qi and Sun proved what they called a "global" theorem [30, Theorem 3.3], which is more in the spirit of Kantorovich's theorem; we will state and prove an improved version of this theorem in the next section.

In this paper we derive Kantorovich-type theorems for a generalized equation: find a point $x \in X$ such that

$$f(x) + F(x) \ni 0, \tag{8}$$

where throughout $f\colon X \to Y$ is a continuous function and $F\colon X \rightrightarrows Y$ is a set-valued mapping with closed graph. Many problems can be formulated as (8), for example, equations, variational inequalities, constraint systems, as well as optimality conditions in mathematical programming and optimal control.

[2]Sometimes one adds to (7) the condition that f is directionally differentiable in every direction.

Newton-type methods for solving nonsmooth equations and variational inequalities have been studied since the 70s. In the last two decades a number of new developments have appeared some of which have been collected in several books [15, 18, 19, 25, 33]. A broad presentation of convergence results for both smooth and nonsmooth problem with particular emphasis on applying Newton-type method to optimization can be found in the recent book [21]. A Kantorovich-type theorem for generalized equations under metric regularity is proven in [13, Theorem 2] using the majorization technique, see also the recent papers [2] and [32]. Related results for particular nonsmooth generalized equations are given in [16] and [28]. In [8] applications of the modified Newton method for solving optimization problems appearing in nonlinear model predictive control are reported.

We adopt the notations used in the book [14]. The set of all natural numbers is denoted by $\mathbb{N}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$; the n-dimensional Euclidean space is $\mathbb{R}^n$. Throughout X and Y are Banach spaces both norms of which are denoted by $\|\cdot\|$. The closed ball centered at x with radius r is denoted as $\mathbb{B}_r(x)$; the unit ball is $\mathbb{B}$. The distance from a point x to a set A is $\operatorname{dist}(x, A) = \inf_{y\in A} \|x - y\|$. A generally set-valued mapping $F\colon X \rightrightarrows Y$ is associated with its graph $\operatorname{gph} F = \big\{ (x, y) \in X \times Y \,\big|\, y \in F(x)\big\}$ and its domain $\operatorname{dom} F = \big\{ x \in X \,\big|\, F(x) \neq \emptyset\big\}$. The inverse of F is $y \mapsto F^{-1}(y) = \big\{ x \in X \,\big|\, y \in F(x)\big\}$. By $\mathcal{L}(X, Y)$ we denote a space of linear bounded operators acting from X into Y equipped with the standard operator norm.

Recall that a set-valued mapping $\Phi\colon X \rightrightarrows Y$ is said to be *metrically regular* at x_0 for y_0 if $y_0 \in \Phi(x_0)$ and there exist neighborhoods U of x_0 and V of y_0 and a positive constant κ such that the set $\operatorname{gph}\Phi \cap (U \times V)$ is closed and

$$\operatorname{dist}\big(x, \Phi^{-1}(y)\big) \leq \kappa \operatorname{dist}\big(y, \Phi(x)\big) \quad \text{for all} \quad (x, y) \in U \times V. \tag{9}$$

The infimum over all $\kappa \geq 0$ in (9) is the regularity modulus of Φ at x_0 for y_0 denoted by $\operatorname{reg}(\Phi; x_0 | y_0)$. If in addition the mapping $\sigma\colon\ V \ni y \mapsto \Phi^{-1}(y) \cap U$ is not multivalued on V, then Φ is said to be strongly metrically regular and then σ is a Lipschitz continuous function on V. More about metric regularity and the related theory can be found in [14].

2. Main theorem

In preparation to our main result presented in Theorem 2.2 we give a strengthened version of [30, Theorem 3.3] for the iteration (6) applied to an equation in Banach spaces.

Theorem 2.1. *Let $f\colon X \to Y$ be a continuous function and let the numbers $a > 0$, $\kappa \geq 0$, $\delta \geq 0$ be such that*

$$\kappa\delta < 1 \quad \text{and} \quad \|f(x_0)\| < (1 - \kappa\delta)\frac{a}{\kappa}. \tag{10}$$

Consider the iteration (6) *with a starting point* x_0 *and a sequence* $\{A_k\}$ *of linear and bounded mappings such that for every* $k \in \mathbb{N}_0$ *we have*

$$\begin{cases} \|A_k^{-1}\| \le \kappa \quad \textit{and} \\ \|f(x) - f(x') - A_k(x - x')\| \le \delta\|x - x'\| \quad \textit{for every} \quad x, x' \in \mathbb{B}_a(x_0). \end{cases} \tag{11}$$

Then there exists a unique sequence satisfying the iteration (6) *with initial point* x_0. *This sequence remains in* $\operatorname{int}\mathbb{B}_a(x_0)$ *and converges to a root* $\bar{x} \in \operatorname{int}\mathbb{B}_a(x_0)$ *of* f *which is unique in* $\mathbb{B}_a(x_0)$; *moreover, the convergence rate is r-linear:*

$$\|x_k - \bar{x}\| < (\kappa\delta)^k a.$$

Proof. Let $\alpha := \kappa\delta$. We will show, by induction, that there is a sequence $\{x_k\}$ with elements in $\operatorname{int}\mathbb{B}_a(x_0)$ satisfying (6) with the starting point x_0 such that

$$\|x_{j+1} - x_j\| \le \alpha^j \kappa \|f(x_0)\| < a\alpha^j(1 - \alpha), \quad j = 0, 1, \ldots. \tag{12}$$

Let $k := 0$. Since A_0 is invertible, there is a unique $x_1 \in X$ such that we get $A_0(x_1 - x_0) = -f(x_0)$. Therefore,

$$\|x_1 - x_0\| = \|A_0^{-1}A_0(x_1 - x_0)\| = \|A_0^{-1}f(x_0)\| \le \kappa\|f(x_0)\| < a(1 - \alpha).$$

Hence $x_1 \in \operatorname{int}\mathbb{B}_a(x_0)$. Suppose that, for some $k \in \mathbb{N}$, we have already found points $x_0, x_1, \ldots, x_k \in \mathbb{B}_a(x_0)$ satisfying (12) for each $j = 0, 1, \ldots, k-1$. Since A_k is invertible, there is a unique $x_{k+1} \in X$ such that $A_k(x_{k+1} - x_k) = -f(x_k)$. Then (12) with $j := k - 1$ implies

$$\begin{aligned} \|x_{k+1} - x_k\| &= \|A_k^{-1}A_k(x_{k+1} - x_k)\| = \|A_k^{-1}f(x_k)\| \le \kappa\|f(x_k)\| \\ &= \kappa\|f(x_k) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1})\| \\ &\le \kappa\delta\|x_k - x_{k-1}\| \le \alpha^k\kappa\|f(x_0)\| < a\alpha^k(1 - \alpha). \end{aligned}$$

From (12), we have

$$\|x_{k+1} - x_0\| \le \sum_{j=0}^{k} \|x_{j+1} - x_j\| \le \sum_{j=0}^{k} \alpha^j\kappa\|f(x_0)\| < a\sum_{j=0}^{\infty} \alpha^j(1 - \alpha) = a, \tag{13}$$

that is, $x_{k+1} \in \operatorname{int}\mathbb{B}_a(x_0)$. The induction step is complete.

For any natural k and p we have

$$\begin{aligned} \|x_{k+p+1} - x_k\| &\le \sum_{j=k}^{k+p} \|x_{j+1} - x_j\| \le \sum_{j=k}^{k+p} \alpha^j\kappa\|f(x_0)\| \\ &< \frac{\alpha^k}{1 - \alpha}\kappa\|f(x_0)\| < a\alpha^k. \end{aligned} \tag{14}$$

Hence $\{x_k\}$ is a Cauchy sequence; let it converge to $\bar{x} \in X$. Passing to the limit with $p \to \infty$ in (14) we obtain

$$\|\bar{x} - x_k\| \le \frac{\alpha^k}{1 - \alpha}\kappa\|f(x_0)\| < a\alpha^k \quad \text{for each} \quad k \in \mathbb{N}_0.$$

In particular, $\bar{x} \in \operatorname{int} I\!B_a(x_0)$. Using (6) and (11), we get

$$\begin{aligned} 0 \le \|f(\bar{x})\| &= \lim_{k\to\infty} \|f(x_k)\| = \lim_{k\to\infty} \|f(x_k) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1})\| \\ &\le \lim_{k\to\infty} \delta\|x_k - x_{k-1}\| = 0. \end{aligned}$$

Hence, $f(\bar{x}) = 0$. Suppose that there is $\bar{y} \in I\!B_a(x_0)$ with $\bar{y} \neq \bar{x}$ and $f(\bar{y}) = 0$. Then

$$\begin{aligned} \|\bar{y} - \bar{x}\| &\le \kappa\|A_0(\bar{y} - \bar{x})\| = \kappa\|f(\bar{y}) - f(\bar{x}) - A_0(\bar{y} - \bar{x})\| \\ &\le \kappa\delta\|\bar{y} - \bar{x}\| < \|\bar{y} - \bar{x}\|, \end{aligned}$$

which is a contradiction. Hence $\bar{x}$ is a unique root of f in $I\!B_a(x_0)$. □

Our main result which follows is an extension of Theorem 2.1 for generalized equations (8). We adopt the following model of an iterative procedure for solving (8). Given $k \in I\!N_0$, based on the current and prior iterates x_n $(n \le k)$ one generates a "feasible" element $A_k \in \mathcal{L}(X, Y)$ and then the next iterate x_{k+1} is chosen according to the following Newton-type iteration:

$$f(x_k) + A_k(x_{k+1} - x_k) + F(x_{k+1}) \ni 0. \tag{15}$$

In order to formalize the choice of A_k we consider a sequence of mappings $A_k : X^k \to \mathcal{L}(X, Y)$, where $X^k = X \times \ldots \times X$ is the product of k copies of X. Thus, A_k does not need to be chosen in advance and may depend on the already obtained iterates. In particular, one may take $A_k = A_0(x_0)$, that is, use the same operator for all iterations, as in the standard chord method. Another possibility is to use $A_k = Df(x_k)$ in the case of a differentiable f or $A_k \in \bar{\partial} f(x_k)$, the Clarke generalized Jacobian if applicable. Intermediate choices are also possible, for example to use the same operator A in m successive steps and then to update it at the current point: $A_k(x_0, \ldots, x_k) = A_{m[k/m]}(x_{m[k/m]})$, where $[s]$ is the integer part of s.

Theorem 2.2. *Let the scalars $a > 0$, $b > 0$, $\kappa > 0$, $\delta \ge 0$ and the points $x_0 \in X$, $y_0 \in f(x_0) + F(x_0)$ be such that*

(A1) *$\kappa\delta < 1$ and $\|y_0\| < (1 - \kappa\delta)\min\{\frac{a}{\kappa}, b\}$.*

Moreover, assume there exists a function $\omega\colon [0, a] \to [0, \delta]$ such that for every $k \in I\!N_0$ and every $x_1, \ldots, x_k \in I\!B_a(x_0)$ the linear and bounded operator $A_k := A_k(x_0, \ldots, x_k)$ appearing in the iteration (15) has the following properties:

(A2) *the mapping*

$$x \mapsto G_{A_k}(x) := f(x_0) + A_k(x - x_0) + F(x) \tag{16}$$

is metrically regular at x_0 for y_0 with constant κ and neighborhoods $I\!B_a(x_0)$ and $I\!B_b(y_0)$;

(A3) $\|f(x)-f(x_k)-A_k(x-x_k)\| \le \omega(\|x-x_k\|)\,\|x-x_k\|$ *for every* $x \in I\!B_a(x_0)$.

Then for every $\alpha \in (\kappa\delta, 1)$ *there exists a sequence* $\{x_k\}$ *generated by the iteration* (15) *with starting point* x_0 *which remains in* $\operatorname{int} I\!B_a(x_0)$ *and converges to a solution* $\bar{x} \in \operatorname{int} I\!B_a(x_0)$ *of* (8)*; moreover, the convergence rate is r-linear; specifically*

$$\|x_k-\bar{x}\| < \alpha^k a \text{ and } \operatorname{dist}(0, f(x_k)+F(x_k)) \le \alpha^k\|y_0\| \text{ for every } k \in I\!N_0. \quad (17)$$

If $\lim_{\xi\to 0}\omega(\xi) = 0$*, then the sequence* $\{x_k\}$ *is convergent r-superlinearly, that is, there exist sequences of positive numbers* $\{\varepsilon_k\}$ *and* $\{\eta_k\}$ *such that* $\|x_k-\bar{x}\| \le \varepsilon_k$ *and* $\varepsilon_{k+1} \le \eta_k\varepsilon_k$ *for all sufficiently large* $k \in I\!N$ *and* $\eta_k \to 0$.

If there exists a constant $L > 0$ *such that* $\omega(\xi) \le \min\{\delta, L\xi\}$ *for each* $\xi \in [0, a]$*, then the convergence of* $\{x_k\}$ *is r-quadratic: specifically, there exists a sequence of positive numbers* $\{\varepsilon_k\}$ *such that for any* $C > \frac{\alpha L}{\delta}$ *we have* $\varepsilon_{k+1} < C\varepsilon_k^2$ *for all sufficiently large* $k \in I\!N$.

If the mapping G_{A_k} *defined in* (16) *is not only metrically regular but also strongly metrically regular with the same constant and neighborhoods, then there is no other sequence* $\{x_k\}$ *satisfying the iteration (15) starting from* x_0 *which stays in* $I\!B_a(x_0)$.

Proof. Choose an $\alpha \in (\kappa\delta, 1)$ and then κ' such that

$$\frac{\alpha}{\delta} \ge \kappa' > \kappa \quad \text{and} \quad \|y_0\| < (1-\alpha)\min\left\{\frac{a}{\kappa'}, b\right\}. \quad (18)$$

Such a choice of κ' is possible for $\alpha > \kappa\delta$ sufficiently close to $\kappa\delta$. We shall prove the claim for an arbitrary value of α for which (18) holds with an appropriately chosen $\kappa' > \kappa$. This is not a restriction, since then (17) will hold for any larger value of α.

We will show that there exists a sequence $\{x_k\}$ with the following properties, for each $k \in I\!N$:

(a) $\|x_k - x_0\| \le \frac{1-\alpha^k}{1-\alpha}\kappa'\|y_0\| < (1-\alpha^k)a$;

(b) $\|x_k - x_{k-1}\| \le \alpha^{k-1}\gamma_0 \dots \gamma_{k-1}\kappa'\|y_0\| < \alpha^{k-1}(1-\alpha)a$,
where $\gamma_0 := 1$, $\gamma_i := \omega(\|x_i - x_{i-1}\|)/\delta$ for $i = 1, \dots, k-1$;

(c) $0 \in f(x_{k-1})+A_{k-1}(x_k-x_{k-1})+F(x_k)$, where $A_{k-1} := A_{k-1}(x_0, \dots, x_{k-1})$.

We use induction, starting with $k = 1$. Since $0 \in I\!B_b(y_0)$ and $y_0 \in G_{A_0}(x_0)$, using (A2) for G_{A_0} we have that

$$\operatorname{dist}\left(x_0, G_{A_0}^{-1}(0)\right) \le \kappa \operatorname{dist}\left(0, G_{A_0}(x_0)\right) \le \kappa\|y_0\|.$$

If $y_0 = 0$, then we take $x_1 = x_0$. If not, we have that

$$\operatorname{dist}\left(x_0, G_{A_0}^{-1}(0)\right) < \kappa'\|y_0\|$$

and then there exists a point $x_1 \in G_{A_0}^{-1}(0)$ such that

$$\|x_1 - x_0\| < \kappa' \|y_0\| < (1-\alpha)a.$$

Clearly, (a)–(c) are satisfied for $k := 1$ and γ_1 is well-defined.

Assume that for some $k \in \mathbb{N}$ the point x_k has already been defined in such a way that conditions (a)–(c) hold. We shall define x_{k+1} so that (a)–(c) remain satisfied for k replaced with $k+1$.

First, observe that (a) implies $x_k \in \mathbb{B}_a(x_0)$. Denote $r_k := f(x_0) - f(x_k) - A_k(x_0 - x_k)$. In view of (a), the fact that $\omega(\|x_0 - x_k\|) \le \delta$ and (A3) with $x = x_0$, we have

$$\begin{aligned}
\|r_k - y_0\| &\le \|y_0\| + \|f(x_0) - f(x_k) - A_k(x_0 - x_k)\| \\
&\le \|y_0\| + \delta\|x_0 - x_k\| \le \|y_0\| + \frac{1-\alpha^k}{1-\alpha}\kappa'\delta\|y_0\| \\
&\le \|y_0\| + \frac{1-\alpha^k}{1-\alpha}\alpha\|y_0\| = \frac{1-\alpha^{k+1}}{1-\alpha}\|y_0\| < b.
\end{aligned}$$

If $r_k \in G_{A_k}(x_k)$ then we take $x_{k+1} = x_k$. If not, by (A2),

$$\operatorname{dist}\left(x_k, G_{A_k}^{-1}(r_k)\right) \le \kappa \operatorname{dist}\left(r_k, G_{A_k}(x_k)\right) < \kappa' \operatorname{dist}\left(r_k, G_{A_k}(x_k)\right).$$

Then there exists a point $x_{k+1} \in G_{A_k}^{-1}(r_k)$ such that

$$\|x_{k+1} - x_k\| < \kappa' \operatorname{dist}\left(r_k, G_{A_k}(x_k)\right).$$

Due to (c), we get $G_{A_k}(x_k) = f(x_0) + A_k(x_k - x_0) + F(x_k) \ni f(x_0) + A_k(x_k - x_0) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1})$.

Using (A3) with $x = x_k$ and then (b) and (18) we have

$$\begin{aligned}
\|x_{k+1} - x_k\| &\le \kappa'\|r_k - [f(x_0) - f(x_{k-1}) + A_k(x_k - x_0) - A_{k-1}(x_k - x_{k-1})]\| \\
&= \kappa'\|f(x_k) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1})\| \\
&\le \kappa'\omega(\|x_k - x_{k-1}\|)\|x_k - x_{k-1}\| = \kappa'\delta\gamma_k\|x_k - x_{k-1}\| \qquad (19) \\
&\le \alpha^k\gamma_0 \dots \gamma_k\kappa'\|y_0\| < \alpha^k(1-\alpha)a. \qquad (20)
\end{aligned}$$

Hence, condition (b) is satisfied for $k+1$ and γ_{k+1} is well-defined. By the choice of x_{k+1} we have

$$r_k \in G_{A_k}(x_{k+1}) = f(x_0) + A_k(x_{k+1} - x_0) + F(x_{k+1}),$$

hence, after rearranging, condition (c) holds for $k+1$. To finish the induction step, use (a) to obtain

$$\begin{aligned}
\|x_{k+1} - x_0\| &\le \|x_{k+1} - x_k\| + \|x_k - x_0\| \le \alpha^k\kappa'\|y_0\| + \frac{1-\alpha^k}{1-\alpha}\kappa'\|y_0\| \\
&= \frac{1-\alpha^{k+1}}{1-\alpha}\,\kappa'\,\|y_0\|.
\end{aligned}$$

Now we shall prove that the sequence $\{x_k\}$ identified in the preceding lines is convergent. By (b) (with γ_i replaced with 1), applied for $k := m$, $n \in \mathbb{N}$ with $m < n$, we have

$$\|x_n - x_m\| \le \alpha^m \frac{1-\alpha^{n-m}}{1-\alpha}\kappa'\|y_0\|,$$

hence $\{x_k\}$ is a Cauchy sequence. Let $\bar{x} = \lim_{k\to\infty} x_k$. Then by (a),

$$\|\bar{x} - x_0\| \le \frac{\kappa'}{1-\alpha}\|y_0\| < a,$$

that is, $\bar{x} \in \operatorname{int} I\!B_a(x_0)$. Using (b), for any $k \in I\!N_0$, and the second inequality in (18), we have

$$\begin{aligned}\|x_k - \bar{x}\| &= \lim_{m\to\infty}\|x_k - x_{k+m}\| \le \lim_{m\to\infty}\sum_{i=k}^{k-1+m}\|x_i - x_{i+1}\| \\ &\le \lim_{m\to\infty}\sum_{i=k}^{k-1+m}\alpha^i\gamma_1\dots\gamma_i\kappa'\|y_0\| \le \alpha^k\gamma_1\dots\gamma_k\lim_{m\to\infty}\sum_{i=k}^{k-1+m}\alpha^{i-k}\kappa'\|y_0\| \\ &\le \alpha^k\gamma_1\dots\gamma_k\frac{\kappa'\|y_0\|}{1-\alpha} \le \alpha^k\gamma_1\dots\gamma_k a =: \varepsilon_k. \end{aligned} \tag{21}$$

By the definition of ε_k we get

$$\varepsilon_{k+1} = \alpha\gamma_{k+1}\varepsilon_k.$$

Since $\gamma_{k+1} \le 1$ we obtain linear convergence in (17). If $\lim_{\xi\to 0}\omega(\xi) = 0$, then $\gamma_k \to 0$ and we have r-superlinear convergence.

Finally, if there exists a constant L such that $\omega(\xi) \le \min\{\delta, L\xi\}$ for each $\xi \in [0, a]$, then for each $k \in I\!N$ condition (b) implies that $\xi := \|x_{k+1} - x_k\| < a$; hence

$$\gamma_{k+1} \le \min\{1, L\|x_{k+1} - x_k\|/\delta\} \le \|x_{k+1} - x_k\|L/\delta \le (\varepsilon_{k+1} + \varepsilon_k)L/\delta.$$

Fix any $C > \alpha L/\delta$. Since the sequence $\{\varepsilon_k\}$ is strictly decreasing and converges to zero, we obtain

$$\varepsilon_{k+1} \le \frac{\alpha L}{\delta}(\varepsilon_k + \varepsilon_{k+1})\varepsilon_k < C\varepsilon_k^2 \quad \text{for all sufficiently large } k \in I\!N.$$

This implies r-quadratic convergence.

To show that $\bar{x}$ solves (8), let $y_k := f(x_k) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1})$ for $k \in I\!N$. From (c) we have $y_k \in f(x_k) + F(x_k)$. Using (A3) with $x = x_k$ and then using (b) we obtain that

$$\begin{aligned}\|y_k\| &= \|f(x_k) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1})\| \\ &\le \delta\|x_k - x_{k-1}\| \le \delta\alpha^{k-1}\kappa'\|y_0\| \le \alpha^k\|y_0\|. \end{aligned} \tag{22}$$

Thus $(x_k, y_k) \to (\bar{x}, 0)$ as $k \to \infty$. Since f is continuous and F has closed graph, we obtain $0 \in f(\bar{x}) + F(\bar{x})$. The second inequality in (17) follows from (22).

In the case of strong metric regularity of G_A the way x_{k+1} is constructed from x_k implies automatically that x_{k+1} is unique in $\mathbb{B}_a(x_0)$. □

Remark 2.3. Suppose that there exist $\beta \in (0, 1]$ and $L > 0$ such that $\omega(\xi) \leq \min\{L\xi^\beta, \delta\}$ for each $\xi \in [0, a]$. Then $\{x_k\}$ converges to $\bar{x}$ with r-rate $1 + \beta$: there exists a sequence of positive numbers $\{\varepsilon_k\}$ converging to zero and $C > 0$ such that $\varepsilon_{k+1} \leq C\varepsilon_k^{1+\beta}$ for all $k \in \mathbb{N}$. Indeed, for each $k \in \mathbb{N}$, (b) implies that $\xi := \|x_{k+1} - x_k\| < a$, hence

$$\gamma_{k+1} \leq \frac{L}{\delta}\|x_{k+1} - x_k\|^\beta \leq \frac{L}{\delta}(\varepsilon_{k+1} + \varepsilon_k)^\beta = \frac{L}{\delta}(1 + \alpha\gamma_{k+1})^\beta\varepsilon_k^\beta \leq \frac{L}{\delta}(1 + \alpha)^\beta\varepsilon_k^\beta.$$

Hence, taking $C := \alpha L(1 + \alpha)^\beta/\delta$ we get

$$\varepsilon_{k+1} = \alpha\gamma_{k+1}\varepsilon_k \leq C\varepsilon_k^{1+\beta} \quad \text{for all } k \in \mathbb{N}.$$

Remark 2.4. Theorem 2.1 follows from the strong regularity part of Theorem 2.2. Indeed, for the case of the equation condition (A1) is the same as (10). The first inequality in (11) means that the mapping G_{A_k} with $F \equiv 0$ is strongly metrically regular uniformly in k, and the second inequality is the same as (A3).

The following corollary is a somewhat simplified version of Theorem 2.2 which may be more transparent for particular cases.

Corollary 2.5. *Let a, b, κ, δ be positive reals and a point $(x_0, y_0) \in \mathrm{gph}(f+F)$ be such that condition* (A1) *in Theorem* 2.2 *holds. Let $\{A_k\}$ be a sequence of bounded linear operators from X to Y such that for every $k \in \mathbb{N}_0$ the mapping G_{A_k} defined in* (16) *is metrically regular at x_0 for y_0 with constant κ and neighborhoods $\mathbb{B}_a(x_0)$ and $\mathbb{B}_b(y_0)$, and*

$$\|f(x) - f(x') - A_k(x - x')\| \leq \delta\|x - x'\| \quad \textit{for any} \quad x, x' \in \mathbb{B}_a(x_0).$$

Then for every $\alpha \in (\kappa\delta, 1)$ there exists a sequence $\{x_k\}$ satisfying (15) *with starting point x_0 which is convergent to a solution $\bar{x} \in \mathrm{int}\,\mathbb{B}_a(x_0)$ of* (8) *with r-linear rate as in* (17).

3. Some special cases

Consider first the generalized equation (8) where the function f is continuously differentiable around the starting point x_0. Then we can take $A_k = Df(x_k)$ in the iteration (15) obtaining

$$f(x_k) + Df(x_k)(x_{k+1} - x_k) + F(x_{k+1}) \ni 0. \tag{23}$$

In the following theorem we obtain q-superlinear and q-quadratic convergence of the iteration (23) by concatenating the main Theorem 2.2 with conventional convergence results from [14], Theorems 6C.1 and 6D.2.

Theorem 3.1. *Consider the generalized equation* (8)*, a point* $(x_0, y_0) \in \mathrm{gph}(f+F)$ *and positive reals* κ*,* δ*,* a *and* b *such that condition (A1) in Theorem 2.2 is satisfied. Suppose that the function* f *is continuously differentiable in an open set containing* $\mathbb{B}_a(x_0)$*, for every* $z \in \mathbb{B}_a(x_0)$ *the mapping*

$$x \mapsto G_z(x) := f(x_0) + Df(z)(x - x_0) + F(x)$$

is metrically regular at x_0 *for* y_0 *with constant* κ *and neighborhoods* $\mathbb{B}_a(x_0)$ *and* $\mathbb{B}_b(y_0)$*, and also*

$$\|f(x) - f(x') - Df(x)(x - x')\| \le \delta\|x - x'\| \quad \text{for all } x, x' \in \mathbb{B}_a(x_0).$$

Then there exists a sequence $\{x_k\}$ *which satisfies the iteration* (23) *with starting point* x_0 *and converges q-superlinearly to a solution* $\bar{x}$ *of* (8) *in* $\operatorname{int}\mathbb{B}_a(x_0)$*. If the derivative mapping* Df *is Lipschitz continuous in* $\mathbb{B}_a(x_0)$*, then the sequence* $\{x_k\}$ *converges q-quadratically to* $\bar{x}$*.*

Proof. Clearly, for any sequence $\{x_k\}$ in $\mathbb{B}_a(x_0)$ and for each $k \in \mathbb{N}_0$ the mapping $A_k := Df(x_k)$ satisfies (A2) and (A3) of Theorem 2.2 with $\omega(\xi) := \delta$, $\xi \ge 0$. From condition (A1) there exists $\alpha \in (\kappa\delta, 1)$ such that

$$\|y_0\| < (1 - \alpha)b. \tag{24}$$

Hence we can apply Theorem 2.2, which yields the existence of a sequence $\{x_k\}$ satisfying (23) and converging to a solution $\bar{x} \in \operatorname{int}\mathbb{B}_a(x_0)$ of (8); furthermore

$$\|\bar{x} - x_0\| \le \frac{\alpha}{\delta(1-\alpha)}\|y_0\|.$$

Hence, for $v_0 := f(\bar{x}) - f(x_0) - Df(\bar{x})(\bar{x} - x_0)$ we have

$$\begin{aligned}\|y_0 + v_0\| &= \|y_0 + f(\bar{x}) - f(x_0) - Df(\bar{x})(\bar{x} - x_0)\| \le \|y_0\| + \delta\|\bar{x} - x_0\| \\ &\le \|y_0\| + \frac{\alpha}{1-\alpha}\|y_0\| = \frac{\|y_0\|}{1-\alpha} < b,\end{aligned}$$

where we use (24). Clearly, the mapping

$$x \mapsto G'(x) := f(\bar{x}) + Df(\bar{x})(x - \bar{x}) + F(x) = v_0 + G_{\bar{x}}(x)$$

is metrically regular at x_0 for $y_0 + v_0$ with constant κ and neighborhoods $\mathbb{B}_a(x_0)$ and $\mathbb{B}_b(y_0 + v_0)$. Let $r, s > 0$ be so small that

$$\mathbb{B}_r(\bar{x}) \subset \mathbb{B}_a(x_0) \quad \text{and} \quad \mathbb{B}_s(0) \subset \mathbb{B}_b(y_0 + v_0).$$

Then since $0 \in G'(\bar{x})$, the mapping G' is metrically regular at $\bar{x}$ for 0 with constant κ and neighborhoods $\mathbb{B}_r(\bar{x})$ and $\mathbb{B}_s(0)$. Hence we can apply Theorems 6C.1, resp. 6D.2, in [14], according to which there exists a neighborhood O of $\bar{x}$ such that for any starting point in O there exists a sequence $\{x'_k\}$ which is q-superlinearly, resp. q-quadratically, convergent to $\bar{x}$. But for some k sufficiently large the iterate x_k of the initial sequence will be in O and hence it can be taken as a starting point of a sequence $\{x'_k\}$ which converges q-superlinearly, resp. q-quadratically, to $\bar{x}$. □

In the theorem coming next we utilize an auxiliary result which follows from Proof I, with some obvious adjustments, of the extended Lyusternik-Graves theorem given in [14, Theorem 5E.1].

Lemma 3.2. *Consider a mapping $F\colon X \rightrightarrows Y$, a point $(x_0, y_0) \in \operatorname{gph} F$ and a function $g\colon X \to Y$. Suppose that there are $a' > 0$, $b' > 0$, $\kappa' \geq 0$, and $\mu \geq 0$ such that F is metrically regular at x_0 for y_0 with constant κ' and neighborhoods $\mathbb{B}_{a'}(x_0)$ and $\mathbb{B}_{b'}(y_0)$, the function g is Lipschitz continuous on $\mathbb{B}_{a'}(x_0)$ with constant μ, and $\kappa'\mu < 1$. Then for any positive constants a and b such that*

$$\begin{cases} \dfrac{1}{1-\kappa'\mu}\left[(1+\kappa'\mu)a + \kappa' b\right] + a < a', \\ b + \mu\left(\dfrac{1}{1-\kappa'\mu}\left[(1+\kappa'\mu)a + \kappa' b\right] + a\right) < b', \end{cases} \tag{25}$$

the mapping $g + F$ is metrically regular at x_0 for $y_0 + g(x_0)$ with any constant $\kappa > \kappa'/(1-\kappa'\mu)$ and neighborhoods $\mathbb{B}_a(x_0)$ and $\mathbb{B}_b(y_0 + g(x_0))$.

Theorem 3.3. *Let the numbers $a > 0$, $b > 0$, $\kappa > 0$ and $\delta > 0$ and the points $x_0 \in X$, $y_0 \in f(x_0) + F(x_0)$ be such that* (A1) *is fulfilled. Let the numbers a', b', κ' be such that:*

$$0 < \kappa' < \frac{\kappa}{1+\kappa\delta}, \quad a' > 2a(1+\kappa\delta) + \kappa b, \quad b' > (2a\delta + b)(1+\kappa\delta). \tag{26}$$

Let f be Fréchet differentiable in an open set containing $\mathbb{B}_a(x_0)$, let $\mathcal{T} \subset \mathcal{L}(X,Y)$, and let $A_k\colon X^k \to \mathcal{T}$ be any sequence with $\sup_{A\in\mathcal{T}} \|A - A_0(x_0)\| \leq \delta$. Assume that

(A2') *the mapping $x \mapsto G(x) := f(x_0) + A_0(x_0)(x - x_0) + F(x)$ is metrically regular with constant κ' and neighborhoods $\mathbb{B}_{a'}(x_0)$ and $\mathbb{B}_{b'}(y_0)$;*

(A3') *$\|A - Df(x)\| \leq \delta$ whenever $A \in \mathcal{T}$ and $x \in \mathbb{B}_a(x_0)$.*

Then the first claim in Theorem 2.2 *holds.*

Proof. We shall prove that conditions (A2) and (A3) in Theorem 2.2 are satisfied. To check (A2), pick any $A \in \mathcal{T}$ and let G_A be the mapping from Theorem 2.2 (with $A_k := A$). Define $g(x) := (A - A_0)(x - x_0)$, $x \in X$, so that

$G_A = G + g$. Then g is Lipschitz continuous with constant δ and we can apply Lemma 3.2 with $\mu := \delta$, which implies (A2).

It remains to check (A3). Let $\omega(\xi) := \delta$ for each $\xi \geq 0$. Pick arbitrary points x_0, x_1, ..., x_k in $\mathbb{B}_a(x_0)$ and set $A_k := A_k(x_0, \ldots, x_k)$. Finally, fix any $x \in \mathbb{B}_a(x_0)$. By the mean value theorem there is $z \in \mathbb{B}_a(x_0)$ such that $f(x) - f(x_k) - Df(z)(x - x_k) = 0$. Hence

$$\|f(x) - f(x_k) - A_k(x - x_k)\| = \|Df(z)(x - x_k) - A_k(x - x_k)\| \leq \delta\|x - x_k\|.$$

This proves (A3) and therefore the theorem. □

Next, we state and prove a theorem regarding convergence of the Newton's method applied to a generalized equation, which is close to the original statement of Kantorovich. The result is somewhat parallel to [13, Theorem 2] but on different assumptions.

Theorem 3.4. *Let the positive scalars L, κ, a, b and the points $x_0 \in X$, $y_0 \in f(x_0) + F(x_0)$ be such that the function f is differentiable in an open neighborhood of the ball $\mathbb{B}_a(x_0)$ and its derivative Df is Lipschitz continuous on $\mathbb{B}_a(x_0)$ with Lipschitz constant L and also the mapping*

$$x \mapsto G(x) := f(x_0) + Df(x_0)(x - x_0) + F(x) \tag{27}$$

is metrically regular at x_0 for y_0 with constant κ and neighborhoods $\mathbb{B}_a(x_0)$ and $\mathbb{B}_b(y_0)$. Furthermore, let $\kappa' > \kappa$ and assume that for $\eta := \kappa'\|y_0\|$ we have

$$h := \kappa' L\eta < \frac{1}{2}, \quad \bar{t} := \frac{1}{\kappa' L}(1 - \sqrt{1 - 2h}) \leq a \quad \text{and} \quad \|y_0\| + L\bar{t}^2 \leq b. \tag{28}$$

Then there is a sequence $\{x_k\}$ generated by the iteration (23) with initial point x_0 which stays in $\mathbb{B}_a(x_0)$ and converges to a solution $\bar{x}$ of the generalized equation (8); moreover, the rate of the convergence is

$$\|x_k - \bar{x}\| \leq \frac{2\sqrt{1 - 2h}\Theta^{2^k}}{\kappa' L(1 - \Theta^{2^k})}, \qquad \text{for } k = 1, 2, \ldots, \tag{29}$$

where

$$\Theta := \frac{1 - \sqrt{1 - 2h}}{1 + \sqrt{1 - 2h}}\ .$$

If the mapping G is not only metrically regular but also strongly metrically regular with the same constant and neighborhoods, then there is no other sequence $\{x_k\}$ generated by the method (23) starting from x_0 which stays in $\mathbb{B}_a(x_0)$.

Proof. In the sequel we will utilize the following inequality for $u, v \in \mathbb{B}_a(x_0)$:

$$\|f(u) - f(v) - Df(v)(u - v)\| = \|\int_0^1 [Df(v + s(u - v)) - Df(v)](u - v)\,\mathrm{d}s\|$$
$$\leq L\|u - v\|^2 \int_0^1 s\,\mathrm{d}s = \frac{L}{2}\|u - v\|^2.$$

We apply a modification of the majorization technique from [17]. Consider a sequence of reals t_k satisfying $t_0 = 0,\ t_{k+1} = s(t_k),\ k = 0, 1, \ldots$, where

$$s(t) = t - (p'(t))^{-1}p(t), \quad p(t) = \frac{\kappa' L}{2}t^2 - t + \eta.$$

It is known from [17] that the sequence $\{t_k\}$ is strictly increasing, convergent to $\bar{t}$, and also

$$t_{k+1} - t_k = \frac{\kappa' L(t_k - t_{k-1})^2}{2(1 - \kappa' L t_k)}, \qquad k = 0, 1, \ldots. \tag{30}$$

Furthermore,

$$\bar{t} - t_k \leq \frac{2\sqrt{1-2h}\Theta^{2^k}}{\kappa' L(1 - \Theta^{2^k})}, \qquad \text{for } k = 0, 1, \ldots. \tag{31}$$

We will show, by induction, that there is a sequence $\{x_k\}$ in $\mathbb{B}_a(x_0)$ fulfilling (23) with the starting point x_0 which satisfies

$$\|x_{k+1} - x_k\| \leq t_{k+1} - t_k, \quad k = 0, 1, \ldots. \tag{32}$$

This implies that $\{x_k\}$ is a Cauchy sequence, hence convergent to some $\bar{x}$, which, by passing to the limit in (23), is a solution of the problem at hand. Combining (31), (30) and (32) we obtain (29).

Let $k = 0$. If $y_0 = 0$ then we take $x_1 = x_0$. If not, since $0 \in \mathbb{B}_b(y_0)$ and $y_0 \in G(x_0)$, from the metric regularity of the mapping G in (27) we obtain

$$\operatorname{dist}(x_0, G^{-1}(0)) \leq \kappa\|y_0\| < \kappa'\|y_0\|,$$

hence there exists $x_1 \in G^{-1}(0)$ such that

$$\|x_1 - x_0\| < \kappa'\|y_0\| = \eta = t_1 - t_0.$$

Suppose that for some $k \in \mathbb{N}$ we have already found points x_0, x_1, …, x_k in $\mathbb{B}_a(x_0)$ generated by (23) such that

$$\|x_j - x_{j-1}\| \leq t_j - t_{j-1} \quad \text{for each} \quad j = 1, \ldots, k.$$

Without loss of generality, let $x_k \neq x_0$; otherwise there is nothing to prove. We have

$$\|x_k - x_0\| \leq \sum_{j=1}^{k} \|x_j - x_{j-1}\| \leq \sum_{j=1}^{k}(t_j - t_{j-1}) = t_k - t_0 = t_k < \bar{t} \leq a.$$

Furthermore, for every $x \in \mathbb{B}_{\bar{t}-t_k}(x_k) \subset \mathbb{B}_{\bar{t}}(x_0)$, we obtain

$$\begin{aligned} &\|f(x_0) + Df(x_0)(x - x_0) - f(x_k) - Df(x_k)(x - x_k)\| \leq \\ \leq\ &\|f(x) - f(x_0) - Df(x_0)(x - x_0)\| + \|f(x) - f(x_k) - Df(x_k)(x - x_k)\| \\ &\leq \frac{L}{2}\big(\|x - x_0\|^2 + \|x - x_k\|^2\big) < L\bar{t}^2 \leq b - \|y_0\|. \end{aligned}$$

In particular, we have $f(x_0)+Df(x_0)(x-x_0)-f(x_k)-Df(x_k)(x-x_k) \in \mathbb{B}_b(y_0)$. Moreover,

$$r := \frac{\frac{1}{2}\kappa' L\|x_k - x_{k-1}\|^2}{1-\kappa' L\|x_k - x_0\|} \le \frac{\kappa' L(t_k - t_{k-1})^2}{2(1-\kappa' L t_k)} = t_{k+1} - t_k.$$

Since $x_k \in \mathbb{B}_a(x_0)$ is generated by (23) from x_{k-1}, we get

$$f(x_0) + Df(x_0)(x_k - x_0) - f(x_{k-1}) - Df(x_{k-1})(x_k - x_{k-1}) \in G(x_k). \quad (33)$$

Now consider the set-valued mapping

$$X \ni x \mapsto \Phi_k(x) := G^{-1}(f(x_0)+Df(x_0)(x-x_0)-f(x_k)-Df(x_k)(x-x_k)) \subset X.$$

If $x_k = x_{k-1}$ then take $x_{k+1} = x_k$. Suppose that $x_k \neq x_{k-1}$. From (33) we obtain

$$\begin{aligned}
\operatorname{dist}(x_k, \Phi_k(x_k)) &= \operatorname{dist}(x_k, G^{-1}\left(f(x_0) + Df(x_0)(x_k - x_0) - f(x_k)\right) \\
&\le \kappa \operatorname{dist}(f(x_0) + Df(x_0)(x_k - x_0) - f(x_k), G(x_k)) \\
&\le \kappa \|f(x_k) - f(x_{k-1}) - Df(x_{k-1})(x_k - x_{k-1})\| \\
&\le \frac{1}{2}\kappa L\|x_k - x_{k-1}\|^2 < \frac{\frac{1}{2}\kappa' L\|x_k - x_{k-1}\|^2}{1-\kappa' L\|x_k - x_0\|}(1-\kappa' L\|x_k - x_0\|) \\
&= r(1-\kappa' L\|x_k - x_0\|).
\end{aligned}$$

Let $u, v \in \mathbb{B}_{\bar{t}-t_k}(x_k)$ and let $z \in \Phi_k(u) \cap \mathbb{B}_{\bar{t}-t_k}(x_k)$. Then

$$f(x_0) + Df(x_0)(u - x_0) - f(x_k) - Df(x_k)(u - x_k) \in G(z).$$

Hence,

$$\begin{aligned}
\operatorname{dist}(z, \Phi_k(v)) &= \operatorname{dist}(z, G^{-1}(f(x_0)+Df(x_0)(v-x_0)-f(x_k)-Df(x_k)(v-x_k)) \\
&\le \kappa \operatorname{dist}(f(x_0) + Df(x_0)(v - x_0) - f(x_k) - Df(x_k)(v - x_k), G(z)) \\
&\le \kappa \|f(x_0) + Df(x_0)(v - x_0) - f(x_k) - Df(x_k)(v - x_k) - \\
&\qquad - (f(x_0) + Df(x_0)(u - x_0) - f(x_k) - Df(x_k)(u - x_k))\| \\
&\le \kappa \|Df(x_0) - Df(x_k)\|\|u - v\| \le (\kappa' L\|x_k - x_0\|)\|u - v\|.
\end{aligned}$$

Since $\mathbb{B}_r(x_k) \subset \mathbb{B}_{\bar{t}-t_k}(x_k)$, by applying the contraction mapping theorem [14, Theorem 5E.2] we obtain that there exists a fixed point $x_{k+1} \in \mathbb{B}_r(x_k)$ of Φ_k. Hence

$$x_{k+1} \in G^{-1}\left(f(x_0) + Df(x_0)(x_{k+1} - x_0) - f(x_k) - Df(x_k)(x_{k+1} - x_k)\right),$$

that is, x_{k+1} is a Newton iterate from x_k according to (23). Furthermore,

$$\|x_{k+1} - x_k\| \le r \le t_{k+1} - t_k.$$

Then

$$\|x_{k+1} - x_0\| \le \sum_{j=1}^{k+1} \|x_j - x_{j-1}\| \le \sum_{j=1}^{k+1} (t_j - t_{j-1}) = t_{k+1} - t_0 = t_{k+1} < \bar{t} \ \le \ a.$$

The induction step is complete and so is the proof. □

At the end of this section we add some comments on the results presented in this paper and give some examples. First, we would like to reiterate that, in contrast to the conventional approach to proving convergence of Newton's method where certain conditions *at a solution* are imposed, the Kantorovich theorem utilizes conditions for *a given neighborhood of the starting point* associated with some constants, the relations among which gives the existence of a solution and convergence towards it. In the framework of the main Theorem 2.2, among the constants taken into account are the radius a of the given neighborhood of the starting point x_0, the norm of the residual $\|y_0\|$ at the starting point, the constant of metric regularity κ, and the constant δ measuring the "quality" of the approximation of the "derivative" of the function f by the operators A_k. These constants are interconnected through relations that cannot be removed even in the particular cases of finite dimensional smooth problems, or nonsmooth problems where elements of the Clarke's generalized Jacobian play the role of approximations. In the smooth case the constant δ may be measured by the diameter of the set $\{\|Df(x)\| : x \in I\!B_a(x_0)\}$ or by La if Df is Lipschitz continuous with a Lipschitz constant L. In the nonsmooth case however, it is not sufficient to assume that the diameter of the generalized Jacobian around x_0 is less than δ. One may argue that for any small δ there exists a positive ε such that the generalized Jacobian has the "strict derivative property" displayed in [14, 6F.3] but in order this to work we need ε to match a. Note that if the residual $\|y_0\| = 0$ then we can always choose the constant a sufficiently small, but this may not be the case for the Kantorovich theorem. It would be quite interesting to know exactly "how far" the conventional and the Kantorovich theorems are from each other in particular for problems involving nonsmooth functions.

Next, we will present some elementary examples that illustrate the difference between the Newton method and the chord method with $A_k = A_0$ for all k, as well as the conditions for convergence appearing in the results presented.

Example 3.5. We start with the smooth one-dimensional example[3] to find a *nonnegative* root of $f(x) := (x-1)^2 - 4$; it is elementary to check that $\bar{x} = 3$ is the only solution. For every $x_0 > 1$ the usual Newton iteration is given by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = \frac{x_k^2 + 3}{2(x_k - 1)}.$$

[3]Note that this problem can be written as a generalized equation.

This iteration is convergent quadratically which agrees with the theory. The chord method,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)} = \frac{2x_0x_k - x_k^2 + 3}{2(x_0 - 1)},$$

converges linearly if there is a constant $c < 1$ and a natural number N such that

$$\frac{|x_{k+1} - 3|}{|x_k - 3|} = \frac{|2x_0 - x_k - 3|}{2|x_0 - 1|} \leq c$$

for every $k \geq N$, but it may not be convergent for x_0 not close enough to 3. For example take $x_0 = 1 + \frac{2}{\sqrt{5}}$. Then the method oscillates between the points $1 + \frac{2}{\sqrt{5}}$ and $1 + \frac{6}{\sqrt{5}}$. The method converges q-superlinearly whenever

$$\lim_{k\to\infty} \frac{|x_{k+1} - 3|}{|x_k - 3|} = \lim_{k\to\infty} \frac{|2x_0 - x_k - 3|}{2|x_0 - 1|} = 0;$$

but this holds only for $x_0 = 3$. Hence, even in the case when there is convergence, it is not q-superlinear.

Let us check the assumptions of Theorem 2.2 with $\omega \equiv \delta$. Given x_0 and $a > 0$ we can calculate how large κ and δ have to be such that conditions (A2) and (A3) are fulfilled. Let us focus on the case $x_0 > 1$. For (A2) to hold we have to assume $a < x_0 - 1$. Then on $\mathbb{B}_a(x_0)$ we have that f' is positive and increasing. Hence (A2) and (A3) are satisfied for $\kappa = 1/f'(x_0 - a) = 1/(2(x_0 - a - 1)$ and $\delta = f'(x_0 + a) - f'(x_0 - a) = 4a$. For fixed x_0 let us find a such that (A1) holds as well, i.e.,

$$\|y_0\| < (1 - \kappa\delta)\frac{a}{\kappa} = 2a(x_0 - 3a - 1). \tag{34}$$

The right hand side is maximal for $a = \frac{x_0 - 1}{6}$. Expressing both sides of this inequality in terms of x_0, we obtain that if $x_0 \in (1 + 2\sqrt{6/7}, 1 + 2\sqrt{6/5})$ then we have convergence.

The following example from [26], see also [25], example BE.1, shows lack of convergence of the nonsmooth Newton method if the function is not semismooth at the solution. But it is also an example which illustrates Corollary 2.5.

Example 3.6. Consider intervals $I(n) = [n^{-1}, (n-1)^{-1}] \subset \mathbb{R}$ and define $c(n) = \frac{1}{2}(n^{-1} + (n-1)^{-1})$ for $n \geq 2$. Let g_n be the linear function through the points $((n-1)^{-1}, (n-1)^{-1})$ and $(-c(n), 0)$, and h_n be the linear function through the points (n^{-1}, n^{-1}) and $(c(2n), 0)$. Then

$$g_n(x) = \frac{2n}{4n-1}x + \frac{2n-1}{(n-1)(4n-1)} \quad \text{and} \quad h_n(x) = \frac{4(2n-1)}{4n-3}x - \frac{4n-1}{n(4n-3)}.$$

Now define $f(x) = \min\{g_n(x), h_n(x)\}$ for $x \in I(n)$, $f(0) = 0$ and for $x < 0$: $f(x) = -f(-x)$. Then the equation $f(x) = 0$ has the single solution $\bar{x} = 0$ and

we have that $\bar{\partial} f(0) = [\frac{1}{2}, 2]$. If we try to apply Corollary 2.5 for a neighborhood that contains $\bar{x} = 0$ we have to choose $\delta \geq \frac{3}{2}$ and $\kappa \geq 2$; but then $\kappa\delta > 1$. In this case for any starting point $x_0 \neq 0$ the Newton iteration does not converge, as shown in [26].

A similar example follows to which Corollary 2.5 can be applied.

Example 3.7. Define

$$g(x) := \begin{cases} 2 \text{ if } x \in \cup_{n\in\mathbb{Z}}[2^{2n-1}, 2^{2n}) \\ 3 \text{ if } x \in \cup_{n\in\mathbb{Z}}[2^{2n}, 2^{2n+1}) \end{cases}.$$

Let $f(x) := \int_0^x g(t)\mathrm{d}t$ for $x \geq 0$ and $f(x) = -f(-x)$ for $x < 0$. The function f is well defined on $\mathbb{R}$ with a unique root at $\bar{x} = 0$. For any starting point x_0 the assumptions for Corollary 2.5 are then fulfilled with $\kappa = \frac{1}{2}$ and $\delta = 1$ and each $a > 0$. Both the Newton and the chord method converge linearly.

4. Nonsmooth inequalities

Suppose that K is a nonempty subset of Y and let $F(x) := K$ for each $x \in X$. Then the generalized equation (8) reads as

$$f(x) + K \ni 0. \tag{35}$$

When $f\colon \mathbb{R}^n \to \mathbb{R}^m$ and $K := \mathbb{R}^m_+$ then the above inclusion corresponds to a system of m nonlinear (possibly nonsmooth) inequalities: find $x \in \mathbb{R}^n$ such that

$$f_1(x) \leq 0, \quad f_2(x) \leq 0, \quad \ldots, \quad f_m(x) \leq 0.$$

Kantorovich-type theorems for exact Newton's method for solving (35) with K being a closed convex cone and f being smooth can be found in [4, Chapter 2.6] and [31]. An inexact Newton's method is treated in a similar way in [16]. The paper [28] deals with a generalized equation of the form

$$g(x) + h(x) + K \ni 0, \tag{36}$$

where $g : X \to Y$ is a smooth function having a Lipschitz derivative on a neighborhood $O \subset X$ of a (starting) point $x_0 \in X$ and the function $h : X \to Y$ is Lipschitz continuous on O. The algorithm proposed therein reads as: given $x_k \in X$ find x_{k+1} satisfying

$$g(x_k) + h(x_k) + g'(x_k)(x_{k+1} - x_k) + K \ni 0. \tag{37}$$

Key assumptions are, similar to [31, 4, 16], that $T := g'(x_0)(\cdot) + K$ maps X onto Y and

$$\|T^{-1}\|^- := \sup_{\|y\|\leq 1} \inf_{x\in T^{-1}(y)} \|x\| \leq b$$

for a sufficiently small number $b > 0$. Then Open Mapping Theorem [5, Theorem 2.2.1] (see also [14, Exercise 5C.4]) implies that T is metrically regular at zero for zero with any constant $\kappa > b$ and neighborhoods X and Y. Moreover, the Lipschitz constants of g' and h are assumed to be small compared to b. Clearly, (37) corresponds to our iteration scheme with $f := g + h$ and $A_k := g'(x_k)$, and, since A_k does not take into account the non-smooth part, it is expected to be slower in general (or not even applicable) as we will show on two toy examples below.

Consider a sequence $\{A_k\}$ in $\mathcal{L}(X, Y)$ and a starting point $x_0 \in X$. Given $k \in \mathbb{N}_0$, $x_k \in X$, and A_k, let

$$\Omega_k := \{u \in X \mid \ g(x_k) + h(x_k) + A_k(u - x_k) + K \ni 0\}.$$

The next iterate x_{k+1} generated by (15), which is sure to exist under the metric regularity assumption in Theorem 2.2, is any point lying in Ω_k such that

$$\|x_{k+1} - x_k\| \leq \kappa' \operatorname{dist}(-g(x_k) - h(x_k), K),$$

where $\kappa' > \kappa$ satisfies (18) and the right-hand side of the above inequality corresponds to a residual at the step k. To sum up, for the already computed x_k, the next iterate x_{k+1} can be found as a solution of the problem:

$$\text{minimize} \quad \varphi_k(x) \quad \text{subject to} \quad x \in \Omega_k,$$

where $\varphi_k : X \to [0, \infty)$ is a suitably chosen function. In [28], $\varphi_k = \| \cdot - x_k\|_2$ is used. In the following examples we solve the linearized problem in MATLAB using either function *fmincon* for $\varphi_k = \| \cdot - x_k\|_2^2$ or *quadprog* for $\varphi_k(x) := \frac{1}{2}x^T x - x_k^T x$. We will compare the following three versions of (15) for solving (36) with different choices of A_k at the step $k \in \mathbb{N}_0$ and current iterate x_k:

(C1) $A_k := g'(x_k)$;

(C2) $A_k \in \bar{\partial}(g + h)(x_k) = g'(x_k) + \bar{\partial}h(x_k)$;

(C3) $A_k := A_0$, where A_0 is a fixed element of $\bar{\partial}(g + h)(x_0) = g'(x_0) + \bar{\partial}h(x_0)$.

Example 4.1. Consider the system from [28]:

$$\begin{aligned} x^2 + y^2 - |x - 0.5| - 1 &\leq 0, \\ x^2 + (y - 1)^2 - |x - 0.5| - 1 &\leq 0, \\ (x - 1)^2 + (y - 1)^2 - 1 &= 0. \end{aligned} \tag{38}$$

Observe that the exact solutions are given by $y = 1 \pm \sqrt{2x - x^2}$ if $0 \leq x \leq (11 - 6\sqrt{3})/26$ and $y = 1 - \sqrt{2x - x^2}$ when $(11 - 6\sqrt{3})/26 \leq x \leq 1/2$, in particular, the points $(x_1^*, y_1^*) := (0.5, 1 - \sqrt{3}/2)$ and $(x_2^*, y_2^*) = (1 - \sqrt{2}/2, 1 - \sqrt{2}/2)$ solve the problem. Then setting $g(x, y) := (x^2 + y^2 - 1, x^2 + (y - 1)^2 - 1, (x - 1)^2 +$

Step	fmincon			quadprog		
k	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
0	5.0 E-2	5.0 E-2	5.0 E-2	5.0 E-2	5.0 E-2	5.0 E-2
1	2.4 E-2	2.0 E-3	2.0 E-3	2.5 E-2	2.0 E-3	2.0 E-3
2	1.2 E-2	2.3 E-6	2.3 E-6	1.3 E-3	2.3 E-6	2.3 E-6
4	3.1 E-3	1.0 E-8	1.0 E-8	3.1 E-3	6.5 E-9	6.5 E-9

Table 4.1: $\|(x_1^*, y_1^*) - (x_k, y_k)\|_\infty$ in Example 4.1 for $(x_0, y_0) = (0.55, 0.1)$.

Step	fmincon			quadprog		
k	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
0	2.9 E-1	2.9 E-1	2.9 E-1	2.9 E-1	2.9 E-1	2.9 E-1
1	4.2 E-2	4.2 E-2	4.2 E-2	4.2 E-2	4.2 E-2	4.2 E-2
2	1.2 E-3	1.2 E-3	1.2 E-3	1.2 E-3	1.2 E-3	1.2 E-3
4	1.1 E-10	5.2 E-10	5.2 E-10	7.9 E-13	7.9 E-13	5.2 E-13
7	1.1 E-10	5.2 E-10	5.2 E-10	1.6 E-16	1.1 E-16	1.1 E-16

Table 4.2: $\|(x_2^*, y_2^*) - (x_k, y_k)\|_\infty$ in Example 4.1 for $(x_0, y_0) = (0, 0)$.

$(y-1)^2-1)$, $h(x,y) := (-|x-0.5|, -|x-0.5|, 0)$, and $K := \mathbb{R}^2_+ \times \{0\}$ we arrive at (36). Denote

$$H(x,y) := \begin{pmatrix} 2x - \operatorname{sgn}(x-0.5) & 2y \\ 2x - \operatorname{sgn}(x-0.5) & 2(y-1) \\ 2(x-1) & 2(y-1) \end{pmatrix},$$

with $\operatorname{sgn}(u) := 1$, if $u > 0$, and $\operatorname{sgn}(u) := -1$ otherwise. In (C2) we set $A_k := H(x_k, y_k)$ for each $k \in \mathbb{N}_0$ and in (C3) we put $A_0 := H(x_0, y_0)$.

From Table 4.1, in which 5.0 E-2 stands for 5.0×10^{-2}, we see that the convergence of (15) with the choice (C1) and the starting point $(0.55, 0.1)$ is much slower than (15) with the choice (C3). Both *quadprog* and *fmincon* are of almost the same efficiency.

From Table 4.2 we see that for the starting point $(0, 0)$ all the choices (C1)–(C3) provide similar accuracy but we get substantially better results when *quadprog* is used to solve the linearized problem.

Example 4.2. Consider the system

$$x^2 + y^2 - 1 \le 0 \quad \text{and} \quad -|x| - |y| + \sqrt{2} \le 0 \tag{39}$$

having four distinct solutions. Set $g(x,y) := (x^2+y^2-1, 0)$, $h(x,y) := (0, -|x|-|y| + \sqrt{2})$, $K := \mathbb{R}^2_+$, and $H(x,y) = \begin{pmatrix} 2x & 2y \\ -\operatorname{sgn}(x) & -\operatorname{sgn}(y) \end{pmatrix}$.

Step k	fmincon		quadprog	
	(C2)	(C3)	(C2)	(C3)
0	7.0 E-1	7.0 E-1	7.0 E-1	7.0 E-1
1	2.5 E-9	2.5 E-9	0	0
2	7.5 E-8	7.5 E-8	0	0
4	1.2 E-8	1.2 E-8	0	0
7	8.5 E-8	8.5 E-8	0	0
10	8.5 E-9	3.7 E-9	0	0

Table 4.3: $\|(-\sqrt{2}/2, -\sqrt{2}/2) - (x_k, y_k)\|_\infty$ in Example 4.2 for $(x_0, y_0) = (0, 0)$.

Step	fmincon			quadprog		
k	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
0	9.9 E 2	9.9 E 2	9.9 E 2	9.9 E 2	9.9 E 2	9.9 E 2
1	4.9 E 2	4.9 E 2	4.9 E 2	–	4.9 E 2	4.9 E 2
4	6.1 E 1	6.1 E 1	6.1 E 1	–	6.1 E 1	6.1 E 1
10	5.0 E-1	6.0 E-1	6.0 E-1	–	5.8 E-1	8.3 E-1
21	7.0 E-1	3.0 E-4	1.5 E-1	–	2.8 E-4	1.4 E 0
40	7.0 E-1	5.3 E-9	1.5 E-1	–	1.0 E-8	1.4 E 0

Table 4.4: $\|(-\sqrt{2}/2, \sqrt{2}/2) - (x_k, y_k)\|_\infty$ in Example 4.2 for $(x_0, y_0) = (99, -999)$.

As before, in (C2) we set $A_k := H(x_k, y_k)$ for each $k \in \mathbb{N}_0$ and in (C3) we put $A_0 := H(x_0, y_0)$.

For the starting point $(0, 0)$ the method (15) with (C1) fails. The convergence for the remaining two choices (C2) and (C3) can be found in Table 4.3. Note that using *quadprog* we find a solution (up to a machine epsilon) after one step and the iteration using *fmincon* gives the precision 10^{-9} at most.

For the starting point $(99, -999)$ the method (15) with (C1) and (C3) do not converge – see Table 4.4. The only convergent scheme is (15) with (C2) (note that we start far away from the solution).

5. Numerical experiments for a model of economic equilibrium

In this section we present numerical results for a model of economic equilibrium presented in [12] and solved by using the Newton, the chord and the hybrid method with various parameter choices. A detailed description of the model is given in [12] so we shall not repeat it here.

The equilibrium problem considered is described by the variational inequality

$$0 \in g(p, m, x, \lambda, m^0, x^0) + N_C(p, m, x, \lambda), \tag{40}$$

where

$$g(p, m, x, \lambda, m^0, x^0) = \begin{pmatrix} \sum_{i=1}^{r}(x_i^0 - x_i) \\ \cdots \\ \lambda_i - \nabla_{m_i} u_i(m_i, x_i) \\ \cdots \\ \lambda_i p - \nabla_{x_i} u_i(m_i, x_i) \\ \cdots \\ m_i^0 - m_i + \langle p, x_i^0 - x_i \rangle \\ \cdots \end{pmatrix}$$

and N_C is the normal cone to the set

$$C = \mathbb{R}^n_+ \times \mathbb{R}^r_+ \times U_1 \times \cdots \times U_r \times \mathbb{R}^r_+.$$

Here r is the number of agents trading n goods, who start with initial vectors of goods x_i^0 and initial amount of money m_i^0. Further, x represents the vector of goods, p is the vector of prices, m is the vector of the amounts of money, U_i are closed subsets of $\mathbb{R}^n_+$. The functions u_i are utility functions and are given by

$$u_i(m_i, x_i) = \alpha_i \ln(m_i) + \chi_{\geq m_i^1}(m_i)\gamma_i(m_i - m_i^1)^2 + \sum_{j=1}^{n} \beta_{ij} \ln(x_{ij})$$

where $\gamma_i \in \mathbb{R}$, α_i, β_{ij} and m_i^1 are positive constants and

$$\chi_{\geq m_i^1}(m_i) = \begin{cases} 1 & m_i \geq m_i^1 \\ 0 & \text{otherwise} \end{cases},$$

that is, when γ_i is different from zero then $\nabla_{m_i} u_i$, and hence g, are not differentiable.

The numerical implementation of Newton's method for this variational inequality has been done in Matlab. Each step of the method reduces to solving a linear complementarity problem (LCP). To solve these problems we used the Path-LCP solver available at [11]. For the linearization for the term involving χ we use the zero vector which is always an element of Clarke's generalized Jacobian of that function.

The computations are done for the following data (similar to [3]). We set the parameters as $n = r = 10$ (so in total we have 130 variables), $\alpha_i = \beta_{ij} = 1$ and $U_i = [0.94, 1.08]^n$ and use random initial endowments $m_i^0 \in [1, 1.3]$ and $x_{ij}^0 \in [0.94, 1.09]$.

First we consider at the smooth problem, that is, with $\gamma_i = 0$ for all $i = 1, 2, \ldots, 10$. We use the Newton method with starting points $p_j^s = m_i^s =$

Step	$k=1$	$k=2$	$k=3$	$k=5$	$k=100$
0	9.7 E-1	9.7 E-1	9.7 E-1	9.7 E-1	9.7 E-1
1	2.0 E-1	2.0 E-1	2.0 E-1	2.0 E-1	2.0 E-1
2	3.9 E-3	3.5 E-2	3.5 E-2	3.5 E-2	3.5 E-2
3	1.5 E-6	1.9 E-4	3.3 E-3	3.3 E-3	3.3 E-3
4	0	2.2 E-6	2.0 E-6	1.2 E-3	1.2 E-3
5	-	0	0	2.1 E-4	2.1 E-4
6	-	-	-	0	2.1 E-5

Table 5.1: Absolute errors with starting values $p_j^s = m_i^s = x_{ij}^s = \lambda_i^s = 1$.

Step	$k=1$	$k=2$	$k=3$	$k=5$	$k=100$
0	1.1 E 0	1.1 E 0	1.1 E 0	1.1 E 0	1.1 E 0
1	1.0 E 0	1.0 E 0	1.0 E 0	1.0 E 0	1.0 E 0
2	1.3 E-1	7.6 E-1	7.6 E-1	7.6 E-1	7.6 E-1
3	1.8 E-3	3.5 E-2	4.2 E-1	4.2 E-1	4.2 E-1
4	0	9.1 E-4	1.7 E-2	2.7 E-1	2.7 E-1
5	-	0	1.4 E-3	1.6 E-1	1.6 E-1
6	-	-	1.9 E-4	2.2 E-3	1.0 E-1

Table 5.2: Absolute errors with starting values $p_j^s = m_i^s = x_{ij}^s = \lambda_i^s = 0.97$.

Step	$k=1$	$k=2$	$k=3$	$k=5$	$k=100$
0	1.2 E 0	1.2 E 0	1.2 E 0	1.2 E 0	1.2 E 0
1	1.7 E 0	1.7 E 0	1.7 E 0	1.7 E 0	1.7 E 0
2	4.3 E-1	1.8 E 0	1.8 E 0	1.8 E 0	1.8 E 0
3	1.6 E-2	2.5 E-1	1.8 E 0	1.8 E 0	1.8 E 0
4	1.1 E-5	2.3 E-2	4.4 E-1	1.8 E 0	1.8 E 0
5	0	2.1 E-5	2.1 E-1	1.8 E 0	1.8 E 0
6	-	0	1.5 E-1	4.7 E-1	1.9 E 0

Table 5.3: Absolute errors with starting values $p_j^s = m_i^s = x_{ij}^s = \lambda_i^s = 0.96$.

Step	$k=1$	$k=2$	$k=3$	$k=5$	$k=100$
0	2.1 E 0	2.1 E 0	2.1 E 0	2.1 E 0	2.1 E 0
1	4.5 E -1	4.5 E -1	4.5 E -1	4.5 E -1	4.5 E -1
2	6.2 E -2	8.2 E -2	8.2 E -2	8.2 E -2	8.2 E -2
3	1.5 E -4	6.9 E -4	2.7 E -2	2.7 E -2	2.7 E -2
4	0	9.1 E -6	5.3 E -5	1.3 E -2	1.3 E -2
5	–	0	5.9 E -7	3.7 E -3	3.7 E -3
6	–	–	0	3.3 E -6	1.1 E -3

Table 5.4: Absolute errors with parameters $m_i^1 = 0.8$ and $\gamma_i = 0.5$.

Step	$k=1$	$k=2$	$k=3$	$k=5$	$k=100$
0	4.1 E 0	4.1 E 0	4.1 E 0	4.1 E 0	4.1 E 0
1	1.5 E 0	1.5 E 0	1.5 E 0	1.5 E 0	1.5 E 0
2	1.2 E 0	2.8 E -1	2.8 E -1	2.8 E -1	2.8 E -1
3	1.3 E -2	3.0 E -2	2.7 E -1	2.7 E -1	2.7 E -1
4	1.1 E -5	5.3 E -3	2.3 E -3	1.4 E -1	1.4 E -1
5	0	0	4.2 E -5	6.9 E -2	6.9 E -2
6	–	–	1.5 E -6	3.8 E -4	8.0 E -2

Table 5.5: Absolute errors with parameters $m_i^1 = 0.8$ and $\gamma_i = 1$.

$x_{ij}^s = \lambda_i^s = 1$, where we update the Jacobian iteration every k steps. For $k = 1, 2, 3, 5, 100$ we get a solution with error $\varepsilon = 10^{-7}$ after $4, 5, 5, 6, 9$ iterations, respectively. Then, while the number of iterations needed increases the number of times to calculate a derivative decreases from 4 to 1. Table 5.1 shows the errors to the solution.

If we change the starting points to $p_j^s = m_i^s = x_{ij}^s = \lambda_i^s = 0.97$ the number of iterations needed increases to $4, 5, 7, 9, 32$. Again, the number of times we update the Jacobian decreases from 4 to 1. The errors are shown in Table 5.2. One can see that, as expected, the choice of the starting point becomes more important if the Jacobian is not updated after every iteration. This is even more evident if we change the starting values to $p_j^s = m_i^s = x_{ij}^s = \lambda_i^s = 0.96$, where the pure chord method without updating of the Jacobian does not converge, see Table 5.3.

Consider now the nonsmooth problem for various values of γ_i and m_i^1. The starting point for the iteration is always $p_j^s = m_i^s = x_{ij}^s = \lambda_i^s = 1$. The results for $m_i^1 = 0.8$ and $\gamma_i = 0.5$ are given in Table 5.4.

If we increase γ_i to 1 the convergence speed in general decreases; the results are in Table 5.5.

Step	$k=1$	$k=2$	$k=3$	$k=5$	$k=100$
0	1.2 E 0	1.2 E 0	1.2 E 0	1.2 E 0	1.2 E 0
1	8.4 E -1	8.4 E -1	8.4 E -1	8.4 E -1	8.4 E -1
2	7.5 E -1	8.0 E -1	8.0 E -1	8.0 E -1	8.0 E -1
3	1.2 E 0	7.6 E -1	7.8 E -1	7.8 E -1	7.8 E -1
4	8.6 E -1	8.5 E -1	8.1 E -1	7.7 E -1	7.7 E -1
8	8.5 E -1	9.1 E -1	1.2 E 0	1.2 E 0	7.6 E -1
13	5.8 E -1	8.6 E -1	1.2 E 0	1.2 E 0	8.2 E -1
23	0	8.6 E -1	1.2 E 0	1.2 E 0	1.2 E -1

Table 5.6: Absolute errors with parameters $m_i^1 = 0.8$ and $\gamma_i = -0.7$.

For negative values of γ_i the model becomes quite unstable. For example if we set $\gamma_i = -0.7$ then for $k = 1$ the method converges after 23 iterations while for $k = 2$ we get a different solution after only 13 iterations and for $k = 3$ we get yet another different solution after 8 iterations. The absolute differences to the solution of the first Newton method are given in Table 5.6.

References

[1] S. Adly, R. Cibulka, H. V. Ngai: Newton's method for solving inclusions using set-valued approximations, SIAM J. Optim. 25 (2015) 159–184.

[2] S. Adly, H. V. Ngai, N. V. Vu: Newton's method for solving generalized equations: Kantorovich's and Smale's approaches, J. Math. Anal. Appl. 439 (2016) 396–418.

[3] F. J. Aragón Artacho, A. Belyakov, A. L. Dontchev, M. Lopez: Local convergence of quasi-Newton methods under metric regularity, Comput. Optim. Appl. 58 (2014) 225–247.

[4] I. K. Argyros: Convergence and Applications of Newton-Type Iterations, Springer, Berlin (2008).

[5] J.-P. Aubin, H. Frankowska: Set-Valued Analysis, Systems and Control: Foundations and Applications, Birkhäuser, Boston (1990).

[6] R. G. Bartle: Newton's method in Banach spaces, Proc. Amer. Math. Soc. 6 (1955) 827–831.

[7] S. C. Billups: Algorithms for Complementarity Problems and Generalized Equations, PhD Thesis, Technical Report 95-14, Computer Sciences Department, University of Wisconsin, Madison (1995).

[8] K. Butts, A. Dontchev, M. Huang, I. Kolmanovsky: A perturbed chord (Newton-Kantorovich) method for constrained nonlinear model predictive control, Proceedings of NOLCOS 2016, IFAC-PapersOnLine 49-18 (2016) 253–258.

[9] P. G. Ciarlet, C. Mardare: On the Newton-Kantorovich theorem, Anal. Appl. (Singapore) 10 (2012) 249–269.

[10] S. P. Dirkse: Robust Solution of Mixed Complementarity Problems, PhD Thesis, Computer Science Departement, University of Wisconsin, Madison (1994).

[11] S. P. Dirkse, M. C. Ferris, T. Munson: http://pages.cs.wisc.edu/~ferris/path.html.

[12] A. L. Dontchev, R. T. Rockafellar: Parametric stability of solutions in models of economic equilibrium, J. Convex Analysis 19 (2012) 975–997.

[13] A. L. Dontchev: Local analysis of a Newton-type method based on partial linearization, in: The Mathematics of Numerical Analysis, (Park City, 1995), Lectures Appl. Math. 32, Amer. Math. Soc., Providence, (1996) 295–306.

[14] A. L. Dontchev, R. T. Rockafellar: Implicit Functions and Solution Mappings. A View from Variational Analysis, 2nd edition, Springer, Berlin (2014).

[15] F. Facchinei, J.-S. Pang: Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer, New York (2003).

[16] O. P. Ferreira, G. N. Silva: Inexact Newton's method to nonlinear functions with values in a cone, preprint, arXiv:1510.01947 (2015).

[17] W. B. Gragg, R. A. Tapia: Optimal error bounds for the Newton-Kantorovich theorem, SIAM J. Numer. Anal. 11 (1974) 10–13.

[18] M. Hintermüller: Semismooth Newton methods and applications, http://www.math.uni-hamburg.de/home/hinze/Psfiles/Hintermueller_OWNotes.pdf, Oberwolfach (2010).

[19] K. Ito, K. Kunisch: Lagrange Multiplier Approach to Variational Problems and Applications, SIAM, Philadelphia (2008).

[20] A. F. Izmailov, A. S. Kurennoy, M. V. Solodov: The Josephy-Newton method for semismooth generalized equations and semismooth SQP for optimization, Set-Valued Var. Anal. 21 (2013) 17–45.

[21] A. F. Izmailov, M. V. Solodov: Newton-Type Methods for Optimization and Variational Problems, Springer, Berlin (2014).

[22] L. V. Kantorovich: On Newton's method for functional equations (Russian), Doklady Akad. Nauk SSSR (N.S.) 59 (1948) 1237–1240.

[23] L. V. Kantorovich, G. P. Akilov: Functional Analysis (Russian), 2nd revised edition, Nauka, Moscow (1977).

[24] C. T. Kelley: Solving Nonlinear Equations with Newton's Method, Fundamentals of Algorithms, SIAM, Philadelphia (2003).

[25] D. Klatte, B. Kummer: Nonsmooth Equations in Optimization. Regularity, Calculus, Methods and Applications, Kluwer, New York (2002).

[26] B. Kummer: Newton's method for non-differentiable functions, in: "Advances in Mathematical Optimization", J. Guddat et al. (eds.), Ser. Math. Res. 45, Akademie-Verlag, Berlin (1988) 114–125.

[27] J. M. Ortega: The Newton-Kantorovich theorem, Amer. Math. Monthly 75 (1968) 658–660.

[28] A. Pietrus: Non differentiable perturbed Newton's method for functions with values in a cone, Investigación Oper. 35 (2014) 58–67.

[29] F. A. Potra, V. Pták: Nondiscrete induction and iterative processes, Research Notes in Mathematics 103, Pitman, Boston (1984).

[30] L. Qi, J. Sun: A nonsmooth version of Newton's method, Math. Programming A 58 (1993) 353–367.

[31] S. M. Robinson: Extension of Newton's method to nonlinear functions with values in a cone, Numer. Math. 19 (1972) 341–347.

[32] G. N. Silva: Kantorovich's theorem on Newton's method for solving generalized equations under the majorant condition, Appl. Math. Comput. 286 (2016) 178–188.

[33] M. Ulbrich: Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces, SIAM, Philadelphia (2011).

[34] T. J. Ypma: Historical development of the Newton-Raphson method, SIAM Rev. 37 (1995) 531–551.

On uniform regularity and strong regularity

R. Cibulka[1], **J. Preininger**[2], **and T. Roubal**[3]

January 31, 2018

Abstract. We investigate uniform versions of (metric) regularity and strong (metric) regularity on compact subsets of Banach spaces, in particular, along continuous paths. These two properties turn out to play a key role in analyzing path-following schemes for tracking a solution trajectory of a parametric generalized equation or, more generally, of a differential generalized equation (DGE). The latter model covers a large territory in control and optimization, such as differential variational inequalities, control systems with constraints, as well as necessary optimality conditions in optimal control. We propose two inexact path-following methods for DGEs having the order of the grid error $O(h)$ and $O(h^2)$, respectively. We provide numerical experiments, comparing the schemes derived, for simple problems arising in physics. Further, we study (metric) regularity of mappings associated with a particular case of the DGE arising in control theory by focusing on the interplay between the pointwise versions of these properties and their infinite-dimensional counterparts.

Key Words. control system, uniform metric regularity, uniform strong metric regularity, discrete approximation, path-following.

AMS Subject Classification (2010) 49k40, 49J40, 49J53, 90c31.

[1]Department of Mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 22, 306 14 Pilsen, Czech Republic, cibi@kma.zcu.cz. Supported by the Czech Science Foundation GA CR, project GA15-00735S.

[2]Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstrasse 8, A-1040 Vienna, preininger.jakob@gmx.at.

[3]Department of Mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 22, 306 14 Pilsen, Czech Republic, roubalt@ntis.zcu.cz. Supported by the Czech Science Foundation GA CR, project GA15-00735S.

1 Introduction

We are going to investigate uniform (metric) regularity and strong (metric) regularity on compact subsets of Banach spaces of mappings which are defined as a sum of a single-valued (possibly non-smooth) mapping and a set-valued mapping with a (locally) closed graph. In the second section, we recall basic definitions from regularity theory and derive a result guaranteeing that a perturbed problem has a solution which is similar to the classical Lyusternik-Graves and Robinson theorem. Conditions ensuring *uniform* [strong] regularity along continuous paths are obtained as a corollary. Roughly speaking, by the word "uniform" we mean that the constants as well as the size of neighborhoods, appearing in the corresponding definitions, remain the same for a certain set of mappings and/or points.

In the third section, we study two (inexact) path-following methods for a *differential generalized equation (DGE)*, a model introduced in [5], which is a problem to find a pair of functions $x : [0, \varepsilon] \to \mathbb{R}^n$ and $u : [0, \varepsilon] \to \mathbb{R}^m$ such that

$$(1) \qquad \begin{cases} \dot{x}(t) &= g(x(t), u(t)), \\ 0 &\in f(x(t), u(t)) + F(u(t)), \qquad \text{for all } t \in [0, \varepsilon], \\ x(0) &= x_{_I}, \end{cases}$$

with a fixed $\varepsilon > 0$, single-valued functions $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, a set-valued mapping $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^d$, and a given initial state $x_{_I} \in \mathbb{R}^n$. This model covers a large territory in control and optimization, such as differential variational inequalities, control systems with constraints, as well as necessary optimality conditions in optimal control (see [5] and references therein). The first scheme, requiring stronger smoothness properties of the solution trajectory of (1), is based on the modified Euler (Euler-Cauchy) method for solving differential equations and is shown to have the grid error of order $O(h^2)$. On the other hand, the latter scheme, based on the Euler method, has the grid error of order $O(h)$ but requires Lipschitz continuity of the solution trajectory only. We provide numerical experiments, comparing the schemes derived and a standard MATLAB function *ODE45*, for two simple problems arising in mechanics and electronics, respectively.

In the fourth section, we study regularity of mappings associated with the problem of *feasibility* in control, which is the problem to find a pair of functions $x : [0, \varepsilon] \to \mathbb{R}^n$ and $u : [0, \varepsilon] \to \mathbb{R}^m$ such that

$$(2) \quad \dot{x}(t) = g(x(t), u(t)) \quad \text{and} \quad f(x(t), u(t)) \in U_{ad} \quad \text{for a.e. } t \in [0, \varepsilon], \quad x(0) = 0,$$

where $\varepsilon > 0$, functions $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and a closed convex subset U_{ad} of $\mathbb{R}^d$ are given. We focus on the interplay between the pointwise conditions and their uniform and infinite-dimensional counterparts. We extend several results from [5].

Basic notation. The *distance* from a point x to a subset A of a metric space (X, ϱ) is $d(x, A) = \inf_{y \in A} \varrho(x, y)$. The closure and the interior of A is denoted by $\operatorname{cl} A$ and $\operatorname{int} A$, respectively. Given sets C, $D \subset X$, the *excess* of C beyond D is defined by $e(C, D) := \sup_{x \in C} d(x, D)$. We use the convention that $\inf \emptyset := +\infty$ and as we work with non-negative quantities we set $\sup \emptyset := 0$. The closed ball centered at a point $x \in X$ with a radius $r > 0$

is denoted by $\mathbb{B}_r(x)$. A set $A \subset X$ is *locally closed* at its point x if there is $r > 0$ such that the set $A \cap \mathbb{B}_r(x)$ is closed. Any singleton set will be identified with its only element, that is, we write a instead of $\{a\}$. Given a (generally set-valued) mapping $F : X \rightrightarrows Y$ between sets X and Y, the *graph*, the *domain*, and the *range* of F are the sets $\operatorname{gph} F := \big\{(x,y) \in X \times Y \,\big|\, y \in F(x)\big\}$, $\operatorname{dom} F := \big\{x \in X \,\big|\, F(x) \neq \emptyset\big\}$, and $\operatorname{rge} F := \big\{y \in Y \,\big|\, \exists x \in X \text{ with } y \in F(x)\big\}$, respectively. The *inverse* of F always exits and is defined as a mapping $Y \ni y \longmapsto F^{-1}(y) := \big\{x \in X \,\big|\, y \in F(x)\big\}$. We write $f : X \to Y$ to emphasize that the mapping f is single-valued. The space of all linear bounded (single-valued) mappings acting from a Banach space X into another Banach space Y equipped with the standard operator norm is denoted by $\mathcal{L}(X,Y)$. The space $\mathbb{R}^n$ is equipped with the Euclidean norm, while the Cartesian product of two or more spaces is considered with the box (product) topology. By a.e. we mean almost every in terms of the Lebesgue measure.

2 Uniform regularity

In our analysis, we employ two key concepts from set-valued analysis called regularity and strong regularity of a set-valued mapping. Let us emphasize that unlike definitions in [13], we prefer not to include the assumption that the mapping under consideration has a locally closed graph in any definition of regularity. Given metric spaces (X,ϱ), (Y,ϱ) and a non-empty subset $U \times V$ of $X \times Y$, a mapping $F : X \rightrightarrows Y$ is said to be *regular on U for V* if there is a constant $\kappa > 0$ such that

$$d\big(x, F^{-1}(y)\big) \leq \kappa\, d(y, F(x) \cap V) \quad \text{for every} \quad (x,y) \in U \times V.$$

If $U = X$ and $V = Y$ then F is said to be *globally regular*. Given $(\bar{x},\bar{y}) \in X \times Y$, the mapping F is said to be *regular at $\bar{x}$ for $\bar{y}$* if $(\bar{x},\bar{y}) \in \operatorname{gph} F$ and there are positive constants a, b, and κ such that

$$d\big(x, F^{-1}(y)\big) \leq \kappa\, d\big(y, F(x)\big) \quad \text{for each} \quad (x,y) \in \mathbb{B}_a(\bar{x}) \times \mathbb{B}_b(\bar{y}).$$

The infimum of $\kappa > 0$ such that the above inequality holds for some $a > 0$ and $b > 0$ is the *regularity modulus* of F at $\bar{x}$ for $\bar{y}$ and is denoted by $\operatorname{reg}(F; \bar{x} \,|\, \bar{y})$. Clearly, if F is regular at $\bar{x}$ for $\bar{y}$ with a constant κ and neighborhoods $\mathbb{B}_a(\bar{x})$ and $\mathbb{B}_b(\bar{y})$, then F is regular on $\mathbb{B}_a(\bar{x})$ for $\mathbb{B}_b(\bar{y})$ with the same constant. On the other hand, when the sets U and V are neighborhoods of points $\bar{x}$ and $\bar{y}$, respectively, and $\bar{y} \in F(\bar{x})$, then regularity of F on U for V implies regularity of F at $\bar{x}$ for $\bar{y}$. The constants are the same again but neighborhoods may differ [13, Proposition 5H.1]. By the Banach open mapping principle, a mapping $A \in \mathcal{L}(X,Y)$ is globally regular if and only if it is surjective. A mapping $F : X \rightrightarrows Y$ is said to be *strongly regular on U for V* if there is a constant $\kappa > 0$ such that the mapping $\sigma : V \ni y \longmapsto F^{-1}(y) \cap U$ is both single-valued and Lipschitz continuous on $V = \operatorname{dom}\sigma$ with the constant κ. The mapping F is said to be *strongly regular* at $\bar{x}$ for $\bar{y}$ if $\bar{y} \in F(\bar{x})$ and there are neighborhoods U of $\bar{y}$ and V of $\bar{x}$ such that F is strongly regular on U for V.

First, we present a statement concerning *perturbed [strong] regularity on a set*.

Theorem 2.1. *Let $(X, \|\cdot\|)$ and $(Y, \|\cdot\|)$ be Banach spaces, let $G : X \rightrightarrows Y$ be a set-valued mapping, and $(\bar{x}, \bar{y}) \in X \times Y$. Assume that there are positive constants a, b, and κ such that the set $\operatorname{gph} G \cap \big(I\!B_a(\bar{x}) \times I\!B_b(\bar{y})\big)$ is closed in $X \times Y$ and G is [strongly] regular on $I\!B_a(\bar{x})$ for $I\!B_b(\bar{y})$ with the constant κ. Let $\mu > 0$ be such that $\kappa\mu < 1$ and let $\kappa' > \kappa/(1-\kappa\mu)$. Then for every positive α and β such that*

$$2\kappa'\beta + \alpha \le a \quad \text{and} \quad \mu(2\kappa'\beta + \alpha) + 2\beta \le b \tag{3}$$

and for every mapping $g : X \to Y$ satisfying

$$\|g(\bar{x})\| \le \beta \quad \text{and} \quad \|g(x) - g(x')\| \le \mu\|x - x'\| \quad \text{for every} \quad x, x' \in I\!B_{2\kappa'\beta+\alpha}(\bar{x}), \tag{4}$$

the mapping $g + G$ has the following property: for every y, $y' \in I\!B_\beta(\bar{y})$ and every $x \in (g+G)^{-1}(y) \cap I\!B_\alpha(\bar{x})$ there exists a [unique] point $x' \in I\!B_{2\kappa'\beta+\alpha}(\bar{x})$ such that

$$y' \in g(x') + G(x') \quad \text{and} \quad \|x - x'\| \le \kappa'\|y - y'\|. \tag{5}$$

Proof. We shall imitate the proof of [13, Theorem 5G.3]. First, suppose that G is regular on $I\!B_a(\bar{x})$ for $I\!B_b(\bar{y})$ with the constant κ. Choose any α and β, and then any g as in the statement. Then

$$y - g(x) \in I\!B_b(\bar{y}) \quad \text{for each} \quad (x, y) \in I\!B_{2\kappa'\beta+\alpha}(\bar{x}) \times I\!B_\beta(\bar{y}). \tag{6}$$

Indeed, fix any such a pair (x, y). Then (4) and (3) imply that

$$\begin{aligned} \|y - g(x) - \bar{y}\| &\le \|g(\bar{x})\| + \|g(\bar{x}) - g(x)\| + \|y - \bar{y}\| \le \beta + \mu\|x - \bar{x}\| + \beta \\ &\le 2\beta + \mu(2\kappa'\beta + \alpha) \le b. \end{aligned}$$

Fix any two distinct y, $y' \in I\!B_\beta(\bar{y})$ and any $x \in (g+G)^{-1}(y) \cap I\!B_\alpha(\bar{x})$. Let $r := \kappa'\|y - y'\|$. As $r \le 2\kappa'\beta$, the first inequality in (3) implies that

$$I\!B_r(x) \subset I\!B_{2\kappa'\beta+\alpha}(\bar{x}) \subset I\!B_a(\bar{x}).$$

Consider the mapping

$$X \ni u \longmapsto \Phi(u) = \Phi_{y'}(u) := G^{-1}\big(y' - g(u)\big) \subset X.$$

It suffices to show that there is a fixed point x' of Φ in $I\!B_r(x)$, because then $x' \in (g+G)^{-1}(y')$ and the desired distance estimate holds. To obtain such a point x' we are going to apply [13, Theorem 5E.2]. The set $\Omega := \operatorname{gph}\Phi \cap (I\!B_r(x) \times I\!B_r(x))$ is closed. Indeed, pick any sequence (x_n, z_n) in Ω converging to a point $(\tilde{x}, \tilde{z}) \in X \times X$. Clearly, $(\tilde{x}, \tilde{z}) \in I\!B_r(x) \times I\!B_r(x)$. The definition of Φ and (6) imply that

$$\big(z_n, y' - g(x_n)\big) \in \operatorname{gph} G \cap (I\!B_r(x) \times I\!B_b(\bar{y})) \subset \operatorname{gph} G \cap (I\!B_a(\bar{x}) \times I\!B_b(\bar{y})) \quad \text{for each} \quad n \in \mathbb{N}.$$

Passing to the limit we get that $\big(\tilde{z}, y' - g(\tilde{x})\big) \in \operatorname{gph} G$, that is, $(\tilde{x}, \tilde{z}) \in \operatorname{gph}\Phi$.

According to (6) we have $y - g(x) \in G(x) \cap I\!B_b(\bar{y})$ and $y' - g(x) \in I\!B_b(\bar{y})$, thus regularity of G on $I\!B_a(\bar{x})$ for $I\!B_b(\bar{y})$ yields that

$$\begin{aligned} d(x, \Phi(x)) &= d\big(x, G^{-1}(y' - g(x))\big) \le \kappa\, d\big(y' - g(x), G(x) \cap I\!B_b(\bar{y})\big) \le \kappa\|y - y'\| \\ &< \kappa'\|y - y'\|(1 - \kappa\mu) = r(1 - \kappa\mu). \end{aligned}$$

Let $u,\, v \in I\!B_r(x)$ be arbitrary. Pick an arbitrary $w \in \Phi(u) \cap I\!B_r(x)$ (if there is any). As $y' - g(u) \in G(w) \cap I\!B_b(\bar{y})$ and $y' - g(v) \in I\!B_b(\bar{y})$, we get

$$d(w, \Phi(v)) \;\; = \;\; d\big(w, G^{-1}(y' - g(v))\big) \le \kappa\, d\big(y' - g(v), G(w) \cap I\!B_b(\bar{y})\big) \le \kappa \|g(u) - g(v)\|.$$

This means that

$$e\big(\Phi(u) \cap I\!B_r(x), \Phi(v)\big) \;\; \le \kappa \|g(u) - g(v)\| \le \kappa\mu\, \|u - v\| \quad \text{whenever} \quad u, v \in I\!B_r(x).$$

The assumptions of [13, Theorem 5E.2] are verified. The existence of $x' \in I\!B_{2\kappa'\beta+\alpha}(\bar{x})$ satisfying (5) is established.

Now, let G be strongly regular on $I\!B_a(\bar{x})$ for $I\!B_b(\bar{y})$ with the constant κ. To prove the uniqueness, it is enough to show that the mapping $I\!B_\beta(\bar{y}) \ni y \longmapsto \sigma(y) := (g + G)^{-1}(y) \cap I\!B_{2\kappa'\beta+\alpha}(\bar{x})$ is nowhere multivalued. Assume on the contrary that for some $y \in I\!B_\beta(\bar{y})$ there are two distinct points $x_1,\, x_2 \in \sigma(y)$. Clearly, $x_1 \in G^{-1}(y - g(x_1)) \cap I\!B_a(\bar{x})$ and $x_2 \in G^{-1}(y - g(x_2)) \cap I\!B_a(\bar{x})$. By (6), the points $y - g(x_1)$ and $y - g(x_2)$ are in $I\!B_b(\bar{y})$. Hence $0 < \|x_1 - x_2\| \le \kappa \|g(x_1) - g(x_2)\| \le \kappa\mu \|x_1 - x_2\| < \|x_1 - x_2\|$, a contradiction. □

If, in addition to the assumptions of Theorem 2.1, we have $(\bar{x}, \bar{y}) \in \operatorname{gph} G$, then we arrive at [9, Theorem 2.3] which is a slight improvement [13, Theorem 5G.3], where it is supposed that G is regular at $\bar{x}$ for $\bar{y}$ with the constant κ and neighborhoods $I\!B_a(\bar{x})$ and $I\!B_b(\bar{y})$.

Remark 2.2. Under the strong regularity, the reasoning used at the end of the proof of Theorem 2.1 implies that the function σ, defined therein, is Lipschitz continuous relative to $\operatorname{dom} \sigma \subset I\!B_\beta(\bar{y})$ with the constant κ'. If, in addition,

$$\big(I\!B_\alpha(\bar{x}) \times I\!B_\beta(\bar{y})\big) \cap \operatorname{gph}(g + G) \ne \emptyset, \tag{7}$$

then $\operatorname{dom} \sigma = I\!B_\beta(\bar{y})$ and consequently $g + G$ is strongly regular on $I\!B_{2\kappa'\beta+\alpha}(\bar{x})$ for $I\!B_\beta(\bar{y})$. Note that (7) holds, for example, when $(\bar{x}, \bar{y}) \in \operatorname{gph} G$.

We also get the following uniformity result.

Corollary 2.3. *Under assumptions of Theorem 2.1, let $\gamma \in [0, \alpha)$, $\delta \in [0, \beta)$, and $(x, y) \in I\!B_\gamma(\bar{x}) \times I\!B_\delta(\bar{y})$ be arbitrary. Then the mapping $g + G$ is regular on $I\!B_{\alpha-\gamma}(x)$ for $I\!B_{\beta-\delta}(y)$ with the constant κ'.*

Proof. Let constants γ and δ along with a pair (x, y) be as in the premise. Set $U := I\!B_{\alpha-\gamma}(x)$ and $V := I\!B_{\beta-\delta}(y)$. We have to show that

$$d\big(u, (g + G)^{-1}(v)\big) \le \kappa'\, d(v, (g + G)(u) \cap V) \quad \text{for every} \quad (u, v) \in U \times V.$$

Fix any such a pair (u, v). Pick an arbitrary $v' \in (g+G)(u) \cap V$ (if there is any). Noting that $U \times V \subset I\!B_\alpha(\bar{x}) \times I\!B_\beta(\bar{y})$, Theorem 2.1 yields $u' \in (g + G)^{-1}(v)$ with $\|u - u'\| \le \kappa' \|v - v'\|$. Hence $d\big(u, (g+G)^{-1}(v)\big) \le \|u - u'\| \le \kappa' \|v - v'\|$. As $v' \in (g + G)(u) \cap V$ was arbitrary, the proof is finished. □

We show now that the regularity at each point of a compact set implies *uniform* regularity, that is, we can choose the same constant and neighborhoods for all points in this set.

Theorem 2.4. *Let (P,ϱ) be a metric space, let $(X,\|\cdot\|)$ and $(Y,\|\cdot\|)$ be Banach spaces, and let Ω be a compact subset of $P\times X$. Consider a set-valued mapping $F:X\rightrightarrows Y$ and a mapping $\sigma:P\times X\to Y$ such that*

(i) for each $z=(p,x)\in\Omega$ the mapping $X\ni v\longmapsto \mathcal{G}_p(v):=\sigma(p,v)+F(v)\subset Y$ has a locally closed graph at $(x,0)$ and is [strongly] regular at x for 0;

(ii) for each $z=(p,x)\in\Omega$ and each $\mu>0$ there is $\delta>0$ such that for each v, $v'\in I\!B_\delta(x)$ and each $p'\in I\!B_\delta(p)$ we have

$$\|[\sigma(p',v')-\sigma(p,v')]-[\sigma(p',v)-\sigma(p,v)]\|\le\mu\|v-v'\|;$$

(iii) for each $x\in X$ the function $\sigma(\cdot,x)$ is continuous.

Then there are positive constants a, b, and κ such that for each $z=(p,x)\in\Omega$ the mapping $\mathcal{G}_p$ is [strongly] regular at x for 0 with the constant κ and neighborhoods $I\!B_a(x)$ and $I\!B_b(0)$.

Proof. Fix any $\bar z=(\bar p,\bar x)\in\Omega$. Using (i), we find positive constants $a_{\bar z}$, $b_{\bar z}$, and $\kappa_{\bar z}$ such that the set $\operatorname{gph}\mathcal{G}_{\bar p}\cap(I\!B_{a_{\bar z}}(\bar x)\times I\!B_{b_{\bar z}}(0))$ is closed in $X\times Y$ and $\mathcal{G}_{\bar p}$ is regular on $I\!B_{a_{\bar z}}(\bar x)$ for $I\!B_{b_{\bar z}}(0)$ with the constant $\kappa_{\bar z}$. Let $\mu_{\bar z}:=1/(2\kappa_{\bar z})$ and $\kappa'_{\bar z}:=3\kappa_{\bar z}$. Then $\kappa_{\bar z}\mu_{\bar z}<1$ and $\kappa'_{\bar z}>2\kappa_{\bar z}=\kappa_{\bar z}/(1-\kappa_{\bar z}\mu_{\bar z})$. In view of (ii), there is $\alpha_{\bar z}\in(0,\min\{a_{\bar z}/2,3\kappa_{\bar z}b_{\bar z}/4\})$ such that for each v, $v'\in I\!B_{2\alpha_{\bar z}}(\bar x)$ and each $p\in I\!B_{\alpha_{\bar z}}(\bar p)$ we have

$$\|[\sigma(p,v)-\sigma(\bar p,v)]-[\sigma(p,v')-\sigma(\bar p,v')]\|\le\mu_{\bar z}\|v-v'\|. \tag{8}$$

Let $\beta_{\bar z}:=\alpha_{\bar z}/(2\kappa'_{\bar z})$. Then

$$2\kappa'_{\bar z}\beta_{\bar z}+\alpha_{\bar z}=2\alpha_{\bar z}<a_{\bar z}\quad\text{and}\quad\mu_{\bar z}(2\kappa'_{\bar z}\beta_{\bar z}+\alpha_{\bar z})+2\beta_{\bar z}=\frac{\alpha_{\bar z}}{\kappa_{\bar z}}+\frac{\alpha_{\bar z}}{3\kappa_{\bar z}}=\frac{4\alpha_{\bar z}}{3\kappa_{\bar z}}<b_{\bar z}. \tag{9}$$

Now, (iii) implies that there is $r_{\bar z}\in(0,\alpha_{\bar z}/2)$ such that

$$\|\sigma(p,\bar x)-\sigma(\bar p,\bar x)\|\le\beta_{\bar z}\quad\text{for all}\quad p\in I\!B_{r_{\bar z}}(\bar p). \tag{10}$$

Pick any $z=(p,x)\in\big(\operatorname{int}I\!B_{r_{\bar z}}(\bar p)\times\operatorname{int}I\!B_{r_{\bar z}}(\bar x)\big)\cap\Omega$. Define a mapping $g_{p,\bar p}:X\to Y$ by

$$g_{p,\bar p}(v):=\sigma(p,v)-\sigma(\bar p,v),\quad v\in X.$$

Then $\mathcal{G}_p=\mathcal{G}_{\bar p}+g_{p,\bar p}$. By (8), for any v, $v'\in I\!B_{2\alpha_{\bar z}}(\bar x)$ we have

$$\|g_{p,\bar p}(v)-g_{p,\bar p}(v')\|\le\mu_{\bar z}\|v-v'\|.$$

Using (10) we get $\|g_{p,\bar p}(\bar x)\|\le\beta_{\bar z}$. Applying Theorem 2.1 we conclude that the following `claim` holds: *for every y, $y'\in I\!B_{\beta_{\bar z}}(0)$ and every $v\in\mathcal{G}_p^{-1}(y')\cap I\!B_{\alpha_{\bar z}}(\bar x)$ there exists $v'\in\mathcal{G}_p^{-1}(y)$ such that $\|v-v'\|\le\kappa'_{\bar z}\,\|y-y'\|$.*

As $z\in\Omega$, we have $0\in\mathcal{G}_p(x)$. We show next that

$$d(v,\mathcal{G}_p^{-1}(y))\le\kappa'_{\bar z}\,d(y,\mathcal{G}_p(v))\quad\text{for all}\quad(v,y)\in I\!B_{\kappa'_{\bar z}\beta_{\bar z}/3}(x)\times I\!B_{\beta_{\bar z}/3}(0). \tag{11}$$

To see this fix any such a pair (v, y). Pick an arbitrary $y' \in \mathcal{G}_p(v)$ (if there is any). The choice of $\beta_{\bar{z}}$ and $r_{\bar{z}}$ implies that

$$\mathbb{B}_{\kappa'_{\bar{z}}\beta_{\bar{z}}}(x) = \mathbb{B}_{\alpha_{\bar{z}}/2}(x) \subset \mathbb{B}_{\alpha_{\bar{z}}}(\bar{x}).$$

First, assume that $\|y'\| \leq \beta_{\bar{z}}$. The claim yields $v' \in \mathcal{G}_p^{-1}(y)$ with $\|v - v'\| \leq \kappa'_{\bar{z}} \|y - y'\|$. Consequently,

$$d(v, \mathcal{G}_p^{-1}(y)) \;\; \leq \;\; \|v - v'\| \leq \kappa'_{\bar{z}} \|y - y'\|.$$

On the other hand, assuming that $\|y'\| > \beta_{\bar{z}}$, we have $\|y' - y\| > \beta_{\bar{z}} - \beta_{\bar{z}}/3 = 2\beta_{\bar{z}}/3$. Then using the claim, with $(y', v) := (0, x)$, we find $v' \in \mathcal{G}_p^{-1}(y)$ such that $\|x - v'\| \leq \kappa'_{\bar{z}} \|y\|$. Consequently,

$$\begin{array}{rcl} d(v, \mathcal{G}_p^{-1}(y)) & \leq & \|v - x\| + d(x, \mathcal{G}_p^{-1}(y)) \leq \|v - x\| + \|x - v'\| \leq \|v - x\| + \kappa'_{\bar{z}} \|y\| \\ & \leq & \kappa'_{\bar{z}}\beta_{\bar{z}}/3 + \kappa'_{\bar{z}}\beta_{\bar{z}}/3 = 2\kappa'_{\bar{z}}\beta_{\bar{z}}/3 < \kappa'_{\bar{z}}\|y - y'\|. \end{array}$$

We have shown that $d(v, \mathcal{G}_p^{-1}(y)) \leq \kappa'_{\bar{z}} \|y - y'\|$ for any $y' \in \mathcal{G}_p(v)$, which proves (11).

Summarizing, for each $z = (p, x) \in \big(\mathrm{int}\mathbb{B}_{r_{\bar{z}}}(\bar{p}) \times \mathrm{int}\mathbb{B}_{r_{\bar{z}}}(\bar{x})\big) \cap \Omega$ the mapping $\mathcal{G}_p$ is regular at x for 0 with the constant $\kappa'_{\bar{z}}$ and neighborhoods $\mathbb{B}_{\kappa'_{\bar{z}}\beta_{\bar{z}}/3}(x)$ and $\mathbb{B}_{\beta_{\bar{z}}/3}(0)$, that is, the size of neighborhoods and the constant of regularity are independent of z in a vicinity of $\bar{z}$. From the open covering $\cup_{\bar{z}=(\bar{p},\bar{x})\in\Omega}\big([\mathrm{int}\mathbb{B}_{r_{\bar{z}}}(\bar{p}) \times \mathrm{int}\mathbb{B}_{r_{\bar{z}}}(\bar{x})] \cap \Omega\big)$ of Ω choose a finite subcovering $\mathcal{O}_i := [\mathrm{int}\mathbb{B}_{r_{\bar{z}_i}}(\bar{p}_i) \times \mathrm{int}\mathbb{B}_{r_{\bar{z}_i}}(\bar{x}_i)] \cap \Omega$, $i = 1, 2, \ldots, k$. Let $a = \min\{\kappa'_{\bar{z}_i}\beta_{\bar{z}_i}/3 \mid i = 1, \ldots, k\}$, $\kappa = \max\{\kappa'_{\bar{z}_i} \mid i = 1, \ldots, k\}$, and $b = \min\{\beta_{\bar{z}_i}/3 \mid i = 1, \ldots, k\}$. For any $z = (p, x) \in \Omega$ there is an index $i \in \{1, \ldots, k\}$ such that $z \in \mathcal{O}_i$. Hence the mapping $\mathcal{G}_p$ is regular at x for 0 with the constant κ and neighborhoods $\mathbb{B}_a(x)$ and $\mathbb{B}_b(0)$.

Under the assumption of strong regularity one uses Remark 2.2 (or the strong regularity part of Theorem 5G.3 in [13]). □

Remark 2.5. Note that (ii) in Theorem 2.4 is satisfied, in particular, when σ has a *point-based approximation on* Ω in the sense of Robinson [17]. Theorem 2.4 yields [9, Lemma 0]. Moreover, given a non-empty subset Ω of a metric space, define the *measure of non-compactness* of Ω by

$$\chi(\Omega) := \inf\{r > 0 \,\big|\, \Omega \subset \Omega_{\mathcal{F}} + \mathbb{B}_r(0) \text{ for some finite subset } \Omega_{\mathcal{F}} \text{ of } \Omega\}.$$

Then Theorem 2.4 holds provided that $\chi(\Omega)$ is strictly smaller than the infimum of the reciprocal values of the regularity moduli of the mappings appearing in (i). This statement is a key element in the proof of the non-smooth versions of Robinson and Lyusternik-Graves theorems, cf. [7, Step 1] and [8, Lemma 12].

Next statement guarantees uniform [strong] regularity along continuous paths.

Theorem 2.6. *Let (T, ϱ) be a compact metric space, let $(X, \|\cdot\|)$ and $(Y, \|\cdot\|)$ be Banach spaces. Consider a set-valued mapping $F : X \rightrightarrows Y$ with closed graph, a mapping $\sigma : T \times X \to Y$, and two continuous mappings $\varphi : T \to X$ and $\psi : T \to Y$ such that*

(i) for each $t \in T$ the mapping $X \ni v \longmapsto \mathcal{G}_t(v) := \sigma(t,v) + F(v) \subset Y$ is [strongly] regular at $\varphi(t)$ for $\psi(t)$;

(ii) for each $t \in T$ and each $\mu > 0$ there is $\delta > 0$ such that for each v, $v' \in I\!B_\delta(\varphi(t))$ and each $t' \in I\!B_\delta(t)$ we have

$$\|[\sigma(t',v') - \sigma(t,v')] - [\sigma(t',v) - \sigma(t,v)]\| \le \mu\|v - v'\|;$$

(iii) for each $x \in X$ the function $\sigma(\cdot, x)$ is continuous.

Then there are positive constants a, b, and κ such that for each $t \in T$ the mapping $\mathcal{G}_t$ is [strongly] regular at $\varphi(t)$ for $\psi(t)$ with the constant κ and neighborhoods $I\!B_a(\varphi(t))$ and $I\!B_b(\psi(t))$.

Proof. Apply Theorem 2.4 with $P := T \times Y$, a (compact) set $\Omega := \bigcup_{t\in T}(t, \psi(t), \varphi(t))$, and $\sigma(p,x) := \sigma(t,x) - y$, $p = (t,y) \in P$, $x \in X$. □

3 Path-following for differential generalized equations

Consider the DGE (1), with $\varepsilon > 0$, single-valued functions $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$, a set-valued mapping $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^d$, and an initial state $x_{_I} \in \mathbb{R}^n$. If it is not clearly indicated otherwise we impose the following:

Standing assumptions (SA)**.** Consider the DGE (1) and suppose that f and g are differentiable functions with a locally Lipschitz continuous derivative, and that F has a closed graph. Further, let a pair of functions $(\bar{x}(\cdot), \bar{u}(\cdot))$ be a solution of (1) such that both of them are differentiable on $[0, \varepsilon]$ and have a Lipschitz continuous derivative on this interval.

For an integer $N > 1$, consider the uniform grid $t_i := ih$, $i \in \{0, 1, \dots, N\}$, with a step size $h := \varepsilon/N$. Given $\Delta > 0$ and points $(e_i)_{i=0}^{N-1}$ in $I\!B_{\Delta h^2}(0)$, consider the following iteration

$$(12) \qquad \begin{cases} \widetilde{x}_{i+1} &= x_i + hg(x_i, u_i), \\ e_i &\in f(\widetilde{x}_{i+1}, u_i) + \nabla_u f(\widetilde{x}_{i+1}, u_i)(u_{i+1} - u_i) + F(u_{i+1}), \\ x_{i+1} &= x_i + \dfrac{h}{2}(g(x_i, u_i) + g(\widetilde{x}_{i+1}, u_{i+1})), \end{cases}$$

with (x_0, u_0) sufficiently close to $(\bar{x}(0), \bar{u}(0))$. The reason for allowing $x_0 \neq x_I$ is that for a given time interval $I := [-\varepsilon, \varepsilon]$, say, one cannot expect that $\bar{u}(\cdot)$ is differentiable on the whole of I in general (for example when a geometric constraint represented by the generalized equation is a variational inequality). However, $\bar{u}(\cdot)$ can be piece-wise smooth on I and the starting point x_0 can be viewed as a final iterate obtained by a numerical algorithm on the previous subinterval $[-\varepsilon, 0]$. One can consider more general Runge-Kutta approximations as in [11] but we prefer to keep the presentation as clear as possible. We use a modification of the classical trapezoidal rule [10] in our analysis.

Lemma 3.1. *Let $\varphi : [a,b] \to \mathbb{R}$ be a function with a Lipschitz continuous derivative on $[a,b]$. Then there is a constant $m > 0$ such that for each t_1, $t_2 \in [a,b]$, with $t_1 < t_2$, we have*

$$\left| \frac{(t_2 - t_1)}{2} \big(\varphi(t_1) + \varphi(t_2)\big) - \int_{t_1}^{t_2} \varphi(t)\,\mathrm{d}t \right| \le m(t_2 - t_1)^3.$$

Proof. Let $\ell > 0$ be a Lipschitz constant of $\dot\varphi$ on $[a,b]$. Fix arbitrary t_1, $t_2 \in [a,b]$ with $t_1 < t_2$ and let $h := t_2 - t_1$. Find τ_1 and τ_2 in $[t_1, t_2]$ such that $\dot\varphi(\tau_1) = \min_{\tau\in[t_1,t_2]} \dot\varphi(\tau)$ and $\dot\varphi(\tau_2) = \max_{\tau\in[t_1,t_2]} \dot\varphi(\tau)$. Consider a function $\psi : [t_1, t_2] \to \mathbb{R}$ defined by

$$\psi(t) := \varphi(t) - \frac{\dot\varphi(\tau_1) + \dot\varphi(\tau_2)}{2} t, \quad t \in [t_1, t_2].$$

For each $t \in [t_1, t_2]$, we have $\dot\varphi(\tau_1) \le \dot\varphi(t) \le \dot\varphi(\tau_2)$, and consequently

$$-\frac{\ell h}{2} \le -\frac{\ell}{2}|\tau_1 - \tau_2| \le \frac{1}{2}(\dot\varphi(\tau_1) - \dot\varphi(\tau_2)) \le \dot\psi(t) \le \frac{1}{2}(\dot\varphi(\tau_2) - \dot\varphi(\tau_1)) \le \frac{\ell}{2}|\tau_2 - \tau_1| \le \frac{\ell h}{2}.$$

Thus $\max_{\tau\in[t_1,t_2]} |\dot\psi(\tau)| \le \ell h/2$. Basic calculus and the mean value theorem imply that

$$\begin{aligned}
\left| \frac{h}{2}\big(\varphi(t_1) + \varphi(t_2)\big) - \int_{t_1}^{t_2} \varphi(t)\,\mathrm{d}t \right| &= \left| \frac{h}{2}\big(\psi(t_1) + \psi(t_2)\big) - \int_{t_1}^{t_2} \psi(t)\,\mathrm{d}t \right| \\
&= \left| \int_{t_1}^{t_1+\frac{h}{2}} [\psi(t_1) - \psi(t)]\,\mathrm{d}t + \int_{t_1+\frac{h}{2}}^{t_2} [\psi(t_2) - \psi(t)]\,\mathrm{d}t \right| \\
&\le \max_{\tau\in[t_1,t_2]} |\dot\psi(\tau)| \left(\int_{t_1}^{t_1+\frac{h}{2}} (t - t_1)\,\mathrm{d}t + \int_{t_1+\frac{h}{2}}^{t_2} (t_2 - t)\,\mathrm{d}t \right) \\
&= \max_{\tau\in[t_1,t_2]} |\dot\psi(\tau)| \left(\frac{h^2}{8} + \frac{h^2}{8} \right) \le \frac{\ell}{8} h^3.
\end{aligned}$$

As ℓ is independent of both t_1 and t_2, setting $m := \ell/8$ we finish the proof. □

Theorem 3.2. *In addition to (SA), suppose that for each $t \in [0,\varepsilon]$ the mapping*

$$\mathbb{R}^m \ni v \longmapsto \mathcal{G}_t(v) := f(\bar x(t), \bar u(t)) + \nabla_u f(\bar x(t), \bar u(t))(v - \bar u(t)) + F(v) \subset \mathbb{R}^d \tag{13}$$

is [strongly] regular at $\bar u(t)$ for 0. Then for any $\Delta > 0$ there are $N_0 \in \mathbb{N}$ and positive constants α and $\bar d$ such that for each $N > N_0$, each $(x_0, u_0) \in I\!B_{\Delta h^2}(\bar x(0)) \times I\!B_{\Delta h^2}(\bar u(0))$, and each $(e_i)_{i=0}^{N-1}$ in $I\!B_{\Delta h^2}(0)$, where $h := \varepsilon/N$, there are [uniquely determined] points $(x_i, u_i) \in \mathbb{R}^n \times \mathbb{R}^m$, $i \in \{1, \dots, N\}$, generated by the iteration (12), with the initial point (x_0, u_0), such that $(x_i, u_i) \in I\!B_\alpha(\bar x(t_i)) \times I\!B_\alpha(\bar u(t_i))$ for each $i \in \{1, \dots, N\}$ satisfying

$$\max_{0\le i\le N} \|x_i - \bar x(t_i)\| \le \bar d h^2 \quad \text{and} \quad \max_{0\le i\le N} \|u_i - \bar u(t_i)\| \le \bar d h^2. \tag{14}$$

Proof. Let a (continuous) function σ be defined by $\sigma(t,v) := f(\bar x(t), \bar u(t)) + \nabla_u f(\bar x(t), \bar u(t))(v - \bar u(t))$, $(t,v) \in [0,\varepsilon] \times \mathbb{R}^m$. For each $t \in [0,\varepsilon]$ and each $\mu > 0$, the continuity of the function $s \longmapsto \nabla_u f(\bar x(s), \bar u(s))$ at t yields a constant $\delta > 0$ such that

$$\|\nabla_u f(\bar x(t'), \bar u(t')) - \nabla_u f(\bar x(t), \bar u(t))\| < \mu \quad \text{whenever} \quad t' \in (t - \delta, t + \delta) \cap [0,\varepsilon],$$

consequently, for any such t' and arbitrary v, $v' \in \mathbb{R}^m$ we have

$$\begin{aligned}\|[\sigma(t',v') - \sigma(t,v')] - [\sigma(t',v) - \sigma(t,v)]\| &= \|[\nabla_u f(\bar{x}(t'),\bar{u}(t')) - \nabla_u f(\bar{x}(t),\bar{u}(t))](v'-v)\| \\ &\le \mu\|v - v'\|.\end{aligned}$$

Theorem 2.6 with $T := [0,\varepsilon]$, $\varphi := \bar{u}(\cdot)$, $\psi \equiv 0$ yields positive constants a, b, and κ such that for each $t \in [0,\varepsilon]$ the mapping $\mathcal{G}_t$ is [strongly] regular at $\bar{u}(t)$ for 0 with the constant κ and neighborhoods $I\!B_a(\bar{u}(t))$ and $I\!B_b(0)$. Find $\ell_1 > 0$ such that both $\bar{x}(\cdot)$ and $\bar{u}(\cdot)$ are Lipschitz continuous on $[0,\varepsilon]$ with the constant ℓ_1. Let $r > 0$ be such that $\bar{x}([0,\varepsilon]) + aI\!B_{\mathbb{R}^n} \subset rI\!B_{\mathbb{R}^n}$ and $\bar{u}([0,\varepsilon]) + aI\!B_{\mathbb{R}^m} \subset rI\!B_{\mathbb{R}^m}$. Pick $\ell_2 > 0$ such that f, g, and $\nabla_u f$ are Lipschitz continuous on the (compact) set $rI\!B_{\mathbb{R}^n} \times rI\!B_{\mathbb{R}^m}$. Let

$$\kappa' := 2\kappa, \quad \mu := 1/(3\kappa), \quad \text{and} \quad \ell := \max\{1,\ell_1,\ell_2\}. \tag{15}$$

By the basic calculus, for every u, $u' \in rI\!B_{\mathbb{R}^m}$ and every $x \in rI\!B_{\mathbb{R}^n}$, we have

$$\|f(x,u') - f(x,u) - \nabla_u f(x,u)(u'-u)\| \le \frac{\ell}{2}\|u'-u\|^2. \tag{16}$$

Let

$$\alpha := \min\{1, a/2, 1/(6\ell\kappa), a/(16\kappa\ell), 3\kappa b/(20\kappa\ell+1)\} \quad \text{and} \quad \beta := 2\ell\alpha. \tag{17}$$

We show the following `claim`: *For any $(t,u,x,y) \in [0,\varepsilon] \times I\!B_\alpha(\bar{u}(t)) \times I\!B_\alpha(\bar{x}(t)) \times I\!B_\beta(0)$, there is a [unique] point $w \in I\!B_\alpha(\bar{u}(t))$ such that $y \in f(x,u) + \nabla_u f(x,u)(w-u) + F(w)$ and*

$$\|w - \bar{u}(t)\| \le \kappa'\ell(\|x - \bar{x}(t)\| + \|u - \bar{u}(t)\|^2 + \|y\|).$$

To prove this, fix any such (t,u,x,y) and consider a function $\varphi : \mathbb{R}^m \to \mathbb{R}^d$ defined by

$$\varphi(v) := f(x,u) + \nabla_u f(x,u)(v-u) - f(\bar{x}(t),\bar{u}(t)) - \nabla_u f(\bar{x}(t),\bar{u}(t))(v - \bar{u}(t)), \quad v \in \mathbb{R}^m.$$

We are going to use Theorem 2.1 (with $G := \mathcal{G}_t$ and $g := \varphi$). Note that $\mathcal{G}_t$ has closed graph. Clearly (15) implies $\kappa\mu < 1$ and $\kappa' > 3\kappa/2 = \kappa/(1-\mu\kappa)$. We also get

$$2\kappa'\beta + \alpha = (8\kappa\ell)\alpha + \alpha \le a/2 + a/2 = a,$$

and, consequently, we obtain that

$$\mu(2\kappa'\beta + \alpha) + 2\beta = \frac{8\kappa\ell\alpha + \alpha}{3\kappa} + 4\alpha\ell = \alpha\frac{20\kappa\ell + 1}{3\kappa} \le b.$$

As $u \in I\!B_\alpha(\bar{u}(t)) \subset I\!B_a(\bar{u}(t)) \subset rI\!B_{\mathbb{R}^m}$ and $x \in I\!B_\alpha(\bar{x}(t)) \subset I\!B_a(\bar{x}(t)) \subset rI\!B_{\mathbb{R}^n}$, by (16) we get

$$\begin{aligned}\|\varphi(\bar{u}(t))\| &= \|f(\bar{x}(t),\bar{u}(t)) - f(x,u) - \nabla_u f(x,u)(\bar{u}(t) - u)\| \\ &\le \|f(\bar{x}(t),\bar{u}(t)) - f(x,\bar{u}(t))\| + \|f(x,\bar{u}(t)) - f(x,u) - \nabla_u f(x,u)(\bar{u}(t) - u)\| \\ &\le \ell\|\bar{x}(t) - x\| + \frac{\ell}{2}\|\bar{u}(t) - u\|^2 < \ell\alpha + \ell\alpha^2 \le 2\ell\alpha = \beta. \qquad (18)\end{aligned}$$

Since $2\ell\alpha \le 1/(3\kappa) = \mu$, for arbitrary v, $v' \in \mathbb{R}^m$, we have

$$\begin{aligned}\|\varphi(v) - \varphi(v')\| &= \|(\nabla_u f(x,u) - \nabla_u f(\bar{x}(t),\bar{u}(t))(v-v')\| \\ &\le \ell\big(\|x-\bar{x}(t)\| + \|u-\bar{u}(t)\|\big)\|v-v'\| \le 2\ell\alpha\|v-v'\| \le \mu\|v-v'\|.\end{aligned}$$

Moreover, observing that $\varphi + \mathcal{G}_t = f(x,u) + \nabla_u f(x,u)(\cdot - u) + F$, we get

$$\begin{aligned}\varphi(\bar{u}(t)) &= f(x,u) + \nabla_u f(x,u)(\bar{u}(t)-u) - f(\bar{x}(t),\bar{u}(t)) \\ &\in f(x,u) + \nabla_u f(x,u)(\bar{u}(t)-u) + F(\bar{u}(t)) = (\varphi+\mathcal{G}_t)\big(\bar{u}(t)\big).\end{aligned}$$

Hence $\bar{u}(t) \in (\varphi+\mathcal{G}_t)^{-1}\big(\varphi(\bar{u}(t))\big)$ and $\varphi(\bar{u}(t)) \in I\!B_\beta(0)$. Remembering that $y \in I\!B_\beta(0)$. Theorem 2.1 implies that there is $w \in (\varphi+\mathcal{G}_t)^{-1}(y)$ such that $\|w-\bar{u}(t)\| \le \kappa'\|y-\varphi(\bar{u}(t))\|$. Then $y \in f(x,u) + \nabla_u f(x,u)(w-u) + F(w)$ and (18) implies that

$$\|w-\bar{u}(t)\| \le \kappa'(\|y\| + \ell\|x-\bar{x}(t)\| + \ell\|u-\bar{u}(t)\|^2),$$

which proves the claim because $\ell \ge 1$.

Use Lemma 3.1 to find $m > 0$ such that for each τ_1, $\tau_2 \in [0,\varepsilon]$, with $\tau_1 < \tau_2$, we have

$$\left\|\frac{(\tau_2-\tau_1)}{2}\big(g(\bar{x}(\tau_1),\bar{u}(\tau_1)) + g(\bar{x}(\tau_2),\bar{u}(\tau_2))\big) - \int_{\tau_1}^{\tau_2} g(\bar{x}(t),\bar{u}(t))\,\mathrm{d}t\right\| \le m(\tau_2-\tau_1)^3. \tag{19}$$

Pick an arbitrary $\Delta > 0$. Let

$$q := \max\{4\ell^2, \Delta, \kappa'\ell, \varepsilon^2, m\}, \quad \lambda := 4q^3, \quad \text{and} \quad \bar{d} := q\big(\varepsilon\lambda e^{\varepsilon\lambda} + 4q\big).$$

Choose $N_0 \in \mathbb{N}$ such that $2\bar{d} < N_0$ and $q\varepsilon \le N_0 \min\{\alpha,\beta\}$. Fix any $N > N_0$ and let $h := \varepsilon/N$. Then

$$h < \frac{\varepsilon}{N_0} \le \frac{\sqrt{q}}{N_0} < \frac{\sqrt{q}}{2\bar{d}} < \frac{1}{2} \quad \text{and} \quad h \le qh < q\frac{\varepsilon}{N_0} \le \min\{\alpha,\beta\}. \tag{20}$$

Let $(x_0,u_0) \in I\!B_{\Delta h^2}(\bar{x}(0)) \times I\!B_{\Delta h^2}(\bar{u}(0))$ and $(e_i)_{i=0}^{N-1}$ in $I\!B_{\Delta h^2}(0)$ be arbitrary. For each $i \in \{0,1,\dots,N\}$, let $t_i := ih$ and $c_i := \lambda i e^{\lambda i h}$. Since $q \ge \Delta$, we have

$$\|x_0 - \bar{x}(0)\| \le qh^2 = (c_0 h + q)h^2 \quad \text{and} \quad \|u_0 - \bar{u}(0)\| \le qh^2 < q(c_0 h + 4q)h^2.$$

As $qh^2 < qh/2 < \alpha/2$ we have $(x_0,u_0) \in I\!B_\alpha(\bar{x}(t_0)) \times I\!B_\alpha(\bar{u}(t_0))$. We proceed by induction. Suppose that for some $i \in \{0,1,\dots,N-1\}$ a point $(x_i,u_i) \in I\!B_\alpha(\bar{x}(t_i)) \times I\!B_\alpha(\bar{u}(t_i))$ verifies

$$\|x_i - \bar{x}(t_i)\| \le (c_i h + q)h^2 \quad \text{and} \quad \|u_i - \bar{u}(t_i)\| \le q(c_i h + 4q)h^2. \tag{21}$$

We will show that there are [uniquely determined] points $\widetilde{x}_{i+1}$, $x_{i+1} \in I\!B_\alpha(\bar{x}(t_{i+1}))$ and $u_{i+1} \in I\!B_\alpha(\bar{u}(t_{i+1}))$ satisfying (12) such that (21) holds for $i := i+1$.

Let $\widetilde{x}_{i+1}$ be defined by the first equality in (12). Clearly, for any $s \in [t_i, t_{i+1}]$, we have

$$\begin{aligned}\|g(x_i,u_i) - g(\bar{x}(s),\bar{u}(s))\| &\le \ell(\|x_i - \bar{x}(s)\| + \|u_i - \bar{u}(s)\|) \\ &\le \ell(\|x_i - \bar{x}(t_i)\| + \ell(s-t_i) + \|u_i - \bar{u}(t_i)\| + \ell(s-t_i)) \\ &= \ell(\|x_i - \bar{x}(t_i)\| + \|u_i - \bar{u}(t_i)\|) + 2\ell^2(s-t_i). \qquad (22)\end{aligned}$$

As $c_ih < \varepsilon\lambda e^{\varepsilon\lambda}$ and $\ell\bar{d}h < q/4$, using (22) and (20) we get

$$\begin{aligned}
\|\widetilde{x}_{i+1} - \bar{x}(t_{i+1})\| &= \left\| x_i + hg(x_i,u_i) - \bar{x}(t_i) - \int_{t_i}^{t_{i+1}} g(\bar{x}(s),\bar{u}(s))\,\mathrm{d}s \right\| \\
&\le \|x_i - \bar{x}(t_i)\| + \int_{t_i}^{t_{i+1}} \|g(x_i,u_i) - g(\bar{x}(s),\bar{u}(s))\|\,\mathrm{d}s \\
&\le \|x_i - \bar{x}(t_i)\| + \ell h(\|x_i - \bar{x}(t_i)\| + \|u_i - \bar{u}(t_i)\|) + \ell^2h^2 \\
&= (1+\ell h)\|x_i - \bar{x}(t_i)\| + \ell h\|u_i - \bar{u}(t_i)\| + \ell^2h^2 \\
&\le (1+\ell h)(c_ih+q)h^2 + \ell\bar{d}h^3 + \ell^2h^2 \\
&= \big(c_ih + \ell(c_ih+q)h + q + \ell\bar{d}h + \ell^2\big)\,h^2 \\
&< \big(c_ih + \ell\bar{d}h + q + \ell\bar{d}h + q/4\big)\,h^2 < (c_ih + q/4 + q + q/4 + q/4)\,h^2 \\
&< (c_ih + 2q)\,h^2 < (\bar{d}/q)h^2 = h(\bar{d}h)/q < h/2 < \alpha/2. \qquad (23)
\end{aligned}$$

In particular $\widetilde{x}_{i+1} \in I\!B_\alpha(\bar{x}(t_{i+1}))$. Remembering that $c_ih < \varepsilon\lambda e^{\varepsilon\lambda}$, (21) and (20) yield that

$$\begin{aligned}
\|u_i - \bar{u}(t_{i+1})\| &\le \|u_i - \bar{u}(t_i)\| + \|\bar{u}(t_i) - \bar{u}(t_{i+1})\| < q(\varepsilon\lambda e^{\varepsilon\lambda} + 4q)h^2 + \ell h \qquad (24)\\
&= (\bar{d}h)h + \ell h < \sqrt{q}h < \alpha.
\end{aligned}$$

Clearly, $e_i \in I\!B_\beta(0)$. The claim with $t := t_{i+1}$, $y := e_i$, $x := \widetilde{x}_{i+1}$, and $u := u_i$ together with (23), (24), and (20) yields a [unique] point $u_{i+1} \in I\!B_\alpha(\bar{u}(t_{i+1}))$ such that

$$e_i \in f(\widetilde{x}_{i+1},u_i) + \nabla_u f(\widetilde{x}_{i+1},u_i)(u_{i+1}-u_i) + F(u_{i+1})$$

satisfying

$$\begin{aligned}
\|u_{i+1} - \bar{u}(t_{i+1})\| &\le q(\|\widetilde{x}_{i+1} - \bar{x}(t_{i+1})\| + \|u_i - \bar{u}(t_{i+1})\|^2 + \|e_i\|) \\
&< q\,(c_ih + 2q + q + \Delta)\,h^2 \le q(c_ih+4q)h^2. \qquad (25)
\end{aligned}$$

As $c_i < c_{i+1}$, we obtain the latter estimate in (21) with $i := i+1$. Let x_{i+1} be defined by the last equality in (12). Now (19), (21), (23), (25), and (20) imply that

$$\begin{aligned}
&\|x_{i+1} - \bar{x}(t_{i+1})\| = \left\| x_i + \frac{h}{2}\big(g(x_i,u_i) + g(\widetilde{x}_{i+1},u_{i+1})\big) - \bar{x}(t_i) - \int_{t_i}^{t_{i+1}} g(\bar{x}(s),\bar{u}(s))\,\mathrm{d}s \right\| \\
&\le \|x_i - \bar{x}(t_i)\| + mh^3 + \frac{h}{2}\,\|g(x_i,u_i) + g(\widetilde{x}_{i+1},u_{i+1}) - g(\bar{x}(t_i),\bar{u}(t_i)) - g(\bar{x}(t_{i+1}),\bar{u}(t_{i+1}))\| \\
&\le (c_ih+q)h^2 + mh^3 + \frac{\ell h}{2}\big(\|x_i - \bar{x}(t_i)\| + \|u_i - \bar{u}(t_i)\| + \|\widetilde{x}_{i+1} - \bar{x}(t_{i+1})\| + \|u_{i+1} - \bar{u}(t_{i+1})\|\big) \\
&\le (c_i+m)h^3 + qh^2 + \frac{\ell h}{2}\big((c_ih+q)h^2 + q(c_ih+4q)h^2 + (c_ih+2q)h^2 + q(c_ih+4q)h^2\big) \\
&< (c_i+q)h^3 + \frac{h^3}{4}\big(q(c_ih+q) + q^2(c_ih+4q) + q(c_ih+2q) + q^2(c_ih+4q)\big) + qh^2 \\
&= c_i\big(1 + (q+q^2)h/2\big)h^3 + (q + 3q^2/4 + 2q^3)h^3 + qh^2 < c_i(1+4q^3h)h^3 + 4q^3h^3 + qh^2 \\
&= c_i\big(1+\lambda h\big)h^3 + \lambda h^3 + qh^2 \le \lambda i e^{\lambda(i+1)h}h^3 + \lambda e^{\lambda(i+1)h}h^3 + qh^2 \\
&= \lambda(i+1)e^{\lambda(i+1)h}h^3 + qh^2 = (c_{i+1}h+q)h^2.
\end{aligned}$$

The first estimate in (21) with $i := i+1$ is proved. Since $(c_{i+1}h+q)h^2 < \bar{d}h^2 < qh/2 < \alpha/2$, we have $x_{i+1} \in I\!B_\alpha(\bar{x}(t_{i+1}))$. The induction step is complete and so is the proof by noting that for each $i \in \{0,1,\ldots,N\}$ we have $c_i h \le \varepsilon\lambda e^{\varepsilon\lambda}$. □

If $\bar{u}(\cdot)$ is only Lipschitz continuous on $[0,\varepsilon]$, one can consider the following iteration:

$$\text{(26)} \qquad \begin{cases} x_{i+1} &= x_i + hg(x_i,u_i), \\ e_i &\in f(x_{i+1},u_i) + \nabla_u f(x_{i+1},u_i)(u_{i+1}-u_i) + F(u_{i+1}), \end{cases}$$

Using a similar technique as in the proof of Theorem 3.2 we obtain:

Theorem 3.3. *Consider the DGE* (1) *and suppose that f and g are differentiable functions with a locally Lipschitz continuous derivative, and that F has a closed graph. Let a pair of functions $(\bar{x}(\cdot),\bar{u}(\cdot))$ be a solution of* (1) *such that both $\bar{x}(\cdot)$ and $\bar{u}(\cdot)$ are Lipschitz continuous on $[0,\varepsilon]$. Suppose that for each $t \in [0,\varepsilon]$ the mapping $\mathcal{G}_t$ in* (13) *is [strongly] regular at $\bar{u}(t)$ for 0. Then for any $\Delta > 0$ there are $N_0 \in \mathbb{N}$ and positive constants α and $\bar{d}$ such that for each $N > N_0$, each $(x_0,u_0) \in I\!B_{\Delta h}(\bar{x}(0)) \times I\!B_{\Delta h}(\bar{u}(0))$, and each $(e_i)_{i=0}^{N-1}$ in $I\!B_{\Delta h}(0)$, where $h := \varepsilon/N$, there are [uniquely determined] points $(x_i,u_i) \in \mathbb{R}^n \times \mathbb{R}^m$, $i \in \{1,\ldots,N\}$, generated by the iteration* (26)*, with the initial point (x_0,u_0), such that $(x_i,u_i) \in I\!B_\alpha(\bar{x}(t_i)) \times I\!B_\alpha(\bar{u}(t_i))$ for each $i \in \{1,\ldots,N\}$ satisfying*

$$\text{(27)} \qquad \max_{0\le i\le N} \|x_i - \bar{x}(t_i)\| \le \bar{d}h \quad \text{and} \quad \max_{0\le i\le N} \|u_i - \bar{u}(t_i)\| \le \bar{d}h.$$

The above statement is a slight extension of [5, Theorem 5.1]. Next, we discuss two basic examples from engineering, which can be formulated either as a DGE or an ODE with a Lipschitz continuous right-hand side. We compare schemes (12) and (26) with the method *ODE45* which is used with the absolute error tolerance 10^{-12} to get a reference solution trajectory. All simulations are performed in MATLAB.

Example 3.4. Consider a particle of mass $m > 0$ connected by a rigid, weightless rod of length $\ell > 0$ to a base by means of a pin joint that can rotate in a plane due to gravity. In addition, the pendulum can have a contact with two walls made of a very flexible material which are at a distance $r > 0$ from a pin joint. The contact force acting on the mass at time t is denoted by $u(t)$; and $\varphi_1(t)$ and $\varphi_2(t)$ denote the angular displacement and the angular velocity at time t, respectively (see Figure 6.1). The equations of motion of the system are

$$\begin{cases} \dot{\varphi}_1(t) &= \varphi_2(t), \\ \dot{\varphi}_2(t) &= -\frac{g}{\ell}\sin\varphi_1(t) - \frac{1}{\ell m}H(\varphi_1(t)), \qquad \text{for all} \quad t \in [0,\varepsilon], \\ \varphi_1(0) &= \gamma_1, \quad \varphi_2(0) = \gamma_2, \end{cases}$$

with given initial conditions γ_1, $\gamma_2 \in \mathbb{R}$, a gravitational acceleration $g = 9.81$ ms^{-2}, and $u(t) = H(\varphi_1(t))$ describing the dependence of the contact force on the angular displacement. We assume that

$$H(\varphi) = \begin{cases} \operatorname{argsinh}(\varphi - \arcsin(r/\ell)) & \text{for} \quad \varphi > \arcsin(\ell/r), \\ \operatorname{argsinh}(\varphi + \arcsin(r/\ell)) & \text{for} \quad \varphi < -\arcsin(\ell/r), \\ 0 & \text{otherwise.} \end{cases}$$

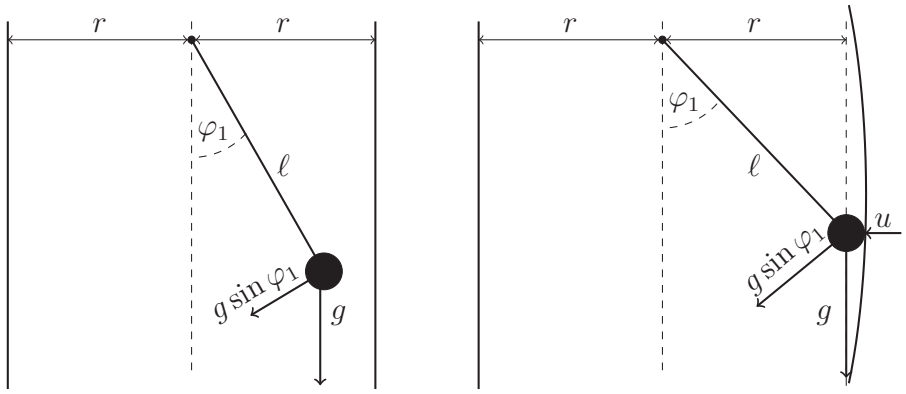


Figure 6.1: Mechanical model from Example 3.4.

The corresponding DGE has form

$$\begin{cases} \dot{\varphi}_1(t) &= \varphi_2(t), \\ \dot{\varphi}_2(t) &= -\frac{g}{\ell}\sin\varphi_1(t) - \frac{1}{\ell m}u(t), \\ 0 &\in -\varphi_1(t) + \sinh u(t) + \arcsin(r/\ell)\ \partial|\cdot|(u(t)), \\ \varphi_1(0) &= \gamma_1, \quad \varphi_2(0) = \gamma_2, \end{cases} \qquad \text{for all} \quad t \in [0,\varepsilon],$$

where ∂ denotes a subdifferential in the sense of convex analysis. The solution for $\ell = m := 1, r := \sin 1$, $\varepsilon := 2, \gamma_1 = \pi/3$, and $\gamma_2 = 0$ is in Figure 6.2. The grid errors with respect to the solution obtained by *ODE45* are in Figure 6.3. For both the schemes (12) and (26), we use the discretion step $h = 10^{-5}$ and $e_i = 0$, $i \in \{0, 1, \ldots, N-1\}$.

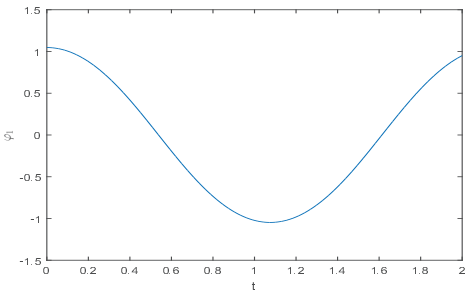
Example 3.5. Consider a circuit in Figure 6.4 involving the four-diodes bridge full-wave rectifier, a resistor with a resistance $R > 0$, a capacitor with the capacitance $C_0 > 0$ and an inductor with the inductance $L > 0$. Denote v_C a voltage across the capacitor, i_C a current through the capacitor, i_L a current through the inductor, $i_{DF1}, i_{DF2}, i_{DR1}, i_{DR2}$ currents through the diodes, and v_{DF1}, v_{DF2}, v_{DR1}, v_{DR2} voltages across the diodes, respectively. Using the Kirchhoff's laws, this problem can be described as a particular DGE (see [4]) called a *differential linear complementarity problem (system)* in the form

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ 0 &\le Cx(t) + Du(t) \perp u(t) \ge 0, \qquad t \in [0,\varepsilon], \\ x(0) &= x_I, \end{cases} \tag{28}$$

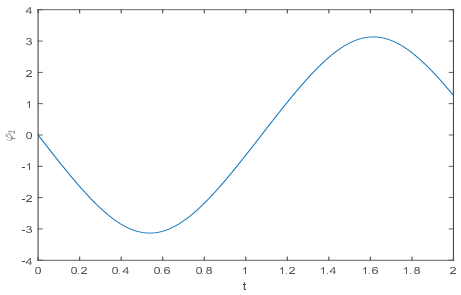
where

$$x := \begin{pmatrix} v_C \\ i_L \end{pmatrix}, A := \begin{pmatrix} 0 & -\frac{1}{C_0} \\ \frac{1}{L} & 0 \end{pmatrix}, B := \begin{pmatrix} 0 & 0 & -\frac{1}{C_0} & \frac{1}{C_0} \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

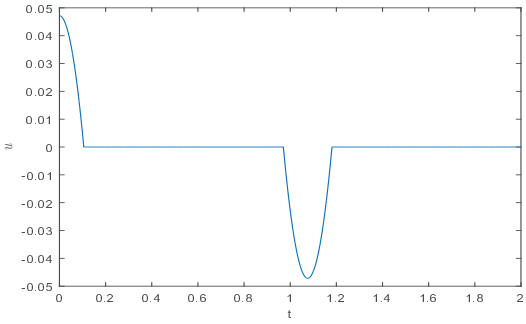
$$u := \begin{pmatrix} -v_{DR1} \\ -v_{DF2} \\ i_{DF1} \\ i_{DR2} \end{pmatrix}, C := \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 1 & 0 \end{pmatrix}, D := \begin{pmatrix} \frac{1}{R} & \frac{1}{R} & -1 & 0 \\ \frac{1}{R} & \frac{1}{R} & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$



(a) The first component φ_1.



(b) The second component φ_2.



(c) The third component u.

Figure 6.2: The solution from Example 3.4.

the symbol $\perp$ denotes a complementarity relation, and inequalities in $\mathbb{R}^4$ are understood coordinate-wise. From (28) we have $v_{DR1}(t) = -\max\{v_C(t), 0\}$, $v_{DF2}(t) = -\max\{-v_C(t), 0\}$, $i_{DF1}(t) = 1/R \max\{v_C(t), 0\}$, and $i_{DR2}(t) = 1/R \max\{-v_C(t), 0\}$ for each $t \in [0, \varepsilon]$. Hence the problem is equivalent to the system of ordinary differential equations, in the form

$$\dot{x}(t) = Ax(t) + Bu(t), \quad t \in [0, \varepsilon], \quad \text{and} \quad x(0) = x_I.$$

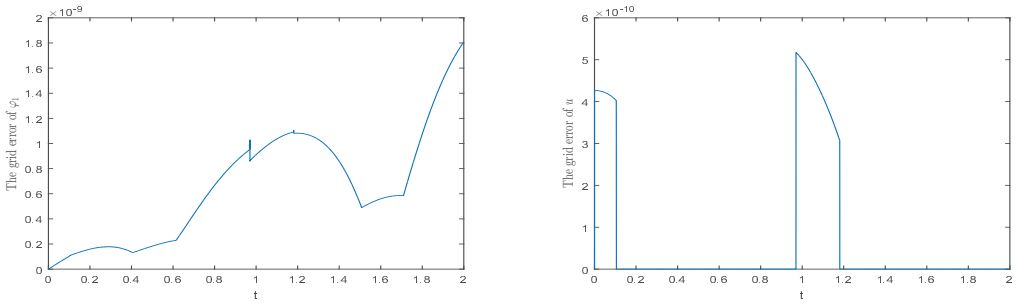
For the simulation we use library LCP[1] and assume that $C_0 := 10^{-6}$, $L := 0.01$, $R := 1000$, $\varepsilon := 0.005$, and $x_I := [10, 0]$. For both the schemes (12) and (26), we use the discretion step $h = 10^{-8}$ and $e_i = 0$, $i \in \{0, 1, \ldots, N-1\}$. Graphs of solution components are in Figure 6.5 while grid errors are in Figure 6.6. We note that the maximal grid error means the biggest error of elements of u or x at the points of the grid.

To conclude this section, let us point out that a similar technique, can be used also in the case of a *parametric generalized equation*, which is a problem for a fixed function $p : [0, \varepsilon] \to \mathbb{R}^n$, find a function $z : [0, \varepsilon] \to \mathbb{R}^n$ such that

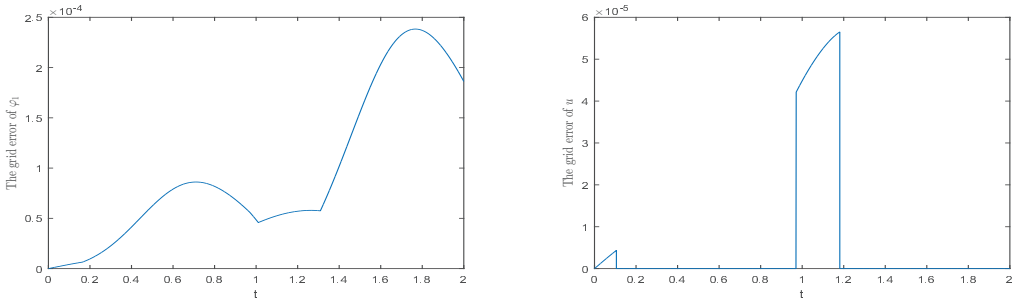
$$p(t) \in f(z(t)) + F(z(t)) \quad \text{for all} \quad t \in [0, \varepsilon], \tag{29}$$

where a constant $\varepsilon > 0$, a function $f : \mathbb{R}^n \to \mathbb{R}^n$ and a set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ are given. This problem can be used, for example, for modeling static problems from electronics, that is, when no capacitors and inductors appear in the circuit [1, 2, 3, 14].

[1]It is available on: https://www.mathworks.com/matlabcentral/fileexchange/20952-lcp—mcp-solver–newton-based-?requestedDomain=www.mathworks.com



(a) Grid errors of the scheme (12).



(b) Grid errors of the scheme (26).

Figure 6.3: Errors of the solution from Example 3.4.

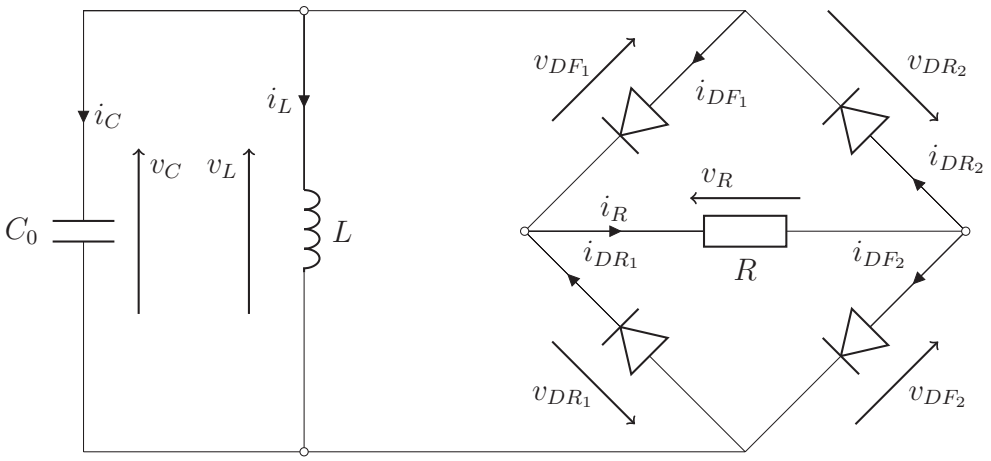
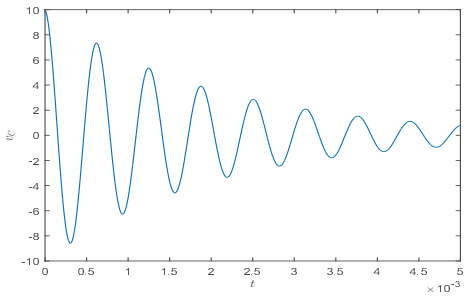


Figure 6.4: The circuit from Example 3.5.

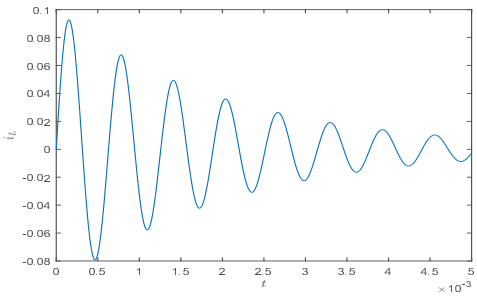
For an integer $N > 1$, define the uniform grid $t_i := ih$, $i \in \{0, 1, \ldots, N\}$, with a step size $h := \varepsilon/N$. Given $\Delta > 0$ and points $(e_i)_{i=0}^N$ in $I\!B_{\Delta h^2}(p(t_{i+1}))$, we study a predictor-corrector scheme in the form

$$\left\{ \begin{array}{lcl} e_i & \in & f(z_i) + \nabla f(z_i)(v_{i+1} - z_i) + F(v_{i+1}), \\ p(t_{i+1}) & \in & f(v_{i+1}) + \nabla f(v_{i+1})(z_{i+1} - v_{i+1}) + F(z_{i+1}), \end{array} \right. \tag{30}$$

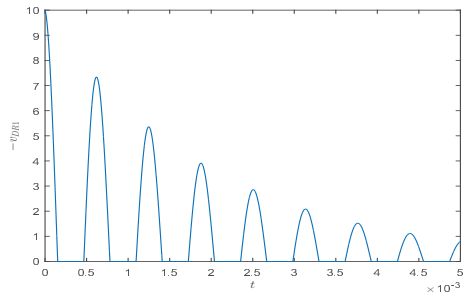
where z_0 is sufficiently close to the exact solution of (29) at time $t := 0$. Uniform regularity



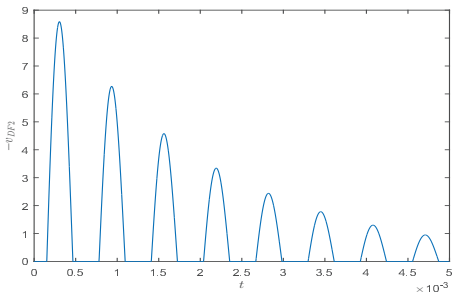
(a) The first component of $x(\cdot)$.



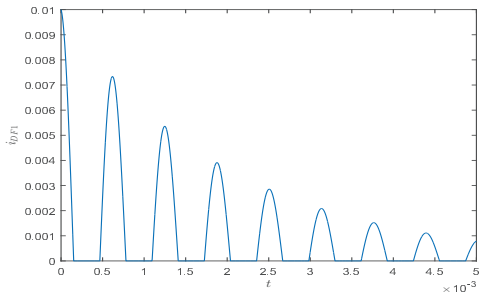
(b) The second component of $x(\cdot)$.



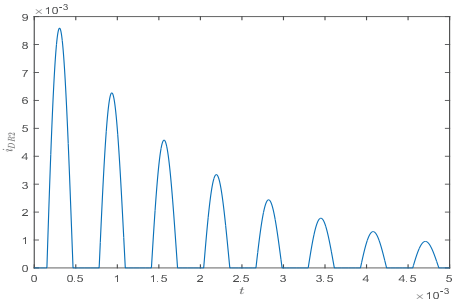
(c) The first component of $u(\cdot)$.



(d) The second component of $u(\cdot)$.



(e) The third component of $u(\cdot)$.



(f) The fourth component of $u(\cdot)$.

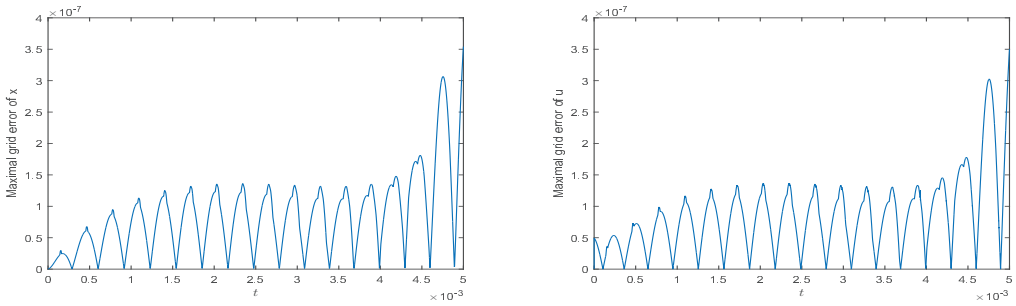
Figure 6.5: Graphs of the solution from Example 3.5.

along a continuous path was used in [6] to obtain the following extension of the main result from [12].

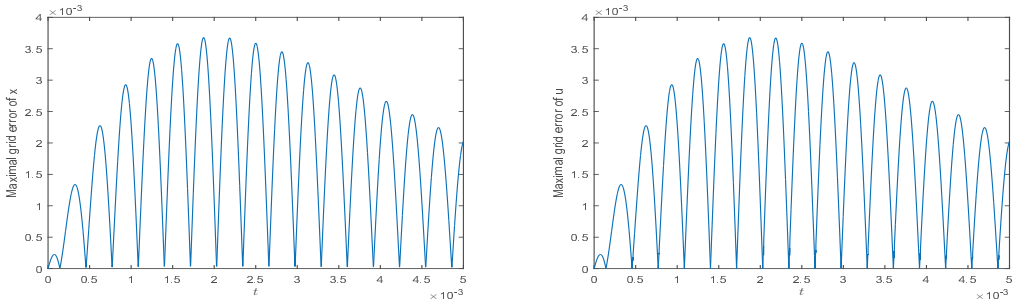
Theorem 3.6. *Let $\bar{z} : [0, \varepsilon] \to \mathbb{R}^n$ be a Lipschitz continuous solution of the problem (29), where $p : [0, \varepsilon] \to \mathbb{R}^n$ is Lipschitz continuous, $f : \mathbb{R}^n \to \mathbb{R}^n$ has a locally Lipschitz continuous derivative on whole of $\mathbb{R}^n$, and $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ has a closed graph. Suppose that for each $t \in [0, \varepsilon]$ the mapping*

$$\mathbb{R}^n \ni v \longmapsto \mathcal{G}_t(v) := f(\bar{z}(t)) + \nabla f(\bar{z}(t))(v - \bar{z}(t)) + F(v) \subset \mathbb{R}^n$$

is [strongly] regular at $\bar{z}(t)$ for $p(t)$. Then there is $\alpha > 0$ such that for any $\Delta > 0$ there are constants $N_0 \in \mathbb{N}$ and $c > 0$ such that for each $N > N_0$ and each $z_0 \in I\!B_{\Delta h^4}(\bar{z}(t_0))$, where $h := \varepsilon/N$, there are [uniquely determined] points $(z_i)_{i=1}^N$ generated by the iteration



(a) Maximal grid error of the scheme (12).



(b) Maximal grid error of the scheme (26).

Figure 6.6: Errors of the solution from Example 3.5.

(30), *with the initial point* z_0 *and arbitrarily chosen points* $(e_i)_{i=0}^{N-1}$ *in* $\mathbb{B}_{\Delta h^2}(p(t_{i+1}))$, *such that* $z_i \in \mathbb{B}_\alpha(\bar{z}(t_i))$ *for each* $i \in \{0, \dots, N\}$ *and*

$$\max_{0 \le i \le N} \|z_i - \bar{z}(t_i)\| \le ch^4. \tag{31}$$

The point e_i appearing in (30) can be interpreted as a sufficiently precise prediction at time t_i of the (possibly unknown) value of $p(t_{i+1})$. Then we wait until the precise value of $p(t_{i+1})$ is known and compute a correction z_{i+1}. On the other hand, taking $e_i := p(t_i) + hp'(t_i)$, $i \in \{0, 1, \dots, N-1\}$, we have $\|e_i - p(t_{i+1})\| \le \Delta h^2$ provided that $p'(\cdot)$ exists and is Lipschitz on $[0, \varepsilon]$ with the constant 2Δ. Hence the algorithm proposed in [13, Section 6G] is a particular case of (30). Finally, instead of $p(t_{i+1})$ in the latter inclusion of (30) one can take any $\tilde{e}_i \in \mathbb{B}_{\Delta h^4}(p(t_{i+1}))$, that is, the corrector step can be done via an inexact method (which is always the case in practice). Finally, let us note that sufficient conditions (of different type) guaranteeing the existence of a Lipschitz continuous solution $\bar{z}(\cdot)$ of (29) can be found either in [6, Theorem 6] or [5, Theorem 11].

4 Uniform regularity and regularity in function spaces

In case that the solution trajectory is not continuous (or even defined) on the whole time interval we can derive the following statement.

Theorem 4.1. *Let $\varepsilon > 0$ and S be a non-empty subset of $[0, \varepsilon]$. Consider a pair of bounded functions $\bar{x} : S \to \mathbb{R}^n$ and $\bar{u} : S \to \mathbb{R}^m$ such that*

$$0 \in f(\bar{x}(t), \bar{u}(t)) + F(\bar{u}(t)) \quad \text{for each} \quad t \in S,$$

with a continuous $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$ having a continuous derivative $\nabla_u f$ and $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^d$ having a closed graph. Let $\Lambda := \cup_{t\in S}(\bar{x}(t), \bar{u}(t))$ and for each $(x, u) \in \operatorname{cl}\Lambda$ define a mapping

$$\mathbb{R}^m \ni v \longmapsto \mathcal{G}_{x,u}(v) := f(x, u) + \nabla_u f(x, u)(v - u) + F(v) \subset \mathbb{R}^d. \tag{32}$$

Then the following statements are equivalent:

- *(i) for each $(x, u) \in \operatorname{cl}\Lambda$ the mapping $\mathcal{G}_{x,u}$ is [strongly] regular at u for 0;*
- *(ii) there are positive constants a, b, and κ such that for each $(x, u) \in \operatorname{cl}\Lambda$ the mapping $\mathcal{G}_{x,u}$ is [strongly] regular at u for 0 with the constant κ and neighborhoods $I\!B_a(u)$ and $I\!B_b(0)$;*
- *(iii) there are positive constants a, b, and κ such that for each $t \in S$ the mapping $\mathcal{G}_t$ in (13) is [strongly] regular at $\bar{u}(t)$ for 0 with the constant κ and neighborhoods $I\!B_a(\bar{u}(t))$ and $I\!B_b(0)$.*

Proof. Assume that (i) holds. Define a (compact) set $\Omega := \operatorname{cl}\big(\cup_{t\in S}(\bar{x}(t), \bar{u}(t), \bar{u}(t))\big)$ and a (continuous) function $\sigma(x, u, v) := f(x, u) + \nabla_u f(x, u)(v - u)$, $(x, u, v) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$. Note that $(x, u, v) \in \Omega$ if and only if $v = u$ and $(x, u) \in \operatorname{cl}\Lambda$. Theorem 2.4 yields positive constants a, b, and κ such that for each $(x, u, u) \in \Omega$ the mapping $\mathcal{G}_{x,u}$ is [strongly] regular at u for 0 with the constant κ and neighborhoods $I\!B_a(u)$ and $I\!B_b(0)$. Since $(\bar{x}(t), \bar{u}(t), \bar{u}(t)) \in \Omega$ and $\mathcal{G}_t = \mathcal{G}_{\bar{x}(t),\bar{u}(t)}$ for each $t \in S$, (iii) is proved.

Assume that (iii) holds. Let $\kappa' := 2\kappa$ and $\mu := 1/(3\kappa)$. Then $\kappa\mu < 1$ and $\kappa' > \kappa/(1-\kappa\mu)$. Pick $r > 0$ such that $\bar{x}(S) + aI\!B_{\mathbb{R}^n} \subset rI\!B_{\mathbb{R}^n}$ and $\bar{u}(S) + aI\!B_{\mathbb{R}^m} \subset rI\!B_{\mathbb{R}^m}$. As f and $\nabla_u f$ are continuous, they are uniformly continuous on a compact set $\Omega := rI\!B_{\mathbb{R}^n} \times rI\!B_{\mathbb{R}^m}$. Find $\beta > 0$ such that both $2\kappa'\beta + \beta < a$ and $\mu(2\kappa'\beta + \beta) + 2\beta < b$; and also that for each $(x, u) \in \Omega$ and each $(x', u') \in (I\!B_{2\kappa'\beta+\beta}(x) \times I\!B_{2\kappa'\beta+\beta}(u)) \cap \Omega$ we have

$$\|\nabla_u f(x', u') - \nabla_u f(x, u)\| < \mu \quad \text{and} \quad \|f(x', u') - f(x, u) - \nabla_u f(x', u')(u' - u)\| < \beta.$$

Fix any $(x, u) \in \operatorname{cl}\Lambda \subset \Omega$. Then $0 \in \mathcal{G}_{x,u}(u)$ since f is continuous and $\operatorname{gph} F$ is closed. Find $\bar{t} \in S$ such that $(x, u) \in I\!B_\beta(\bar{x}(\bar{t})) \times I\!B_\beta(\bar{u}(\bar{t}))$. Then $\mathcal{G}_{x,u} = \mathcal{G}_{\bar{t}} + g$, with

$$g(v) = f(x, u) + \nabla_u f(x, u)(v - u) - f(\bar{x}(\bar{t}), \bar{u}(\bar{t})) - \nabla_u f(\bar{x}(\bar{t}), \bar{u}(\bar{t}))(v - \bar{u}(\bar{t})), \quad v \in \mathbb{R}^m.$$

Then $\|g(\bar{u}(\bar{t}))\| = \|f(x, u) - f(\bar{x}(\bar{t}), \bar{u}(\bar{t})) - \nabla_u f(x, u)(u - \bar{u}(\bar{t}))\| < \beta$. Moreover, for any v, $v' \in \mathbb{R}^m$ we have $\|g(v) - g(v')\| = \|[\nabla_u f(x, u) - \nabla_u f(\bar{x}(\bar{t}), \bar{u}(\bar{t}))](v - v')\| \le \mu\|v - v'\|$. Applying Theorem 2.1, with $\alpha := \beta$, and using a similar reasoning as in the proof of Theorem 2.4 we conclude that the mapping $\mathcal{G}_{x,u}$ is [strongly] regular at u for 0 uniformly in $(x, u) \in \operatorname{cl}\Lambda$. Hence (ii) holds. Clearly, (ii) implies (i). □

The above statement is a generalization of [5, Theorem 7], where strong regularity is considered only, because it requests point-wise [strong] regularity on the closure of the range of the solution instead of on the closure of its graph. The function $\bar{x}(\cdot)$ can be either an input signal in a parametric generalized equation (29) or a state trajectory of the DGE (1). In the latter case, $\bar{x}(\cdot)$ is continuous on $S = [0, \varepsilon]$, so if $\bar{u}(\cdot)$ has closed range, then the uniform [strong] regularity of $\mathcal{G}_t$ in (13) on S is equivalent to its point-wise [strong] regularity on S. We also get the following *uniform* version of the Lyusternik-Graves and Robinson theorem which implies [5, Theorem 9] under substantially weaker assumptions.

Theorem 4.2. *Let ε, S, $\bar{x}(\cdot)$, $\bar{u}(\cdot)$, f, and F be as in Theorem 4.1. Then the mapping $G_t = f(\bar{x}(t), \cdot) + F$ is [strongly] regular at $\bar{u}(t)$ for 0 uniformly in $t \in S$ if and only if so is the mapping $\mathcal{G}_t$ in* (13).

Proof. Suppose that there are positive constants a, b and κ such that for each $t \in S$ the mapping $\mathcal{G}_t$ in (13) is [strongly] regular at $\bar{u}(t)$ for 0 with the constant κ and neighborhoods $I\!B_a(\bar{u}(t))$ and $I\!B_b(0)$. Let β, κ', μ, r, Ω be as in the proof of $(iii) \Rightarrow (ii)$ in Theorem 4.1. Fix any $t \in S$. Let $g_t(v) := f(\bar{x}(t), v) - f(\bar{x}(t), \bar{u}(t)) - \nabla_u f(\bar{x}(t), \bar{u}(t))(v - \bar{u}(t))$, $v \in \mathbb{R}^m$. Then $g_t(\bar{u}(t)) = 0$ and for any v, $v' \in I\!B_{2\kappa'\beta+\beta}(\bar{u}(t))$ we have

$$\begin{aligned}
\|g_t(v) - g_t(v')\| &= \|f(\bar{x}(t), v) - f(\bar{x}(t), v') - \nabla_u f(\bar{x}(t), \bar{u}(t))(v - v')\| \\
&= \|\int_0^1 \big(\nabla_u f(\bar{x}(t), v' + s(v - v')) - \nabla_u f(\bar{x}(t), \bar{u}(t))\big)(v - v')\, \mathrm{d}s\| \\
&\leq \mu\|v - v'\|.
\end{aligned}$$

As in Theorem 4.1 we conclude that the mapping $G_t = g_t + \mathcal{G}_t$ is [strongly] regular at $\bar{u}(t)$ for 0 uniformly in $t \in S$. The converse implication follows in the same way. □

Before continuing we set up notions used later.

Notation (N)**.** Let a constant $\varepsilon > 0$, twice differentiable functions $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, and a closed convex subset U_{ad} of $\mathbb{R}^d$ be given. Consider the problem (2). The controls $u(\cdot)$ are assumed to be in $\mathcal{U} := \mathcal{L}^\infty([0, \varepsilon], \mathbb{R}^m)$, the space of essentially bounded and measurable functions on $[0, \varepsilon]$ with values in $\mathbb{R}^m$ considered with the norm $\|u(\cdot)\|_\infty := \operatorname{ess\,sup}\|u(\cdot)\|$, $u(\cdot) \in \mathcal{U}$. The state trajectories $x(\cdot)$ belong to $\mathcal{X} := \mathcal{W}_0^{1,\infty}([0, \varepsilon], \mathbb{R}^n)$, the space of Lipschitz continuous functions on $[0, \varepsilon]$ with values in $\mathbb{R}^n$ satisfying $x(0) = 0$ equipped with the norm $\|x(\cdot)\|_\mathcal{X} = \|x(\cdot)\|_\infty + \|\dot{x}(\cdot)\|_\infty$, $x(\cdot) \in \mathcal{X}$. Let $\mathcal{V} := \mathcal{X} \times \mathcal{U}$, $\mathcal{R} := \mathcal{L}^\infty([0, \varepsilon], \mathbb{R}^n)$, $\mathcal{P} := \mathcal{L}^\infty([0, \varepsilon], \mathbb{R}^d)$,

$$\mathcal{U}_{ad} := \{u(\cdot) \in \mathcal{U} \,\big|\, u(t) \in U_{ad} \text{ for a.e. } t \in [0, \varepsilon]\},$$

and $\mathcal{W} := \mathcal{R} \times \mathcal{P}$. Given a solution $(\bar{x}(\cdot), \bar{u}(\cdot)) \in \mathcal{V}$ of (2) we set $A(t) = \nabla_x g(\bar{x}(t), \bar{u}(t))$, $B(t) = \nabla_u g(\bar{x}(t), \bar{u}(t))$, $C(t) = \nabla_x f(\bar{x}(t), \bar{u}(t))$, $D(t) = \nabla_u f(\bar{x}(t), \bar{u}(t))$, and $\bar{f}(t) = f(\bar{x}(t), \bar{u}(t))$ for a.e. $t \in [0, \varepsilon]$. Let Φ be the fundamental matrix solution of the linear equation $\dot{z} = A(t)z$, that is, $\frac{d}{dt}\Phi(t, \tau) = A(t)\Phi(t, \tau)$, $\Phi(\tau, \tau) = I$.

Consider a set-valued mapping $H : \mathcal{V} \rightrightarrows \mathcal{W}$ defined by

$$\mathcal{V} \ni (x(\cdot), u(\cdot)) \longmapsto H(x(\cdot), u(\cdot)) := \begin{pmatrix} \dot{x}(t) - g(x(t), u(t)) \\ f(x(t), u(t)) - \mathcal{U}_{ad} \end{pmatrix} \subset \mathcal{W}$$

along with its shifted partial linearization $\mathcal{H}$ at $(\bar{x}(\cdot),\bar{u}(\cdot))$ defined for each $(z(\cdot),v(\cdot))\in\mathcal{V}$ by

$$\mathcal{H}(z(\cdot),v(\cdot)):=\left(\begin{array}{c}\dot{z}(t)-A(t)\,z(t)-B(t)\,v(t)\\ \bar{f}(t)+C(t)\,z(t)+D(t)\,v(t)-\mathcal{U}_{ad}\end{array}\right)\subset\mathcal{W},$$

a mapping $\mathcal{K}:\mathcal{U}\rightrightarrows\mathcal{P}$ defined as

$$\mathcal{K}[v(\cdot)](t):=\bar{f}(t)+C(t)\int_0^t\Phi(t,\tau)B(\tau)v(\tau)\,\mathrm{d}\tau+D(t)\,v(t)-\mathcal{U}_{ad},\quad v(\cdot)\in\mathcal{U},$$

and mappings G_t, $\mathcal{G}_t:\mathbb{R}^m\to\mathbb{R}^d$, $t\in S$, defined, respectively, for each $v\in\mathbb{R}^m$ by

$$G_t(v):=f(\bar{x}(t),v)-U_{ad}\quad\text{and}\quad\mathcal{G}_t(v):=\bar{f}(t)+D(t)\,(v-\bar{u}(t))-U_{ad}.$$

Now we are ready to formulate and prove the main result of this section generalizing [5, Theorem 3].

Theorem 4.3. *Under the notation (N), the following assertions are equivalent:*

(i) H is regular at $(\bar{x}(\cdot),\bar{u}(\cdot))$ for 0;

(ii) $\mathcal{H}$ is regular at $(0,0)$ for 0;

(iii) $\mathcal{K}$ is regular at 0 for 0;

(iv) there is a subset S of $[0,\varepsilon]$ having full Lebesgue measure such that the mapping G_t is regular at $\bar{u}(t)$ for 0 uniformly in $t\in S$;

(v) there is a subset S of $[0,\varepsilon]$ having full Lebesgue measure such that the mapping $\mathcal{G}_t$ is regular at $\bar{u}(t)$ for 0 uniformly in $t\in S$;

(vi) there is $\delta>0$ such that for every $w(\cdot)\in\mathcal{P}$ with $\|w(\cdot)\|_\infty<\delta$ there is $v(\cdot)\in\mathcal{U}$ with $\|v(\cdot)\|_\infty\le 1$ such that

$$\bar{f}(t)+C(t)\int_0^t\Phi(t,\tau)B(\tau)v(\tau)\,\mathrm{d}\tau+D(t)\,v(t)+w(t)\in U_{ad}\quad\text{for a.e.}\quad t\in[0,\varepsilon];$$

(vii) there are $\delta>0$ and $r>0$ such that for every $w(\cdot)\in\mathcal{P}$ with $\|w(\cdot)\|_\infty<\delta$ there is a pair $(z(\cdot),v(\cdot))\in r\mathbb{B}_{\mathcal{X}}\times r\mathbb{B}_{\mathcal{U}}$ such that

$$\bar{f}(t)+C(t)\,z(t)+D(t)\,v(t)+w(t)\in U_{ad}\quad\text{for a.e.}\quad t\in[0,\varepsilon].$$

Proof. Define a bounded linear mapping $\mathcal{Q}:\mathcal{R}\to\mathcal{X}$ by $\mathcal{Q}[r(\cdot)](t)=\int_0^t\Phi(t,\tau)\,r(\tau)\,\mathrm{d}\tau$ for $t\in[0,\varepsilon]$. Let $\nu:=\max\{\|A(\cdot)\|_\infty,\|B(\cdot)\|_\infty,\|C(\cdot)\|_\infty,\|D(\cdot)\|_\infty,\|\bar{x}(\cdot)\|_\infty,\|\bar{u}(\cdot)\|_\infty\}$.

Applying the Lyusternik-Graves theorem [13, Theorem 5E.6] and substituting $z(\cdot)=x(\cdot)-\bar{x}(\cdot)$ and $v(\cdot):=u(\cdot)-\bar{u}(\cdot)$, we obtain that $(i)\Leftrightarrow(ii)$. By Theorem 4.2 we have $(iv)\Leftrightarrow(v)$ because $\bar{x}(\cdot)$ is continuous and $\bar{u}(\cdot)$ is essentially bounded.

To prove that $(ii)\Leftrightarrow(iii)$, note that given $r(\cdot)\in\mathcal{R}$, one has that $\dot{z}(t)-A(t)z(t)=r(t)$ for a.e. $t\in[0,\varepsilon]$ and $z(0)=0$ if and only if $z(t)=\mathcal{Q}[r(\cdot)](t)$, $t\in[0,\varepsilon]$. This implies

that having $(r(\cdot),p(\cdot)) \in \mathcal{H}(z(\cdot),v(\cdot))$ is the same as having $w(t) \in \mathcal{K}[v(\cdot)](t)$ for $w(t) = p(t) - C(t)\mathcal{Q}[r(\cdot)](t)$, that is, we can replace the differential expression in $\mathcal{H}$ with the integral one and then drop the variable z. Moreover, $\|w(\cdot)\|_\infty$ is bounded by a quantity proportional to $\|(r(\cdot),p(\cdot))\|_{\mathcal{W}}$.

As $\mathcal{K}$ has a closed convex graph, $(iii) \Leftrightarrow (vi)$ by Robinson-Ursescu theorem [13, Theorem 5B.4]. If (vi) holds then setting $z(t) := \mathcal{Q}[B(\cdot)v(\cdot)](t)$, $t \in [0,\varepsilon]$, we get (vii) with $r := \max\{1, \nu\|\mathcal{Q}\|\}$.

Suppose that (vii) holds. We shall establish (v). Pick $\beta > 0$ such that $\bar{w}_\beta(\cdot) \equiv (\beta,\beta,\dots,\beta) \in \mathbb{R}^d$ has $\|\bar{w}_\beta(\cdot)\|_\infty < \delta$. Let $\{w_1, w_2, \dots\}$ be a countable dense subset of $I\!B_\beta(0)$. For any $i \in \mathbb{N}$, the function $w_i(\cdot) \equiv -w_i$ has $\|w_i(\cdot)\|_\infty \le \|\bar{w}_\beta(\cdot)\|_\infty < \delta$, thus there is a subset S_i of $[0,\varepsilon]$ having a full Lebesgue measure along with a pair $(z_i(\cdot), v_i(\cdot)) \in rI\!B_{\mathcal{X}} \times rI\!B_{\mathcal{U}}$ such that

$$\bar{f}(t) + C(t)\, z_i(t) + D(t)\, v_i(t) - w_i \in U_{ad} \quad \text{for all} \quad t \in S_i.$$

Without any loss of generality assume that $\|z_i(t)\| \le r$ and $\|v_i(t)\| \le r$ whenever $t \in S_i$. Then $S := \cap_{i=1}^\infty S_i$ has a full Lebesgue measure. Without any loss of generality assume that $\|C(t)\| \le \nu$ and $\bar{u}(t)$ is defined whenever $t \in S$. Fix any $t \in S$. Define a mapping $\mathcal{F}_t(z,v) := \bar{f}(t) + C(t)\, z + D(t)\, v - U_{ad}$, $(z,v) \in \mathbb{R}^n \times \mathbb{R}^m$. For every $i \in \mathbb{N}$ we have $w_i \in \mathcal{F}_t(rI\!B_{\mathbb{R}^n} \times rI\!B_{\mathbb{R}^m})$. Hence the image of $rI\!B_{\mathbb{R}^n} \times rI\!B_{\mathbb{R}^m}$ under $\mathcal{F}_t$ (having a closed convex graph) is dense in $I\!B_\beta(0)$, and consequently applying Robinson-Ursescu theorem [15, Theorem 6.22] we get that $\mathcal{F}_t$ is regular at $(0,0)$ for 0 with modulus r/β. In particular, the regularity modulus does not depend on the choice of $t \in S$. Let Λ be the set in Theorem 4.1. Fix any $(x,u) \in \operatorname{cl}\Lambda$. Let

$$\mathcal{F}_{x,u}(z,v) := f(x,u) + \nabla_x f(x,u)z + \nabla_u f(x,u)v - U_{ad}, \quad (z,v) \in \mathbb{R}^n \times \mathbb{R}^m.$$

Then $0 \in \mathcal{F}_{x,u}(0,0)$ since f is continuous and U_{ad} is closed. Since $\nabla_x f$ and $\nabla_u f$ are continuous, the uniformity of the regularity moduli of mappings $\mathcal{F}_t$ and the Lyusternik-Graves theorem imply that $\mathcal{F}_{x,u}$ is regular at $(0,0)$ for 0. Thus the mapping $\mathcal{F}'_{x,u}(z,v) := \mathcal{F}_{x,u}(z, v-u)$, $(z,v) \in \mathbb{R}^n \times \mathbb{R}^m$, is regular at $(0,u)$ for 0. Since $w \in \mathcal{F}'_{x,u}(z,v)$ if and only if $w - \nabla_x f(x,u)z \in \mathcal{G}_{x,u}(v)$, where $\mathcal{G}_{x,u}$ is the mapping in (32) with $F \equiv -U_{ad}$, we conclude that $\mathcal{G}_{x,u}$ is regular at u for 0. Theorem 4.1 implies that (v) holds.

Suppose that (v) holds. We shall establish (ii) and the theorem will be proved. Assume without any loss of generality that

$$\sup\{\|A(t)\|, \|B(t)\|, \|C(t)\|, \|D(t)\|, \|\bar{u}(t)\|, \|\bar{x}(t)\|\} \le \nu \quad \text{for each} \quad t \in S.$$

Theorem 4.1 implies that there are positive constants a, b and κ such that for any $(x,u) \in \operatorname{cl}\Lambda$, with $\Lambda := \cup_{t\in S}(\bar{x}(t), \bar{u}(t))$, the mapping

$$\mathcal{G}_{x,u}(v) := f(x,u) + \nabla_u f(x,u)(v-u) - U_{ad}, \quad v \in \mathbb{R}^m,$$

is regular at u for 0 with the constant κ and neighborhoods $I\!B_a(u)$ and $I\!B_b(0)$. Pick $\ell > \kappa$ and then $\beta \in (0, \min\{a/\ell, b\}/2)$. Let $\Omega := I\!B_\beta(0) \times \operatorname{cl}\Lambda$ and consider a mapping

$$\Omega \ni (y,x,u) \longmapsto \Sigma(y,x,u) := \mathcal{G}^{-1}_{x,u}(y) \cap I\!B_{\ell\|y\|}(u) \subset \mathbb{R}^m.$$

Given $w := (y, x, u) \in \Omega$, the regularity of $\mathcal{G}_{x,u}$ at u for 0 implies that there is $v \in \mathcal{G}_{x,u}^{-1}(y)$ such that $\|u - v\| \leq \ell\|y\|$ (with the strict inequality when $y \neq 0$), which means that $v \in \Sigma(w)$. The set U_{ad} is both closed and convex hence so is $\mathcal{G}_{x,u}^{-1}(y)$, and consequently also $\Sigma(w)$. We showed that $\operatorname{dom}\Sigma = \Omega$ and Σ has closed convex values.

Since $\Sigma(w) \subset I\!B_{\ell\|y\|}(u)$ for any $w \in \Omega$ and $\Sigma(0, \bar{x}, \bar{u}) = \{\bar{u}\}$ for each $(\bar{x}, \bar{u}) \in \operatorname{cl}\Lambda$, the mapping Σ is continuous at any point of the set $\Omega_0 := \{0\} \times \operatorname{cl}\Lambda$. We will show that Σ is inner semi-continuous on $\Omega \setminus \Omega_0$. To see this fix an arbitrary $\bar{w} = (\bar{y}, \bar{x}, \bar{u}) \in \Omega \setminus \Omega_0$ and then any $\bar{v} \in \Sigma(\bar{y}, \bar{x}, \bar{u})$. Let $\mathcal{O}_{\bar{v}}$ be any open set containing $\bar{v}$.

First, assume that $\|\bar{v} - \bar{u}\| < \ell\|\bar{y}\|$. As $\bar{v} \in I\!B_{\ell\|\bar{y}\|}(\bar{u}) \subset I\!B_{a/2}(\bar{u})$ and $\bar{y} \in I\!B_{\beta}(0) \subset I\!B_{b/2}(0)$ the mapping $\mathcal{G}_{\bar{x},\bar{u}}$ is regular at $\bar{v}$ for $\bar{y}$ with the constant κ (cf. Corollary 2.3). Thus the mapping $\Phi := \mathcal{G}_{\bar{x},\bar{u}}(\cdot) - \bar{y}$ is regular at $\bar{v}$ for 0 with the same constant. Define the function g for each $w = (y, x, u) \in \Omega$ and each $v \in \mathbb{R}^m$ by

$$g(w, v) := f(x, u) + \nabla_u f(x, u)(v - u) - y - f(\bar{x}, \bar{u}) - \nabla_u f(\bar{x}, \bar{u})(v - \bar{u}) + \bar{y}.$$

Let $\mathcal{S}(w) := \{v \in \mathbb{R}^m \mid 0 \in \mathcal{G}_{x,u}(v) - y = \Phi(v) + g(w, v)\}$, $w = (y, x, u) \in \Omega$. The continuity of $\nabla_u f$ and the implicit form of the Lyusternik-Graves theorem [13, Theorem 5E.5] imply that there are positive constants $\lambda_{\bar{w}}$ and $\delta_{\bar{w}}$ such that

$$\mathcal{S}(w') \cap I\!B_{\delta_{\bar{w}}}(\bar{v}) \subset \mathcal{S}(w) + \lambda_{\bar{w}}\|w - w'\|I\!B_{\mathbb{R}^m} \quad \text{whenever} \quad w, w' \in I\!B_{\delta_{\bar{w}}}(\bar{w}) \cap \Omega.$$

As $\mathcal{S}(\bar{w}) = \Phi^{-1}(0) \ni \bar{v}$, taking $w' := \bar{w}$ we get a function $s : I\!B_{\delta_{\bar{w}}}(\bar{w}) \cap \Omega \to \mathbb{R}^m$ such that

$$y \in \mathcal{G}_{x,u}\big(s(w)\big) \quad \text{and} \quad \|s(w) - \bar{v}\| \leq \lambda_{\bar{w}}\|w - \bar{w}\| \quad \text{for each} \quad w = (y, x, u) \in I\!B_{\delta_{\bar{w}}}(\bar{w}) \cap \Omega.$$

As $\|\bar{v} - \bar{u}\| < \ell\|\bar{y}\|$ and the function s is continuous at $\bar{w}$ with $s(\bar{w}) = \bar{v}$, there is a neighborhood $\mathcal{O}_{\bar{w}}$ of $\bar{w} = (\bar{y}, \bar{x}, \bar{u})$ with $\mathcal{O}_{\bar{w}} \subset I\!B_{\delta_{\bar{w}}}(\bar{w})$ such that

$$s(w) \in \mathcal{O}_{\bar{v}} \quad \text{and} \quad \|s(w) - u\| < \ell\|y\| \quad \text{for each} \quad w = (y, x, u) \in \mathcal{O}_{\bar{w}} \cap \Omega.$$

Consequently, $s(w) \in \mathcal{G}_{x,u}^{-1}(y) \cap I\!B_{\ell\|y\|}(u) \cap \mathcal{O}_{\bar{v}} = \Sigma(w) \cap \mathcal{O}_{\bar{v}}$ for each $w = (y, x, u) \in \mathcal{O}_{\bar{w}} \cap \Omega$. So $\Sigma(w) \cap \mathcal{O}_{\bar{v}} \neq \emptyset$ for each $w \in \mathcal{O}_{\bar{w}} \cap \Omega$.

On the other hand, if $\|\bar{v} - \bar{u}\| = \ell\|\bar{y}\|$ then find $\hat{v} \in \Sigma(\bar{w})$ with $\|\hat{v} - \bar{u}\| < \ell\|\bar{y}\|$ (which exists as we have seen right after the definition of Σ). Since the set $\Sigma(\bar{w})$ is convex and contains both $\hat{v}$ and $\bar{v}$, there exists $\tilde{v} \in \Sigma(\bar{w}) \cap \mathcal{O}_{\bar{v}}$ such that $\|\tilde{v} - \bar{u}\| < \ell\|\bar{y}\|$. By the previous case, there is a neighborhood $\mathcal{O}_{\bar{w}}$ of $\bar{w}$ such that $\Sigma(w) \cap \mathcal{O}_{\bar{v}} \neq \emptyset$ for every $w \in \mathcal{O}_{\bar{w}} \cap \Omega$.

In both the cases we showed that Σ is inner semi-continuous at $(\bar{w}, \bar{v})$. Hence Σ is inner semi-continuous on whole of Ω. Michael selection theorem [13, Theorem 5J.5] yields a continuous mapping σ such that

$$\sigma(y, x, u) \in \mathcal{G}_{x,u}^{-1}(y) \quad \text{and} \quad \|\sigma(y, x, u) - u\| \leq \ell\|y\| \quad \text{for each} \quad (y, x, u) \in I\!B_{\beta}(0) \times \operatorname{cl}\Lambda.$$

Let $c \in (0, \beta/(\nu + 1))$ and $\Omega_c := \{(z, t, p) \in \mathbb{R}^{n+1+d} \mid t \in S, \|z\| \leq c, \|p\| \leq c\}$. Clearly, for each $(z, t, p) \in \Omega_c$ we have $p - C(t)z \in I\!B_{\beta}(0)$. Define the function

$$\Omega_c \ni (z, t, p) \longmapsto u(z, t, p) := \sigma(p - C(t)z, \bar{x}(t), \bar{u}(t)).$$

Then for any $t \in S$ (hence for a.e. $t \in [0,\varepsilon]$), the function $(z,p) \longmapsto u(z,t,p)$ is continuous. For every $\{(z,p) \mid (z,t,p) \in \Omega_c \text{ for some } t \in S\}$, the function $S \ni t \longmapsto u(z,t,p)$ is measurable as a composition of a continuous function and a measurable function; and

$$\|u(z,t,p) - \bar{u}(t)\| = \|u(z,t,p) - u(0,t,0)\| \le \ell(\|p\| + \nu\|z\|) \quad \text{whenever} \quad (z,t,p) \in \Omega_c.$$

Choose $\Delta > 0$ such that

$$\Delta\varepsilon(1+\ell\nu)e^{\nu(1+\ell\nu)\varepsilon} < c. \tag{33}$$

Fix arbitrary functions $p(\cdot) \in \mathcal{P}$ and $r(\cdot) \in \mathcal{R}$ with $\|p(\cdot)\|_\infty < \Delta$ and $\|r(\cdot)\|_\infty < \Delta$. Consider the initial value problem

$$\dot{z}(t) = A(t)z(t) + B(t)(u(z(t),t,p(t)) - \bar{u}(t)) + r(t) \quad \text{for a.e.} \quad t \in [0,\varepsilon], \quad z(0) = 0. \tag{34}$$

The right-hand side of this differential equation is a Carathèodory function, and also the initial condition $z(0) = 0 \in \operatorname{int} I\!B_c(0)$. Hence there is a maximal interval $[0,\tau] \subset [0,\varepsilon]$ such that there exists a solution $z(\cdot) \in \mathcal{X}$ of (34) on $[0,\tau]$ with values in $I\!B_c(0)$, and if $\tau < \varepsilon$ then $\|z(\tau)\| = c$. Suppose that $\tau < \varepsilon$. Then for each $t \in [0,\tau]$ we have

$$\|z(t)\| \le \int_0^t \big(\nu\|z(s)\| + \nu\ell(\Delta + \nu\|z(s)\|) + \Delta\big)\,\mathrm{d}s < \Delta\varepsilon(1+\ell\nu) + \nu(1+\ell\nu)\int_0^t \|z(s)\|\,\mathrm{d}s.$$

Applying the Grönwall lemma and using (33), we get $\|z(t)\| < \Delta\varepsilon(1+\ell\nu)e^{\nu(1+\ell\nu)\varepsilon} < c$ for each $t \in [0,\tau]$. In particular, $\|z(\tau)\| < c$, a contradiction. Hence $\tau = \varepsilon$ and there exists a solution $z(\cdot)$ of (34) on the entire interval $[0,\varepsilon]$ such that $z(t) \in \operatorname{int} I\!B_c(0)$ for each $t \in [0,\varepsilon]$. Let $v(t) := u(z(t),t,p(t)) - \bar{u}(t)$, $t \in [0,\varepsilon]$. Then $(z(\cdot),v(\cdot)) \in \mathcal{V}$, $z(0) = 0$, and

$$\begin{array}{l} \dot{z}(t) = A(t)z(t) + B(t)v(t) + r(t), \\ p(t) \in \bar{f}(t) + C(t)z(t) + D(t)v(t) - U_{ad}, \end{array} \qquad \text{for a.e.} \quad t \in [0,\varepsilon].$$

Hence $(r(\cdot),p(\cdot)) \in \mathcal{H}(z(\cdot),v(\cdot)))$. As $\mathcal{H}$ has a closed convex graph, Robinson-Ursescu theorem implies (ii). □

It seems that one can formulate a similar statement when a constant mapping $F \equiv -U_{ad}$ is replaced by a general $F : \mathbb{R}^m \to \mathbb{R}^d$ with a closed convex graph, which would be a regularity version of [5, Theorem 13]. This is out of the scope of the current work and is a subject for future research.

References

[1] S. Adly, R. Cibulka, Quantitative stability of a generalized equation. Application to non-regular electrical circuits, *J. Optim. Theory Appl.* **160** (2014) 90–110.

[2] S. Adly, R. Cibulka, H. Massias, Variational analysis and generalized equations in electronics. Stability and simulation issues, *Set-Valued Var. Anal.* **21** (2013) 333–358.

[3] S. Adly, J.V. Outrata, Qualitative stability of a class of non-monotone variational inclusions. Application in electronics, *J. Convex Anal.* **20** (2013) 43–66.

[4] R. Cibulka, *Differential Variational Inequalities. A Gentle Introduction*, Proceedings SDE 2014, University of West Bohemia in Pilsen 2016.

[5] R. Cibulka, A.L. Dontchev, M.I. Krastanov, V.M. Veliov, Metrically regular differential generalized equations, *SIAM J. Optim.* (2018), to appear.

[6] R. Cibulka, T. Roubal, Solution stability and path-following for a class of generalized equations, Lecture Notes in Economics and Mathematical Systems, Control Systems and Mathematical Methods in Economics, Springer, 2018, to appear.

[7] R. Cibulka, A.L. Dontchev, A nonsmooth Robinson's inverse function theorem in Banach spaces, *Math. Program.*, Ser. A (2016) 257–270.

[8] R. Cibulka, A.L. Dontchev, V.M. Veliov, Lyusternik–Graves theorems for the sum of a Lipschitz function and a set-valued mapping, *SIAM J. Control Optim.* **54** (2016) 3273–3296.

[9] R. Cibulka, M. Fabian, On primal regularity estimates for set-valued mappings, *J. Math. Anal. Appl.* **438** (2016) 444–464.

[10] D. Cruz-Uribe, C.J. Neugebauer, An elementary proof of error estimates for the trapezoidal rule, *Mathematics Magazine* **76** (2003) 303–306.

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

[12] A.L. Dontchev, M.I. Krastanov, R.T. Rockafellar, V.M. Veliov, An Euler–Newton continuation method for tracking solution trajectories of parametric variational inequalities, *SIAM J. Control Optim.* **51** (2013) 1823–1840.

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

[14] D. Goeleven, Existence and uniqueness for a linear mixed variational inequality arising in electrical circuits with transistors, *J. Optim. Theory Appl.* **138** (2008) 397–406.

[15] A.D. Ioffe, *Variational Analysis of Regular Mappings. Theory and Applications,* Springer International Publishing 2017.

[16] W. Kelley, A. Peterson, *Theory of Differential Equations: Classical and Qualitative*, Second Edition, Springer 2010.

[17] S.M. Robinson, Newton's method for a class of nonsmooth functions, *Set-Valued Analysis* **2** (1994) 291–305.

Metric regularity properties in bang-bang type linear-quadratic optimal control problems*

J. Preininger†, T. Scarinci‡, and V.M. Veliov§

Abstract

The paper investigates the Lipschitz/Hölder stability with respect to perturbations of optimal control problems with linear dynamic and cost functional which is quadratic in the state and linear in the control variable. The optimal control is assumed to be of bang-bang type and the problem to enjoy certain convexity properties. Conditions for bi-metric regularity and (Hölder) metric sub-regularity are established, involving only the order of the zeros of the associated switching function and smoothness of the data. These results provide a basis for the investigation of various approximation methods. They are utilized in this paper for the convergence analysis of a Newton-type method applied to optimal control problems which are affine with respect to the control.

Key words: variational analysis, optimal control, linear control systems, bang-bang controls, metric regularity, stability analysis, Newton's method.
AMS subject classifications: 49J30, 49K40, 49M15, 49N05, 47J07.

1 Introduction

Stability analysis of solutions is a crucial topic in optimization theory due, in particular, to its applications for obtaining error estimates of numerical approximations. Although related investigations in optimal control theory accompany its development from its early stages, the systematic analysis of (Lipschitz) stability in the area started with the works of Dontchev, Hager and Malanowski (see [10, 12]). In these papers, the authors prove Lipschitz dependence of the solutions with respect to perturbations, under a strict coercivity condition which also implies Lipschitz continuity of the optimal control.

In contrast, in the present paper we investigate a class of problems in which the control appears linearly, therefore the strict coercivity fails. Moreover, when the control set is the m-dimensional

*This research is supported by the Austrian Science Foundation (FWF) under grant No P26640-N25. The second author is also supported by the Doctoral Program "Vienna Graduate School on Computational Optimization" funded by the Austrian Science Fund (FWF), project No W1260-N35.

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

‡Dept. of Statistics and Operations Research, University of Vienna, Austria, teresa.scarinci@univie.ac.at.

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

hypercube $[-1,1]^m$, each component of the optimal control generally switches from ± 1 to ∓ 1, possibly concatenating with arcs with values in the interior of $[-1,1]$. That is, the optimal control is typically discontinuous.

Problems which are affine with respect to the control variable arise in many applications, such as engineering, biology and medicine (see e.g. [22, 21, 23, 25]). Nevertheless, only few papers address the stability analysis in case of non-coercive problems and such with discontinuous optimal controls; in fact, many relevant questions still remain unanswered. Recent progress was made in [16, 24, 18, 17] for control-affine problems and in [27] for problems with linear dynamics, and we build on these papers. We mention also the paper [29] and the references therein for problems with group sparsity. Applications to error estimates for time-discretization schemes are discussed in [32, 2, 19, 3, 29, 26] for linear systems or problems of the type (P) below. We mention also the paper [5], where stability analysis is discussed for control-affine systems with bang-singular optimal controls.

In the present paper we focus our attention on the following class of optimal control problems:

$$\begin{array}{ll} \text{minimize} & J(x,u) \\ \text{subject to} & \dot{x}(t) = A(t)x(t) + B(t)u(t) + d(t), \quad t \in [0,T], \\ & u(t) \in U := [-1,1]^m, \\ & x(0) = x_0, \end{array} \tag{P}$$

where

$$J(x,u) := g(x(T)) + \int_0^T \left(\frac{1}{2} x(t)^\top W(t) x(t) + x(t)^\top S(t) u(t) \right) dt.$$

Here, $u(t) \in U$ and $x(t) \in \mathbb{R}^n$ denote the control and the state of the system at time $t \in [0,T]$, the function $g : \mathbb{R}^n \to \mathbb{R}$ is given, as well as $A(t), W(t) \in \mathbb{R}^{n\times n}$, $B(t), S(t) \in \mathbb{R}^{n\times m}$ and $d(t) \in \mathbb{R}^n$, $t \in [0,T]$. The set of admissible controls in Problem (P), further denoted by $\mathcal{U}$, consists of all measurable functions u satisfying $u(t) \in U$ for almost every $t \in [0,T]$,

$$\mathcal{U} = \{u \in L^\infty([0,T], \mathbb{R}^m) : u(t) \in U \text{ a.e. on } [0,T]\}.$$

Linear terms in u or x are not included in the integrand, which is not a restriction of generality, since such terms can be shifted in a standard way into the differential equation.

The stability properties of the solution(s) of (P) will be analyzed through the Pontryagin minimum principle, which states that for any optimal pair $(\hat{x}, \hat{u})$, there exists an absolutely continuous function $\hat{p} : [0,T] \to \mathbb{R}^n$ such that the triple $(\hat{x}, \hat{p}, \hat{u})$ solves the following system a.e. on $[0,T]$:

$$\begin{array}{l} 0 = \dot{x}(t) - A(t)x(t) - B(t)u(t) - d(t), \\ 0 = \dot{p}(t) + A(t)^\top p(t) + W(t)x(t) + S(t)u(t), \\ 0 \in B(t)^\top p(t) + S(t)^\top x(t) + N_U(u(t)), \\ 0 = p(T) - \nabla g(x(T)). \end{array} \tag{PMP}$$

Here $N_U(u)$ is the normal cone to U at u defined in the usual way:

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

It will be assumed (see the next sections for precise formulations) that the data are smooth enough, Problem (P) satisfies some convexity-like assumptions, the (reference) optimal control is piece-wise constant with each component taking only the values -1 and 1. Moreover, it will be assumed that each component of the associated "switching function", $t \mapsto B(t)^\top p(t) + S(t)^\top x(t)$, satisfies at its zeros a certain growth condition, characterized by a number $\kappa \geq 1$ (κ can be regarded as the multiplicity of the zeros if the switching function is smooth).

We recast the system (PMP) as the generalized equation

$$0 \in F(x, p, u), \tag{1.1}$$

where F is the set-valued mapping

$$F(x, p, u) := \begin{pmatrix} \dot{x} - Ax - Bu - d \\ \dot{p} + A^\top p + Wx + Su \\ B^\top p + S^\top x + N_{\mathcal{U}}(u) \\ p(T) - \nabla g(x(T)) \end{pmatrix} \tag{1.2}$$

acting in a suitable Banach space $\mathcal{X} \ni (x, p, u)$ with values in a linear normed space $\mathcal{Y}$. The set $N_{\mathcal{U}}(u)$ in (1.2) is a functional replacement for the point-wise cones $N_U(u(t))$ in (PMP) and will be strictly defined in the next section together with the spaces $\mathcal{X}$ and $\mathcal{Y}$.

As usual, we investigate the stability of the solution of problem (P) by introducing a perturbation $y \in \mathcal{Y}$ in the system of necessary optimality conditions, that is, considering the perturbed inclusion $y \in F(x, p, u)$. Under the assumptions briefly mentioned above, the unperturbed system $0 \in F(x, p, u)$, that is the system of necessary optimality conditions (PMP), has a unique solution $(\hat{x}, \hat{p}, \hat{u})$.

Two main concepts of stability are investigated in the paper.

The first concept is a stronger version of the *Hölder strong metric sub-regularity* (see the recent paper [9]). Roughly speaking, we prove that for all sufficiently small perturbations y, the inclusion $y \in F(x, p, u)$, associated with problem (P), has a solution and all the solutions are at distance (in the space $\mathcal{X}$) at most proportional to $\|y\|^{1/\kappa}$ from the unique solution $(\hat{x}, \hat{p}, \hat{u})$ of the inclusion $0 \in F(x, p, u)$. We mention that a similar result was proved in [3, Theorem 9], but with different functional spaces and on slightly stronger assumptions. Moreover, the claim in our result is somewhat stronger, which is rather essential for the analysis of the strong bi-metric regularity and the convergence of Newton's method which will be discussed below.

The second concept extends the standard *strong metric regularity* introduced in the seminal paper [28] by Robinson (see also [13, Chapter 3.7]). The new feature is that a second metric space $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$ is involved (presumably with a non-equivalent and larger metric than that in $\mathcal{Y}$) and only disturbances from this space are considered. Roughly, strong bi-metric regularity relative to $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$ of F at $\hat{z} := (\hat{x}, \hat{p}, \hat{u})$ means that the inverse mapping $\mathcal{Y} \ni y \mapsto F^{-1}(y) = \{z \in \mathcal{X} : y \in F(z)\}$ is locally (around $\hat{z}$) single-valued when restricted to a sufficiently small ball in $\widetilde{\mathcal{Y}}$, centered at $y = 0$. Moreover, this single-valued mapping is Lipschitz continuous with respect to the metric of $\mathcal{Y}$. In the terminology of [13], this means that F has a single-valued localization in $\mathcal{X} \times \widetilde{\mathcal{Y}}$ and it is Lipschitz continuous, but the Lipschitz property holds with respect to the metric of $\mathcal{Y}$.

The general notion of strong bi-metric regularity was introduced in somewhat more restrictive form in [27], where applications to Mayer's type problems for linear control systems were in the focus. Similarly as the strong metric regularity, it has the important property to be invariant with respect to small (in an appropriate sense) functional perturbations of F. This property is often referred to as Lyusternik-Graves type theorem, see e.g. [13, Chapter 5.5]. In the present paper we prove a general Lyusternik-Graves type theorem for strong bi-metrically regular inclusions, which is a substantial improvement of the one in [27], since most of the assumptions are now formulated in terms of the (smaller) metric of $\mathcal{Y}$ rather than in the metric of $\widetilde{\mathcal{Y}}$, as in [27].

We prove strong bi-metric regularity of the mapping F associated with Problem (P), which extends the result in [27] concerning Mayer's problems. This extension is nontrivial, since, technically speaking, the integral cost introduces the state variable in the switching function, making this function nonsmooth. This forces us, among other things, to consider the present slightly more general notion of bi-metric regularity compared with the one in [27]. As an application we give a Lipschitz stability result with respect to small non-linear perturbations in the differential equation.

In the last section of the paper, we investigate the convergence of a Newton-type method (as interpreted in the context of generalized equations, see e.g. [13, Chapter 6.3]) applied to a class of control-affine problems for which (P) can be regarded as a linearization. Notice that the known convergence results (cf. [10]) are inapplicable for non-coercive problems, where the strong metric regularity in the usual space settings fails. We will give sufficient conditions under which the considered Newton's method converges, and does so quadratically. The proof is based on a strengthen version of the metric sub-regularity proved in the present paper for Problem (P). We mention that the stability analysis and the convergence properties of Newton methods still remain not fully understood when singular arcs occur. Some advances have been done recently in [17] for the first issue, and in [5, 15] for the latter. However, these issues remain as interesting topics for future research.

The paper is organized as follows. In Section 2, we recall some basic facts and introduce the main assumptions on Problem (P) together with some notations. Section 3 is devoted to the proof of the Hölder sub-regularity of Problem (P) (actually, of the associated mapping F). In Section 4, we introduce the definition of strong bi-metric regularity, and prove an extension of the Lyusternik-Graves theorem suitable to this new notion. After that, we prove the strong bi-metric regularity of the mapping F resulting from problem (P) and give a result about the invariance of this property under a class of non-linear perturbations. In Section 5, we investigate the convergence of a Newton-type method applied to some control-affine problems with bang-bang solutions.

2 Preliminaries

Throughout the paper we use the following common notations. The standard n-dimensional Euclidean space is denoted by $\mathbb{R}^n$, with the scalar product and norm denoted by $\langle\cdot,\cdot\rangle$ and $|\cdot|$, respectively. The superscript $^\top$ denotes transposition. Further, $L^1([0,T],\mathbb{R}^n)$ and $L^\infty([0,T],\mathbb{R}^n)$ are the spaces of all measurable and absolutely integrable, respectively essentially bounded, functions with the corresponding norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$, which sometimes will be abbreviated as L^1 and L^∞, respectively. Moreover, $W^{1,k}([0,T],\mathbb{R}^n)$ is the space of all absolutely continuous functions from $[0,T]$ to $\mathbb{R}^n$

whose first derivatives belonging to L^k, $k \in \{1, \infty\}$. The corresponding norms are denoted by $\|\cdot\|_{1,1}$ and $\|\cdot\|_{1,\infty}$, respectively. We also denote $W^{1,1}_{x_0}([0,T],\mathbb{R}^n) := \{x \in W^{1,1}([0,T]),\mathbb{R}^n) : x(0) = x_0\}$.

We introduce the following assumptions, some of which will be strengthened in the next sections.

Assumption (A1). The matrix-functions B and S are continuous, A, W and d are measurable and bounded. The matrix $W(t)$ is symmetric for every $t \in [0,T]$. The function g is differentiable with globally Lipschitz continuous gradient ∇g.

We stress that the assumption about *global* Lipschitz continuity of ∇g is made for technical convenience only and is not a real restriction. Since the reachable set in Problem (P) is compact, any modification of g outside a neighborhood of the reachable set does not affect the problem.

For every $u \in \mathcal{U}$ the differential equation in problem (P) with the given initial condition has a unique (absolutely continuous) solution x on $[0,T]$. Every such pair (x,u) is called "admissible", and the set of all admissible pairs is denoted by $\mathcal{F}$.

Thanks to Assumption (A1), a standard compactness argument implies the existence of an optimal solution of Problem (P). In what follows we consider a fixed optimal solution $(\hat{x},\hat{u})$.

Assumption (A2). For every admissible pair $(x,u) \in \mathcal{F}$ it holds that

$$\langle \nabla g(x(T)) - \nabla g(\hat{x}(T)), \Delta x(T)\rangle + \int_0^T (\langle W(t)\Delta x, \Delta x\rangle + 2\langle S(t)\Delta u, \Delta x\rangle)dt \geq 0,$$

where $\Delta x(T) := x(T) - \hat{x}(T)$, $\Delta x := x(t) - \hat{x}(t)$ and $\Delta u := u(t) - \hat{u}(t)$.

Let $\hat{p}$ be a co-state function for $(\hat{x},\hat{u})$, i.e. $(\hat{x},\hat{p},\hat{u})$ solves (PMP). We recall that

$$\hat{\sigma} := B^\top \hat{p} + S^\top \hat{x}$$

is the so-called *switching function* corresponding to the triple $(\hat{x},\hat{p},\hat{u})$. For every $j \in \{1,\dots,m\}$ denote by $\hat{\sigma}_j$ its j-th component. Notice that $\hat{\sigma}$ is continuous due to Assumption (A1).

In the next assumption we postulate that the optimal control $\hat{u}$ is *strictly bang-bang*, with a finite number of *switching times* on $[0,T]$, and that the switching function exhibits a certain growth in a neighborhood of any zero.

Assumption (A3). 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 $\hat{\sigma}_j(s) = 0$ we have

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

A similar assumption is introduced in [16] in the case $\kappa = 1$ and in [27, 30] for $\kappa \geq 1$. The set $\mathcal{U}$ of admissible controls will be considered as a metric space with the metric induced by the L^1-norm. For this reason we define

$$\mathcal{X} := W^{1,1}_{x_0}([0,T],\mathbb{R}^n) \times W^{1,1}([0,T],\mathbb{R}^n) \times L^1([0,T],\mathbb{R}^m),$$

with the usual norm: for $(x,p,u) \in \mathcal{X}$,

$$\|(x,p,u)\| := \|x\|_{1,1} + \|p\|_{1,1} + \|u\|_1.$$

Next, we denote by $\mathcal{Y}$ the space

$$\mathcal{Y} := L^1([0,T],\mathbb{R}^n) \times L^1([0,T],\mathbb{R}^n) \times L^\infty([0,T],\mathbb{R}^m) \times \mathbb{R}^n, \tag{2.1}$$

with the usual norm: for $(\xi,\pi,\rho,\nu) \in \mathcal{Y}$,

$$\|(\xi,\pi,\rho,\nu)\| := \|\xi\|_1 + \|\pi\|_1 + \|\rho\|_\infty + |\nu|.$$

We denote by $d_{\mathcal{Y}}$ the distance induced by $\|\cdot\|$.

As in the introduction, we recast the first order optimality conditions (Pontryagin system) (PMP) for Problem (P) as the generalized equation

$$0 \in F(x,p,u), \tag{2.2}$$

where $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is defined in (1.2). The normal cone $N_{\mathcal{U}}(u)$ appearing there is defined in the standard way: for $u \in L^1([0,T],\mathbb{R}^m)$,

$$N_{\mathcal{U}}(u) := \{v \in L^\infty([0,T],\mathbb{R}^m) : v(t) \in N_U(u(t)) \text{ for a.e. } t \in [0,T]\}.$$

Notice that this definition is consistent with the general definition of a normal cone if $\mathcal{U}$ is considered as a subset of the space L^1 (although $\mathcal{U}$ is also contained in L^∞; but then $N_{\mathcal{U}}(u)$ should be a cone in the dual space to L^∞).

In the following sections, given a perturbation $y = (\xi,\pi,\rho,\nu) \in \mathcal{Y}$, we will study the inclusion

$$y \in F(x,p,u), \tag{2.3}$$

which, written in detail, looks as follows: for a.e. $t \in [0,T]$,

$$\begin{aligned}
0 &= \dot{x}(t) - A(t)x(t) - B(t)u(t) - d(t) - \xi(t),\\
0 &= \dot{p}(t) + A(t)^\top p(t) + W(t)x(t) + S(t)u(t) - \pi(t),\\
0 &\in B(t)^\top p(t) + S(t)^\top x(t) - \rho(t) + N_U(u(t)),\\
0 &= p(T) - \nabla g(x(T)) - \nu.
\end{aligned} \tag{2.4}$$

3 Strong metric sub-regularity

In this section we prove an important regularity property of the mapping F defined in (1.2), related to, but stronger than, *strong Hölder metric sub-regularity*, see [9].

We begin with some important properties of switching functions that fulfill Assumption (A3) . First we fix some notations. Given any continuous function $\sigma : [0,T] \to \mathbb{R}^m$ (σ_j will denote its j-th component) satisfying Assumption (A3) with constants κ, α and τ, and a real number $\delta > 0$, we define

$$I_j(\sigma,\delta) := \bigcup_{s\in[0,T]:\,\sigma_j(s)=0} (s-\delta, s+\delta) \cap [0,T], \qquad I(\sigma,\delta) := \bigcup_{1\le j\le m} I_j(\sigma,\delta),$$

and

$$l_{\min}(\sigma,\delta) := \min_{1\le j\le m} \min_{t\in[0,T]\setminus I_j(\sigma,\delta)} |\sigma_j(t)| > 0. \tag{3.1}$$

Note that this minimum always exists and is indeed positive since σ is continuous and $[0,T]\setminus I_j(\sigma,\tau)$ is compact for any $j\in\{1,\dots,m\}$.

Now we state an auxiliary result which presents an inverse integral inequality for functions satisfying Assumption (A3) of the type of those developed in Theorem 2.1 and Corollary 2.1 and 2.2 in [31]. It extends [30, Lemma 1.3], which in its turn originates from [16, Lemma 3.3].

Lemma 3.1. *Let $\sigma:[0,T]\to\mathbb{R}^m$ be any continuous function satisfying Assumption (A3). Then there exists a constant $c_0>0$ such that*

$$\|v\|_\infty^\kappa\int_0^T\sum_{j=1}^m|\sigma_j(t)v_j(t)|\,dt\geq c_0\|v\|_1^{\kappa+1}\quad\textit{for any } v\in L^\infty([0,T],\mathbb{R}^m). \tag{3.2}$$

Remark 3.2. Carefully following the proof below we can establish that the constant c_0 in the lemma only depends on the numbers κ, α, τ and $l_{\min}(\sigma,\tau)$. Thus Lemma 3.1 can be reformulated in the following more precise form: for any given positive real numbers $\kappa\geq 1$, $\alpha,\tau>0$ and $m_0>0$ there exists a constant $c_0>0$ such that the claim (3.2) holds for any continuous function $\sigma:[0,T]\to\mathbb{R}^m$ satisfying Assumption (A3) with constants κ, α, τ, and with $l_{\min}(\sigma,\tau)\geq m_0$.

Proof. If $v=0$, then the inequality in Lemma 3.1 is fulfilled. If $v\neq 0$ then due to the homogeneity of order $\kappa+1$ of the two sides of (3.2) with respect to v, it is enough to prove the lemma in the case of $\|v\|_\infty=1$, which will be assumed in the remaining part of the proof.

Now we choose $\bar\delta\in(0,\tau)$ such that $\alpha\bar\delta^\kappa<l_{\min}(\sigma,\tau)$. Then for all $\delta\in(0,\bar\delta]$ and $j\in\{1,\dots,m\}$ we have

$$|\sigma_j(t)|\geq\alpha\delta^\kappa\quad\forall t\in[0,T]\setminus I(\sigma,\delta). \tag{3.3}$$

Indeed, if $t\in I_j(\sigma,\tau)\setminus I(\sigma,\delta)$ then inequality (3.3) follows from (A3) and if $t\notin I_j(\sigma,\tau)$ then $|\sigma_j(t)|\geq l_{\min}(\sigma,\tau)>\alpha\bar\delta^\kappa\geq\alpha\delta^\kappa$.

Using (3.3) we obtain that

$$\begin{aligned}\varphi(v):=\int_0^T\sum_{j=1}^m|\sigma_j(t)v_j(t)|\,dt&\geq\int_{[0,T]\setminus I(\sigma,\delta)}\sum_{j=1}^m|\sigma_j(t)v_j(t)|\,dt\\ \geq\alpha\delta^\kappa\sum_{j=1}^m\int_{[0,T]\setminus I(\sigma,\delta)}|v_j(t)|\,dt&\geq\alpha\delta^\kappa\left(\|v\|_1-\sum_{j=1}^m\int_{I(\sigma,\delta)}|v_j(t)|\,dt\right)\geq\alpha\delta^\kappa(\|v\|_1-2\lambda\delta),\end{aligned}$$

where λ is the sum of the number of zeros of σ_j for all $j\in\{1,\dots,m\}$. (Notice that Assumption (A3) implies $\lambda\leq\frac{mT}{2\tau}+m$.) If $\|v\|_1\geq 4\lambda\bar\delta$, we choose $\delta:=\bar\delta$ to get

$$\varphi(v)\geq\frac{\alpha\bar\delta^\kappa}{2}\|v\|_1$$

and since $\|v\|_1\leq T\|v\|_\infty=T$ we have that $\varphi(v)\geq\frac{\alpha\bar\delta^\kappa}{2T^\kappa}\|v\|_1^{\kappa+1}$. If $\|v\|_1\leq 4\lambda\bar\delta$, we choose $\delta:=\frac{\|v\|_1}{4\lambda}\leq\bar\delta$ to get

$$\varphi(v)\geq\frac{\alpha}{2^{2\kappa+1}\lambda^\kappa}\|v\|_1^{\kappa+1}.$$

Hence, by defining $c_0 := \min\left\{\frac{\alpha\bar{\delta}^\kappa}{2T^\kappa}, \frac{\alpha}{2^{2\kappa+1}\lambda^\kappa}\right\}$ we obtain that

$$\varphi(v) \geq c_0\|v\|_1^{\kappa+1},$$

which proves (a).

Since we can choose $\bar{\delta}$ to only depend on κ, α, τ and m_0 and there is an upper bound to λ which only depends on m, T and τ, the constant c_0 also only depends on m, T, κ, α, τ and m_0. This proves Remark 3.2. Q.E.D.

The following theorem establishes a stability property of the mapping F associated with system (PMP) which is a somewhat stronger form of the well known property of *metric sub-regularity*, [13, Section 3I]. It extends [3, Theorem 8] in that Assumption (A3) is weaker than the corresponding assumption there (since we allow 0 and T to be feasible zeros of some components of the switching function), the norm in the space $\mathcal{Y}$ is somewhat weaker, and the function g is not necessarily quadratic. Most importantly, the size of the disturbance y for which the claim of the theorem holds is not a priori restricted (as in the definition of metric sub-regularity, [13, Section 3H] and in [3, Theorem 8]).

Theorem 3.3. *Let $(\hat{x}, \hat{p}, \hat{u})$ be a solution of* (PMP) *such that Assumptions (A1)–(A3) are fulfilled. Then for any $b > 0$ there exists $c > 0$ such that for any $y \in \mathcal{Y}$ with $\|y\| \leq b$, there exists a triple $(x, p, u) \in \mathcal{X}$ solving $y \in F(x, p, u)$, and any such triple satisfies*

$$\|x - \hat{x}\|_{1,1} + \|p - \hat{p}\|_{1,1} + \|u - \hat{u}\|_1 \leq c\|y\|^{\frac{1}{\kappa}}. \tag{3.4}$$

Remark 3.4. Due to further needs, in the proof of the above theorem we will care about how the constant c depends on the data of the problem and the associated switching function $\hat{\sigma}$. More precisely, the following statement will be proved.

Let the natural numbers n, m and the real number $T > 0$ be fixed. Given constants $\kappa \geq 1$, $\alpha > 0$, $\tau > 0$, $m_0 > 0$, $b > 0$ and K, there exists a number $c > 0$ with the following property[1].

Let the $(n \times n)$-matrix functions $A(t)$ and $W(t)$ the $(n \times m)$-matrix functions $B(t)$ and $S(t)$ be defined on $[0, T]$, and $g : \mathbb{R}^n \to \mathbb{R}$ be such that Assumption (A1) is fulfilled, and in addition,

$$\|A\|_\infty,\ \|B\|_\infty,\ \|W\|_\infty,\ \|S\|_\infty,\ \|d\|_\infty,\ \leq K,\ \ \nabla g \text{ is Lipschitz with constant } K. \tag{3.5}$$

Let $(\hat{x}, \hat{p}, \hat{u})$ be a solution of (PMP) *(i.e. of* (1.1)*) such that Assumption (A2) holds, the corresponding switching function $\hat{\sigma}$ fulfills Assumption (A3) with constants κ, α and τ, and $l_{\min}(\hat{\sigma}, \tau) \geq m_0$. Then for every $y \in \mathcal{Y}$ with $\|y\| \leq b$ the inclusion $y \in F(x, p, u)$ (with F defined in* (1.2)*) has a solution and for every solution (x, p, u) the estimation* (3.4) *holds.*

Proof. First of all, we note that the inclusion $y \in F(x, p, u)$, for any $y = (\xi, \pi, \rho, \nu) \in \mathcal{Y}$, represents the system of necessary optimality conditions of the following problem:

$$\min\left\{g(x(T)) - \nu^\top x(T) + \int_0^T \Big(\frac{1}{2}x(t)^\top W(t)x(t) + x(t)^\top S(t)u(t) - \rho^\top(t)u(t) - \pi^\top(t)x(t)\Big)dt\right\} \tag{3.6}$$

[1] If $\kappa = 1$, then the constant c can be chosen independent of b.

subject to

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

Due to the linearity in u and the convexity and compactness of the constraining set U this problem has a solution, hence also the inclusion $y \in F(x,p,u)$.

Now, let $b > 0$ be arbitrarily chosen and let (x,p,u) be a solution of $y \in F(x,p,u)$, where $y = (\xi,\pi,\rho,\nu) \in \mathcal{Y}$ and $\|y\| \le b$. The following notations will be used. As before, $\hat{\sigma}(t) := B(t)^\top \hat{p}(t) + S(t)^\top \hat{x}(t)$, while $\sigma(t) := B(t)^\top p(t) + S(t)^\top x(t) - \rho(t)$. Furthermore, we denote $\Delta x(t) := x(t) - \hat{x}(t)$, $\Delta p(t) = p(t) - \hat{p}(t)$, $\Delta u(t) := u(t) - \hat{u}(t)$ and $\Delta\sigma(t) := \sigma(t) - \hat{\sigma}(t)$ and skip the argument t whenever this does not lead to ambiguity.

Integrating by parts, we have

$$\int_0^T \langle \Delta\dot{p}, \Delta x\rangle \, dt = \langle \Delta p(T), \Delta x(T)\rangle - \int_0^T \langle \Delta p, \Delta\dot{x}\rangle \, dt.$$

Substituting here the expressions for Δx and Δp resulting from the inclusions $y \in F(x,p,u)$ and $0 \in F(\hat{x},\hat{p},\hat{u})$ in view of (1.2) we obtain that

$$\int_0^T \langle -A^\top \Delta p - W\Delta x - S\Delta u + \pi, \Delta x\rangle \, dt$$
$$= \langle \nabla g(x(T)) - \nabla g(\hat{x}(T)) + \nu, \Delta x(T)\rangle - \int_0^T \langle \Delta p, A\Delta x + B\Delta u + \xi\rangle \, dt.$$

Rearranging the terms in this equality and using Assumption (A2) we get

$$\int_0^T (\langle \Delta p, B\Delta u\rangle + \langle S\Delta u, \Delta x\rangle) \, dt + \int_0^T (\langle \pi, \Delta x\rangle + \langle \xi, \Delta p\rangle) \, dt - \langle \nu, \Delta x(T)\rangle$$
$$= \langle \nabla g(x(T)) - \nabla g(\hat{x}(T)), \Delta x(T)\rangle + \int_0^T (\langle W\Delta x, \Delta x\rangle + 2\langle S\Delta u, \Delta x\rangle) \, dt \ge 0.$$

Using this inequality and the definitions of the functions σ and $\hat{\sigma}$ we obtain

$$\int_0^T \langle \Delta\sigma, \Delta u\rangle \, dt = \int_0^T \langle B^\top \Delta p + S^\top \Delta x - \rho, \Delta u\rangle \, dt \ge$$
$$\ge \int_0^T (-\langle \pi, \Delta x\rangle - \langle \xi, \Delta p\rangle - \langle \rho, \Delta u\rangle) \, dt + \langle \nu, \Delta x(T)\rangle. \tag{3.7}$$

The third component of the inclusion $y \in F(x,p,u)$ reads as $-\sigma(t) \in N_U(u(t))$, which implies $\langle -\sigma(t),\, \hat{u}(t) - u(t)\rangle \le 0$. Then

$$-\int_0^T \langle \Delta\sigma, \Delta u\rangle \, dt = \int_0^T [-\langle \sigma, \Delta u\rangle + \langle \hat{\sigma},\, \Delta u\rangle] \, dt \ge \int_0^T \langle \hat{\sigma},\, \Delta u\rangle \, dt.$$

From here, using that $-\hat{\sigma}_j(t) \in N_{[-1,1]}(\hat{u}_j(t))$, hence $\hat{\sigma}_j(t)\,\Delta u_j(t) \ge 0$ for each j, Lemma 3.1 implies that

$$-\int_0^T \langle \Delta\sigma, \Delta u\rangle \, dt \ge \int_0^T \sum_{j=1}^m |\hat{\sigma}_j\,\Delta u_j| \, dt \ge c_0 \|\Delta u\|_1^{\kappa+1},$$

where the constant c_0 only depends on κ, α, τ and m_0 (see Remark 3.4). Then using (3.7) and the Hölder inequality we obtain

$$\|\pi\|_1 \, \|\Delta x\|_\infty + \|\xi\|_1 \, \|\Delta p\|_\infty + |\nu| \, |\Delta x(T)| + \|\rho\|_\infty \, \|\Delta u\|_1 \geq c_0 \|\Delta u\|_1^{\kappa+1}. \tag{3.8}$$

Using Assumption (A1) and the solution formula of the Cauchy problem for Δx and Δp we get

$$\|\Delta x\|_\infty \leq c_1(\|\xi\|_1 + \|\Delta u\|_1), \quad \|\Delta p\|_\infty \leq c_2(\|\xi\|_1 + \|\pi\|_1 + \|\Delta u\|_1 + |\nu|) \tag{3.9}$$

for some constants c_1 and c_2 that only depend on K (see (3.5) in Remark 3.4). (We mention that for the estimation of $\|\Delta p\|_\infty$ we use the estimation for $|\Delta x(T)|$ and the Lipschitz continuity of the gradient ∇g appearing in the end-point conditions for p and $\hat{p}$ in (1.1).) Therefore, by (3.8)–(3.9) we obtain that

$$(\|y\|^2 + \|y\| \|\Delta u\|_1) \geq c_3 \|\Delta u\|_1^{\kappa+1} \tag{3.10}$$

for some constant c_3, only depending on c_0, c_1 and c_2. Now, we distinguish two cases. First, if $\|y\| \leq \|\Delta u\|_1$ then

$$2\|y\| \|\Delta u\|_1 \geq c_3 \|\Delta u\|_1^{\kappa+1},$$

which implies

$$\|\Delta u\|_1 \leq \left(\frac{2}{c_3}\|y\|\right)^{1/\kappa}. \tag{3.11}$$

Otherwise, if $\|\Delta u\|_1 \leq \|y\| \leq b$ then

$$\|\Delta u\|_1 \leq \|y\|^{1/\kappa} \|y\|^{(\kappa-1)/\kappa} \leq b^{(\kappa-1)/\kappa} \|y\|^{1/\kappa}. \tag{3.12}$$

Inequalities (3.11) and (3.12) imply that for any $b > 0$ there exists $c_4 > 0$, depending on c_3 and b such that

$$\|\Delta u\|_1 \leq c_4 \|y\|^{1/\kappa}.$$

Then the claim of the theorem follows with a suitable constant c (depending only on c_1, c_2 and c_4) from the above estimation together with (3.9).

Notice that c_4, hence also c, depend on b only due to the term $b^{(\kappa-1)/\kappa}$ in estimation (3.12), which equals 1 in the case $\kappa = 1$. This justifies Footnote 1. Q.E.D.

Remark 3.5. Clearly, the property established in Theorem 3.3 implies that $(\hat{x}, \hat{p}, \hat{u})$ is the unique solution of (PMP), thus $(\hat{x}, \hat{u})$ is the unique solution of problem (P). Therefore, (PMP), together with Assumptions (A1)–(A3), is a sufficient optimality condition.

4 Bi-metric regularity

The notion of strong bi-metric regularity was introduced in [27] in order to grasp in a relevant way the dependence on perturbations of the solutions of Mayer's type optimal control problems for linear systems. Its extension to the Bolza problem considered in this paper is more complicated due to the missing smoothness of the switching function associated with the optimal control. In this section we present such an extension, starting from the abstract definition of strong bi-metric regularity and a new, substantially strengthened version of the Lyusternik-Graves type theorem proved in [27].

4.1 The abstract setting

First, we give the definition of strong bi-metric regularity, which is a more convenient extension of the one introduced in [27].

Let (X, d_X), (Y, d_Y) and $(\widetilde{Y}, \widetilde{d}_Y)$ be metric spaces, with $\widetilde{Y} \subset Y$ and $d_Y \le \widetilde{d}_Y$ on $\widetilde{Y}$. Denote by $B_X(\bar{x}; a)$ and $B_{\widetilde{Y}}(\bar{y}; b)$ the closed balls in the metric spaces (X, d_X) and $(\widetilde{Y}, \widetilde{d}_Y)$ with radius $a > 0$ and $b > 0$ centered at $\bar{x}$ and $\bar{y}$, respectively. We will suppose that the metric d_Y and $\tilde{d}_Y$ are *shift-invariant*, which means, in terms of the metric d_Y, that

$$d_Y(y + z, y' + z) = d_Y(y, y'), \quad \forall y, y', z \in Y.$$

Definition 4.1. The map $\Phi : X \rightrightarrows Y$ is strongly bi-metrically regular relative to $\widetilde{Y} \subset Y$ at $\bar{x} \in X$ for $\bar{y} \in \widetilde{Y}$ with constants $\varsigma \ge 0$, $a > 0$ and $b > 0$ if $(\bar{x}, \bar{y}) \in \mathrm{graph}(\Phi)$ and the following properties are fulfilled:

1. the mapping $B_{\widetilde{Y}}(\bar{y}; b) \ni y \mapsto \Phi^{-1}(y) \cap B_X(\bar{x}; a)$ is single-valued, and
2. for all $y, y' \in B_{\widetilde{Y}}(\bar{y}; b)$,

$$d_X(\Phi^{-1}(y) \cap B_X(\bar{x}; a), \Phi^{-1}(y') \cap B_X(\bar{x}; a)) \le \varsigma d_Y(y, y'). \tag{4.1}$$

It is important to notice that in this definition the "disturbances" y, y' are taken from the smaller space $\widetilde{Y}$ (and are sufficiently small in the metric of this space), but the Lipschitz property (4.1) holds with the (smaller) metric d_Y. This is the crucial difference with the standard definition of strong metric regularity (see e.g. [13, Section 3G] and [20]), where the spaces Y and $\widetilde{Y}$ coincide.

The next result resembles the main features of the Lyusternik-Graves-type theorem proved in [27, Theorem 2.1], but under substantially weakened requirements, as explained in the comments after the proof.

Theorem 4.2. *Let X be a complete metric space, Y be a linear space, $\widetilde{Y}$ be a subspace of Y, and let both metrices, d_Y in Y and $\widetilde{d}_Y$ in $\widetilde{Y}$, be shift-invariant and $d_Y \le \widetilde{d}_Y$ on $\widetilde{Y}$. Let the set-valued map $\Phi : X \rightrightarrows Y$ be strongly bi-metrically regular at $\bar{x}$ for $\bar{y}$ with constants ς, a, b. Let $\mu > 0$ and ς' be such that $\varsigma\mu < 1$ and $\varsigma' \ge \varsigma/(1 - \varsigma\mu)$. Then for every positive constants a', b', and γ satisfying*

$$a' \le a, \quad b' + \gamma \le b, \quad \varsigma b' \le (1 - \varsigma\mu)a', \tag{4.2}$$

and for every function $\varphi : X \to \widetilde{Y}$ such that

$$\widetilde{d}_Y(\varphi(\bar{x}), \varphi(x)) \le \gamma \quad \forall x \in B_X(\bar{x}; a'), \tag{4.3}$$

and

$$d_Y(\varphi(x), \varphi(x')) \le \mu d_X(x, x') \quad \forall x, x' \in B_X(\bar{x}; a'), \tag{4.4}$$

the mapping $B_{\widetilde{Y}}(\bar{y} + \varphi(\bar{x}); b') \ni y \mapsto (\varphi + \Phi)^{-1}(y) \cap B_X(\bar{x}; a')$ is single-valued and Lipschitz continuous with constant ς' with respect to the metric d_Y.

Proof. Let us fix μ, ς', a', b' and γ as in the theorem. Take an arbitrary function $\varphi : X \to \widetilde{Y}$ such that (4.3) and (4.4) are fulfilled.

By assumption, the mapping $B_{\widetilde{Y}}(\bar{y}; b) \ni y \mapsto s(y) := \Phi^{-1}(y) \cap B_X(\bar{x}; a)$ is a Lipschitz continuous function (with respect to the metric d_Y in $\widetilde{Y}$) with Lipschitz constant ς. For any $x \in B_X(\bar{x}; a')$ and $y \in B_{\widetilde{Y}}(\bar{y} + \varphi(\bar{x}); b')$ we have

$$\widetilde{d}_Y(y - \varphi(x), \bar{y}) \le \widetilde{d}_Y(y, \bar{y} + \varphi(\bar{x})) + \widetilde{d}_Y(\varphi(\bar{x}), \varphi(x)) \le b' + \gamma \le b. \tag{4.5}$$

Thus $s(y - \varphi(x))$ is defined for all such pairs (x, y).

For an arbitrarily fixed $y \in B_{\widetilde{Y}}(\bar{y} + \varphi(\bar{x}); b')$ we consider the mapping $B_X(\bar{x}; a') \ni x \mapsto Z_y(x) := s(y - \varphi(x))$. We shall prove that the mapping Z_y has a unique fixed point by using the contraction mapping theorem in the form of [13, Theorem 1A.2]. For this we denote $\lambda = \varsigma\mu < 1$ and estimate

$$\begin{aligned} d_X(\bar{x}, Z_y(\bar{x})) &= d_X(s(\bar{y}), s(y - \varphi(\bar{x}))) \le \varsigma d_Y(\bar{y} + \varphi(\bar{x}), y) \\ &\le \varsigma b' \le (1 - \varsigma\mu)a' = (1 - \lambda)a'. \end{aligned}$$

Moreover, for x, $x' \in B_{d_X}(\bar{x}; a')$ we have

$$\begin{aligned} d_X(Z_y(x), Z_y(x')) &= d_X(s(y - \varphi(x)), s(y - \varphi(x'))) \le \varsigma d_Y(y - \varphi(x), y - \varphi(x')) \\ &= \varsigma d_Y(\varphi(x), \varphi(x')) \le \varsigma\mu d_X(x, x') = \lambda d_X(x, x'). \end{aligned}$$

Then, according to [13, Theorem 1A.2], there exists a unique $x = x(y) \in B_X(\bar{x}; a')$ such that $x = s(y - \varphi(x))$. The latter implies that $y - \varphi(x) \in \Phi(x)$, hence $x \in (\varphi + \Phi)^{-1}(y) \cap B_X(\bar{x}; a')$. Moreover, $x(y)$ is the unique element of $(\varphi + \Phi)^{-1}(y) \cap B_X(\bar{x}; a')$. Indeed, if $x \in (\varphi + \Phi)^{-1}(y) \cap B_X(\bar{x}; a')$, then $y \in \varphi(x) + \Phi(x)$, hence $y - \varphi(x) \in \Phi(x)$, and since as in (4.5) we have $y - \varphi(x) \in B_{\widetilde{Y}}(\bar{y}; b)$ and $x \in B_X(\bar{x}; a') \subset B_X(\bar{x}; a)$, it also holds that $x = s(y - \varphi(x))$. Thus $x = x(y)$. Thus the mapping $B_{\widetilde{Y}}(\bar{y} + \varphi(\bar{x}); b') \ni y \mapsto (\varphi + \Phi)^{-1}(y) \cap B_X(\bar{x}, a')$ is single-valued.

Now, take two arbitrary elements y, $y' \in B_{\widetilde{Y}}(\bar{y} + \varphi(\bar{x}); b')$ and let $x = s(y - \varphi(x))$ and $x' = s(y' - \varphi(x'))$ be the unique solutions of $y \in \varphi(x) + \Phi(x)$ in $B_X(\bar{x}; a')$ corresponding to y and y', respectively. Then

$$\begin{aligned} d_X(x, x') &= d_X(s(y - \varphi(x)), s(y' - \varphi(x'))) \le \varsigma d_Y(y - \varphi(x), y' - \varphi(x')) \\ &\le \varsigma d_Y(y, y') + \varsigma d_Y(\varphi(x), \varphi(x')) \le \varsigma d_Y(y, y') + \varsigma\mu d_X(x, x'). \end{aligned}$$

Hence,

$$d_X(x, x') \le \frac{\varsigma}{1 - \varsigma\mu} d_Y(y, y') \le \varsigma' d_Y(y, y'),$$

which completes the proof. Q.E.D.

The main improvement in the above theorem, compared with [27, Theorem 2.1], is that the Lipschitz property (4.4) is required in [27, Theorem 2.1] to be fulfilled in the stronger metric $\widetilde{d}_Y$, which makes the theorem unusable in several applications, including that presented in Subsection 4.3.

4.2 Strong bi-metric regularity of the linear-quadratic problem

Now consider again Problem (P). First we will present a result about stability under perturbations of Assumption (A3) in the case $\kappa = 1$, where the following strengthened form of Assumption (A1) will be used.

Assumption (A1'). The functions A, W and d are continuous, B and S have continuous first derivatives. The matrices $W(t)$ and $S^\top(t)B(t)$ are symmetric for every $t \in [0,T]$. The function g is differentiable with (globally) Lipschitz continuous gradient.

Furthermore we introduce the subspace

$$\widetilde{\mathcal{Y}} := L^\infty([0,T],\mathbb{R}^n) \times L^\infty([0,T],\mathbb{R}^n) \times W^{1,\infty}([0,T],\mathbb{R}^m) \times \mathbb{R}^n$$

of $\mathcal{Y}$ endowed with the usual norm of $y = (\xi,\pi,\rho,\nu) \in \widetilde{\mathcal{Y}}$:

$$\|(\xi,\pi,\rho,\nu)\|_\sim := \|\xi\|_\infty + \|\pi\|_\infty + \|\rho\|_{1,\infty} + |\nu|. \tag{4.6}$$

We denote by $\widetilde{d}_{\mathcal{Y}}$ the distance induced by $\|\cdot\|_\sim$.

Proposition 4.3. *(Stability of Assumption (A3).) Let Assumption (A1') be fulfilled. Let $(\hat{x},\hat{p},\hat{u})$ be a solution of* (PMP)*, and let Assumption (A2) and Assumption (A3) with $\kappa = 1$ be fulfilled. Then Assumption (A3) is stable under perturbations in the following sense: there exist constants $\widetilde{b} > 0$, $\widetilde{\alpha} > 0$, $\widetilde{\tau} > 0$ and $\widetilde{m}_0 > 0$ such that if $(\xi,\pi,\rho,\nu) = y \in \widetilde{\mathcal{Y}}$ with $\|y\|_\sim \le \widetilde{b}$, then for any triple $(x,p,u) \in \mathcal{X}$ solving $y \in F(x,p,u)$ the function $\sigma := B^\top p + S^\top x - \rho$ satisfies Assumption (A3) with $\kappa = 1$ and constants $\widetilde{\alpha}$ and $\widetilde{\tau}$ replacing α and τ, respectively, and $l_{\min}(\sigma,\widetilde{\tau}) \ge \tilde{m}_0$ (see* (3.1)*).*

Proof. Let α and τ be the constants appearing in Assumption (A3), and let $j \in \{1,\dots,m\}$ be arbitrary. Further, we consider only disturbances $y \in \widetilde{\mathcal{Y}}$ satisfying $\|y\|_\sim \le 1$.

First, observe that for all $t \in [0,T]$ it holds that

$$|\sigma_j(t) - \hat{\sigma}_j(t)| \le \left| \Big(B(t)^\top (p(t) - \hat{p}(t)) + S(t)^\top (x(t) - \hat{x}(t)) \Big)_j \right| + |\rho_j(t)|.$$

Using this inequality and Theorem 3.3 (applied with $b = 1$), we obtain that there is a constant c_1 such that

$$|\sigma_j(t) - \hat{\sigma}_j(t)| \le c_1 \|y\|$$

for all $j = 1,\dots,m$, $t \in [0,T]$, and $y \in \widetilde{\mathcal{Y}}$ with $\|y\|_\sim \le 1$. Hence,

$$|\sigma_j(t)| \ge |\hat{\sigma}_j(t)| - c_1\|y\|, \quad t \in [0,T], \quad j \in \{1,\dots,m\}. \tag{4.7}$$

Consider (skipping the argument t) the derivative

$$\begin{aligned}
\dot{\hat{\sigma}}_j &= \left[\dot{B}^\top \hat{p} + B^\top \dot{\hat{p}} + \dot{S}^\top \hat{x} + S^\top \dot{\hat{x}} \right]_j \\
&= \left[\dot{B}^\top \hat{p} + B^\top (-A^\top \hat{p} - W\hat{x} - S\hat{u}) + \dot{S}^\top \hat{x} + S^\top (A\hat{x} + B\hat{u} + d) \right]_j \\
&= \left[\dot{B}^\top \hat{p} + B^\top (-A^\top \hat{p} - W\hat{x}) + \dot{S}^\top \hat{x} + S^\top (A\hat{x} + d) \right]_j,
\end{aligned} \tag{4.8}$$

where in the last inequality we use the symmetricity of $B^\top S$. This implies, in particular, that $\dot{\hat\sigma}_j$ is continuous. Then there exists $\tau_1 \in (0,\tau]$ such that $|\dot{\hat\sigma}_j(\theta_1) - \dot{\hat\sigma}_j(\theta_2)| \le \alpha/4$ whenever θ_1, $\theta_2 \in [0,T]$ and $|\theta_1 - \theta_2| < \tau_1$. Hence, using (4.8) and Assumption (A3) we obtain that for any $j \in \{1,\dots,m\}$, for any zero $\hat s$ of $\hat\sigma_j$ and arbitrary $t \in (\hat s - \tau_1, \hat s + \tau_1) \cap [0,T]$

$$\alpha|t-\hat s| \le |\hat\sigma_j(t) - \hat\sigma_j(\hat s)| = \Big|\int_{\hat s}^t \dot{\hat\sigma}_j(\theta)\,d\theta\Big| \le \Big|\int_{\hat s}^t \dot{\hat\sigma}_j(\hat s)\,d\theta\Big| + \Big|\int_{\hat s}^t \frac{\alpha}{4}\,d\theta\Big|,$$

hence $|\dot{\hat\sigma}_j(\hat s)| \ge 3\alpha/4$ for any zero $\hat s$ of $\hat\sigma_j$, $j = 1,\dots,m$.

The equality (4.8) holds also for σ_j (where $(\hat x, \hat p)$ is replaced with (x,p)), with the additional term $\Big[B^\top\pi + S^\top\xi - \dot{\hat\rho}\Big]_j$ in the right-hand side. Then using Assumption (A1'), and the estimation in Theorem 3.3, we obtain that

$$\|\dot\sigma_j - \dot{\hat\sigma}_j\|_\infty \le c_2(\|y\| + \|\xi\|_\infty + \|\pi\|_\infty + \|\dot\rho\|_\infty) \le c_3\|y\|_\sim, \tag{4.9}$$

where c_2 and c_3 are independent of j and $y \in \widetilde{\mathcal{Y}}$, $\|y\|_\sim \le 1$.

Define $\widetilde\tau := \tau_1/2$ and choose the number $\widetilde b > 0$ in such a way that

$$c_1\widetilde b \le \min\left\{\frac{l_{\min}(\hat\sigma,\widetilde\tau/2)}{2}, \frac{\alpha\widetilde\tau}{4}\right\} \quad \text{and} \quad 4c_3\widetilde b \le \alpha, \quad \widetilde b \le 1, \tag{4.10}$$

and let $\|y\|_\sim \le \widetilde b$. Since from (4.7) and the first inequality in (4.10) we have that for $t \in [0,T] \setminus I_j(\hat\sigma,\widetilde\tau/2)$

$$|\sigma_j(t)| \ge |\hat\sigma_j(t)| - c_1\|y\| \ge l_{\min}(\hat\sigma,\widetilde\tau/2) - c_1\widetilde b \ge \frac{l_{\min}(\hat\sigma,\widetilde\tau/2)}{2} > 0,$$

we obtain that any zero s of σ_j is contained in $I_j(\hat\sigma,\widetilde\tau/2)$. Thus $s \in (\hat s - \widetilde\tau/2, \hat s + \widetilde\tau/2) \cap [0,T]$ for some zero $\hat s$ of $\hat\sigma_j$.

Now take an arbitrary $t \in (s-\widetilde\tau, s+\widetilde\tau) \cap [0,T]$. Then t, $s \in (\hat s - \tau_1, \hat s + \tau_1) \cap [0,T]$ and using (4.9) and the second inequality in (4.10) we obtain that

$$\begin{aligned}
|\sigma_j(t)| &= \left|\int_s^t \dot\sigma_j(\theta)\,d\theta\right| = \left|\int_s^t \Big[\dot{\hat\sigma}_j(\hat s) + (\dot{\hat\sigma}_j(\theta) - \dot{\hat\sigma}_j(\hat s)) + (\dot\sigma_j(\theta) - \dot{\hat\sigma}_j(\theta))\Big]\,d\theta\right| \\
&\ge |\dot{\hat\sigma}_j(\hat s)||t-s| - \frac{\alpha}{4}|t-s| - c_3\|y\|_\sim|t-s| \\
&\ge \frac{3\alpha}{4}|t-s| - \frac{\alpha}{4}|t-s| - \frac{\alpha}{4}|t-s| \ge \frac{\alpha}{4}|t-s|.
\end{aligned}$$

Thus (A3) holds for σ with $\kappa = 1$ and constants $\widetilde\alpha = \alpha/4$ and $\widetilde\tau$.

Further for $t \in I(\hat\sigma,\widetilde\tau) \setminus I(\sigma,\widetilde\tau)$ we have

$$|\sigma_j(t)| \ge \alpha|t-\hat s| - c_1\|y\| \ge \alpha|t-s| - \alpha|s-\hat s| - c_1\|y\| \ge \frac{\alpha\widetilde\tau}{4}$$

for some zeros $\hat s$ and s of $\hat\sigma$ and σ respectively. So if we set $m_0 := \min\{\frac{\alpha\widetilde\tau}{4}, l_{\min}(\hat\sigma,\widetilde\tau)\}$ then $l_{\min}(\sigma,\widetilde\tau) \ge m_0$. Q.E.D.

Proposition 4.3 allows to extend the result for strong bi-metric regularity of F, obtained in [27] for Mayer's problems for linear systems, to the present Bolza problem. For that we need the following stronger version of Assumption (A2).

Assumption (A2'). For every couple of admissible pairs $(x,u),\ (x',u') \in \mathcal{F}$ it holds that

$$\langle \nabla g(x(T)) - \nabla g(x'(T)), x(T) - x'(T)\rangle + \int_0^T \left(\langle W(t)\left(x(t)-x'(t)\right), x(t)-x'(t)\rangle + 2\langle S(t)\left(u(t)-u'(t)\right), x(t)-x'(t)\rangle\right) dt \geq 0.$$

Remark 4.4. Standard convex analysis shows that Assumption (A2') is equivalent to the fact that the functional J is convex on the set $\mathcal{F}$, or equivalently, the mapping $L^1([0,T],\mathbb{R}^m) \ni u \mapsto J(x(u),u)$ is convex on the set of admissible controls $\mathcal{U}$, where $x(u)$ denotes the solution of the Cauchy problem $\dot{x} = Ax + Bu,\ x(0) = 0$.

To prove strong bi-metric regularity of (PMP) we first have to introduce the following additional spaces. First we consider the set $\mathcal{U} = L^\infty([0,T],U)$ as a metric space with the metric

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

where "meas" stands for the Lebesgue measure in $[0,T]$. This metric is shift-invariant and we shall shorten $d^\#(u_1,u_2) = d^\#(u_1-u_2,0) =: d^\#(u_1-u_2)$. Moreover, $\mathcal{U}$ is a complete metric space with respect to $d^\#$ (see [14, Lemma 7.2]). Then the triple (x,p,u) is considered as an element of the space

$$\widetilde{\mathcal{X}} = W^{1,1}_{x_0}([0,T],\mathbb{R}^n) \times W^{1,1}([0,T],\mathbb{R}^n) \times \mathcal{U}, \tag{4.11}$$

endowed with the (shift-invariant) metric

$$\widetilde{d}_{\mathcal{X}}(x,p,u) = \|x\|_{1,1} + \|p\|_{1,1} + d^\#(u). \tag{4.12}$$

Clearly $\widetilde{\mathcal{X}}$ is a complete metric space.

Theorem 4.5 (Bi-metric regularity)**.** *Let Assumptions (A1') and (A2') be fulfilled. Let $(\hat{x},\hat{p},\hat{u})$ be a solution of* (PMP) *such that Assumption (A3) is fulfilled with $\kappa = 1$. Then the mapping $F : \widetilde{\mathcal{X}} \rightrightarrows \mathcal{Y}$ introduced in* (1.2) *is strongly bi-metrically regular relative to $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$ at $\hat{z} := (\hat{x},\hat{p},\hat{u}) \in \widetilde{\mathcal{X}}$ for $0 \in \widetilde{\mathcal{Y}}$.*

Proof. We shall prove that F^{-1} is single-valued in $B_{\widetilde{\mathcal{Y}}}(0;\widetilde{b})$ and

$$\widetilde{d}_{\mathcal{X}}(F^{-1}(y'),F^{-1}(y)) \leq c\, d_{\mathcal{Y}}(y',y), \tag{4.13}$$

for all $y,y' \in B_{\widetilde{\mathcal{Y}}}(0;\widetilde{b})$, where $\widetilde{b}$ and c are as in Proposition 4.3. Thus the conditions in Definition 4.1 will be fulfilled even with $a = +\infty$.

Let us start by giving a reformulation of the perturbed version of (P), which will turn out to be useful in the sequel. Let us take an arbitrary $y = (\xi,\pi,\rho,\nu) \in \mathcal{Y}$. Then the perturbed system $y \in F(x,p,u)$ is the set of necessary conditions for the problem (3.6) introduced in the proof of Theorem 3.3. Notice that (3.6) is exactly of the same form as (P) with the state and co-state variables augmented by one dimension, and the data A, B, d, W, S and g replaced with

$$\widetilde{A} = \begin{pmatrix} A & 0 \\ \pi^\top & 0 \end{pmatrix}, \widetilde{B} = \begin{pmatrix} B \\ \rho^\top \end{pmatrix}, \widetilde{d} = \begin{pmatrix} d+\xi \\ 0 \end{pmatrix}, \widetilde{W} = \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix}, \widetilde{S} = \begin{pmatrix} S \\ 0 \end{pmatrix},$$

and $\widetilde{g}(x(T), x_{n+1}(T)) = g(x(T)) - \nu^\top x(T) - x_{n+1}(T)$, respectively. Thus, (x, p, u) is a solution of $y \in F(x, p, u)$ if and only if the triple

$$(\widetilde{x}(\cdot), \widetilde{p}(\cdot), \widetilde{u}(\cdot)) = \left(\begin{pmatrix} x(\cdot) \\ \int_0^\cdot \left(\pi^\top x + \rho^\top u\right) dt \end{pmatrix}, \begin{pmatrix} p(\cdot) \\ -1 \end{pmatrix}, u(\cdot) \right) \tag{4.14}$$

is a solution of the system

$$\begin{aligned} 0 &= \dot{\widetilde{x}}(t) - \widetilde{A}(t)\widetilde{x}(t) - \widetilde{B}(t)\widetilde{u}(t) - \widetilde{d}(t) \\ 0 &= \dot{\widetilde{p}}(t) + \widetilde{A}(t)^\top \widetilde{p}(t) + \widetilde{W}(t)\widetilde{x}(t) + \widetilde{S}(t)\widetilde{u}(t) \\ 0 &\in \widetilde{B}(t)^\top \widetilde{p}(t) + \widetilde{S}(t)^\top \widetilde{x}(t) + N_U(\widetilde{u}(t)) \\ 0 &= \widetilde{p}(T) - \nabla \widetilde{g}(\widetilde{x}(T)). \end{aligned} \tag{4.15}$$

The above system can be recast as a generalized inclusion

$$0 \in \widetilde{F}_y(\widetilde{x}, \widetilde{p}, \widetilde{u}) \tag{4.16}$$

where $\widetilde{F}_y$ is defined as in (1.2) replacing A by $\widetilde{A}$, and similarly for the other data. $\widetilde{F}_y$ maps the space

$$\hat{\mathcal{X}} := W^{1,1}_{\widetilde{x}_0}([0,T], \mathbb{R}^{n+1}) \times W^{1,1}([0,T], \mathbb{R}^{n+1}) \times L^1([0,T], \mathbb{R}^m)$$

to

$$\hat{\mathcal{Y}} := L^1([0,T], \mathbb{R}^{n+1}) \times L^1([0,T], \mathbb{R}^{n+1}) \times L^\infty([0,T], \mathbb{R}^m) \times \mathbb{R}^{n+1},$$

where $\widetilde{x}_0 := (x_0^\top, 0)^\top$. In few words, the dimension of the state and co-state variable is augmented to $n+1$ and the additional initial condition $x_{n+1}(0) = 0$ is added.

Note that by construction for any $y \in \widetilde{\mathcal{Y}}$ Assumption (A1) and Assumption (A2') are fulfilled for (4.16). Choose $\widetilde{b}$, $\widetilde{\alpha}$, $\widetilde{\tau}$ and m_0 as in Proposition 4.3. Then there exists a constant K such that for any y with $\|y\|_\sim \le \widetilde{b}$ we have

$$\|\widetilde{A}\|_\infty, \|\widetilde{B}\|_\infty\|, \|\widetilde{d}\|_\infty,\ \|\widetilde{W}\|_\infty,\ \|\widetilde{S}\|_\infty \le K, \quad \nabla\tilde{g} \text{ is Lipschitz with constant } K.$$

Then by Proposition 4.3 for any $y = (\xi, \pi, \rho, \nu) \in B_{\widetilde{\mathcal{Y}}}(0; \widetilde{b})$ and any solution (x, p, u) of the perturbed problem $y \in F(x, p, u)$ Assumption (A3) is satisfied by $\sigma := B^\top p + Sx - \rho$ with constants $\widetilde{\alpha}$, $\widetilde{\tau}$ and $l_{\min}(\sigma, \widetilde{\tau}) \ge m_0$. An easy calculation shows that the switching function of the solution $(\widetilde{x}, \widetilde{p}, \widetilde{u})$ (given by (4.14)) of (4.16) is given by $\widetilde{B}^\top \widetilde{p} + \widetilde{S}^\top \widetilde{x} = B^\top p + S^\top x - \rho = \sigma$. Then Theorem 3.3 in the detailed form in Remark 3.4 is applicable to (4.16) with the constant c independent of the particular $y \in B_{\widetilde{\mathcal{Y}}}(0; \widetilde{b})$. In particular, this implies that $(\widetilde{x}, \widetilde{p}, \widetilde{u})$ is the unique solution for (4.16). Therefore, $\widetilde{u} = u$ is bang-bang and F^{-1} is single valued on $B_{\widetilde{\mathcal{Y}}}(0; \widetilde{b})$. For any $y' = (\xi', \pi', \rho', \nu') \in B_{\widetilde{\mathcal{Y}}}(0; \widetilde{b})$ and its solution (x', p', u') of $y' \in F(x', p', u')$ we define

$$(\widetilde{x}', \widetilde{p}', \widetilde{u}') := ((x', \int_0^\cdot (\pi^\top x' + \rho^\top u')), (p', -1), u'), \quad \widetilde{y}' := ((\xi' - \xi, 0), (\pi' - \pi, 0), \rho' - \rho, (\nu' - \nu, 0)).$$

An easy calculation shows the inclusion $\widetilde{y}' \in \widetilde{F}_y(\widetilde{x}', \widetilde{p}', \widetilde{u}')$. Then Theorem 3.3 (in the form in Remark 3.4) implies

$$\|\widetilde{x}' - \widetilde{x}\|_{1,1} + \|\widetilde{p}' - \widetilde{p}\|_{1,1} + \|\widetilde{u}' - \widetilde{u}\|_1 \le c\|\widetilde{y}'\|_{\hat{\mathcal{Y}}}, \tag{4.17}$$

where $\|\cdot\|_{\hat{\mathcal{Y}}}$ denotes the norm of $\hat{\mathcal{Y}}$. Hence by (4.17) we have

$$\begin{aligned}\|x-x'\|_{1,1}+\|p-p'\|_{1,1}+\|u-u'\|_1 &\le \|\widetilde{x}'-\widetilde{x}\|_{1,1}+\|\widetilde{p}'-\widetilde{p}\|_{1,1}+\|\widetilde{u}'-\widetilde{u}\|_1\\ &\le c\|\widetilde{y}'\|_{\hat{\mathcal{Y}}}=c\|y-y'\|.\end{aligned}$$

Since u, u' are bang-bang, similar to [27, p. 4130] we have $\|u-u'\|_1 \ge 2d^{\#}(u-u')$ which proves (4.13). Q.E.D.

We mention that the strong bi-metric regularity for Mayer's problems is proved in [27] for a general polyhedral set U and also in the case $\kappa > 1$. Extension of Theorem 4.5 to a general compact polyhedral U set is a matter of modification of Assumption (A3) and technicalities that we avoid in this paper, while the case $\kappa > 1$ is still open and challenging for the Bolza problem.

4.3 Stability of bi-metric regularity under perturbations

In this subsection, we will apply Theorem 4.2 to prove that the strong bi-metric regularity property is stable under some class of nonlinear perturbations.

Along with problem (P) we consider the following perturbed problem:

$$\begin{array}{ll}\text{minimize} & \widetilde{J}(x,u)\\ \text{subject to} & \dot{x}(t)=A(t)x(t)+\widetilde{a}(x(t),t)+B(t)u(t)+\widetilde{B}(x(t),t)u(t), \quad t\in[0,T],\\ & u(t)\in U:=[-1,1]^m,\\ & x(0)=x_0,\end{array} \tag{4.18}$$

where

$$\begin{aligned}\widetilde{J}(x,u):=&g(x(T))+\widetilde{g}(x(T))+\\ &\int_0^T\left(\frac{1}{2}x(t)^\top W(t)x(t)+\widetilde{w}(x(t),t)+x(t)^\top S(t)u(t)+\langle\widetilde{s}(x(t),t),u(t)\rangle\right)dt.\end{aligned}$$

Here $\widetilde{a}:\mathbb{R}^n\times[0,T]\to\mathbb{R}^n$, $\widetilde{B}:\mathbb{R}^n\times[0,T]\to\mathbb{R}^{n\times m}$, $\widetilde{g}:\mathbb{R}^n\to\mathbb{R}$, $\widetilde{w}:\mathbb{R}^n\times[0,T]\to\mathbb{R}$, $\widetilde{s}:\mathbb{R}^n\times[0,T]\to\mathbb{R}^m$ are continuously differentiable functions. All these functions will be assumed "small" in a sense clarified in the theorem below.

The system of necessary optimality conditions for problem (4.18) is given by

$$\begin{aligned}0 &= \dot{x}(t)-A(t)x(t)-\widetilde{a}(x(t),t)-B(t)u(t)-\widetilde{B}(x(t),t)u(t),\\ 0 &= \dot{p}(t)+\big(A(t)+\widetilde{a}_x(x(t),t)+(\widetilde{B}(x(t),t)u(t))_x\big)^\top p(t)+W(t)x(t)+\widetilde{w}_x(x(t),t)^\top+S(t)u(t),\\ &\quad +\widetilde{s}_x(x(t),t)^\top u(t)\\ 0 &\in (B(t)+\widetilde{B}(x(t),t))^\top p(t)+S(t)^\top x(t)+\widetilde{s}(x(t),t)+N_U(u(t)),\\ 0 &= p(T)-\nabla g(x(T))-\nabla\widetilde{g}(x(T)),\end{aligned} \tag{4.19}$$

where the subscript x (as in $\widetilde{a}_x$) means differentiation with respect to x.

The system (4.19) can be recast as

$$0\in f(x,p,u)+F(x,p,u), \tag{4.20}$$

where F (corresponding to the non–perturbed system) is given by (1.2) and f is defined by

$$f(x,p,u)(t) = \begin{pmatrix} -\widetilde{a}(x,t) - \widetilde{B}(x,t)u \\ \big(\widetilde{a}_x(x,t) + (\widetilde{B}(x,t)u)_x\big)^\top p + \widetilde{w}_x(x,t)^\top + \widetilde{s}_x(x,t)^\top u \\ \widetilde{B}(x,t)^\top p + \widetilde{s}(x,t) \\ -\nabla\widetilde{g}(x(T)) \end{pmatrix}.$$

As before we consider F as a set-valued mapping $\widetilde{\mathcal{X}} \rightrightarrows \mathcal{Y}$, where the spaces $\widetilde{\mathcal{X}}$ and $\mathcal{Y}$ are defined in (4.11) and (2.1), respectively. We fix a solution $\hat{z} := (\hat{x}, \hat{p}, \hat{u})$ of the inclusion $0 \in F(x,p,u)$.

Assumption (B). The mapping $F : \widetilde{\mathcal{X}} \rightrightarrows \mathcal{Y}$ is strongly bi-metrically regular relative to $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$ at $\hat{z} \in \widetilde{\mathcal{X}}$ for $0 \in \widetilde{\mathcal{Y}}$.

We recall that sufficient conditions for strong bi-metric regularity of F are given in Theorem 4.5.

Our purpose will be to prove that the strong bi-metric regularity of F is not destroyed by the disturbance f, provided that the disturbances in (4.18) are sufficiently "small". Notice that the space $\widetilde{\mathcal{X}}$ contains elements (x,p,u) for which some of the norms $\|x\|_\infty$, $\|p\|_\infty$, $\|\dot{x}\|_\infty$, $\|\dot{p}\|_\infty$, may be arbitrarily large or even infinite (the latter applies to the derivatives), that is, elements which are irrelevant to the linear-quadratic problem to which F is associated. Moreover, the image $f(\widetilde{\mathcal{X}})$ is not necessarily contained in $\widetilde{\mathcal{Y}}$, which is important from a technical point of view. Therefore, for a given compact set $D \subset \mathbb{R}^n$ we introduce the complete metric space (with the metric $\widetilde{d}_X$)

$$\widetilde{\mathcal{X}}_D := \{(x,p,u) \in \widetilde{\mathcal{X}} : x(t), p(t), \dot{x}(t), \dot{p}(t) \in D \text{ for any } t \in [0,T]\}.$$

Also, denote by $F_D := F|_{\widetilde{\mathcal{X}}_D} : \widetilde{\mathcal{X}}_D \rightrightarrows \mathcal{Y}$ and $f_D := f|_{\widetilde{\mathcal{X}}_D} : \widetilde{\mathcal{X}}_D \to \mathcal{Y}$ the restrictions of F and f to $\mathcal{X}_D$.

Lemma 4.6. *Let Assumption (A1) be fulfilled, let $\hat{z} = (\hat{x}, \hat{p}, \hat{u})$ be a solution of the non-perturbed system* (PMP)*, and let Assumption (B) be fulfilled. Then there exists a compact set $D_0 \subset \mathbb{R}^n$ such that for every compact set $D \subset \mathbb{R}^n$ containing D_0 the restriction f_D maps $\widetilde{\mathcal{X}}_D$ into $\widetilde{\mathcal{Y}}$ and the mapping $F_D : \widetilde{\mathcal{X}}_D \rightrightarrows \mathcal{Y}$ is strongly bi-metrically regular relative to $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$ at $\hat{z} \in \widetilde{\mathcal{X}}$ for $0 \in \widetilde{\mathcal{Y}}$.*

Proof. First note that because of continuity of $\widetilde{a}$, $\widetilde{B}$, $\widetilde{s}$, $\widetilde{a}_x$, $\widetilde{B}_x$, $\widetilde{w}_x$ and $\widetilde{s}_x$ we have that for every compact set D_0 the first three components of f_{D_0} are in L^∞. Moreover the third component is differentiable in t and since $(\widetilde{B}(x,t)^\top p)_x$ is continuous as a function in x, p and t, and $\widetilde{s}_x$ is continuous this derivative lies in L^∞. Hence f_{D_0} maps into $\widetilde{\mathcal{Y}}$.

Further let $\varsigma \geq 0$, $a > 0$ and $b > 0$ be the constants corresponding the strong bi-metric regularity of F. Let $y = (\xi, \pi, \rho, \nu) \in B_{\widetilde{\mathcal{Y}}}(0;b)$ and $(x,p,u) \in \widetilde{\mathcal{X}}$ be a solution the generalized equation $y \in F(x,p,u)$ (i.e. of (2.4)). Moreover we denote $\Delta x(t) := x(t) - \hat{x}(t)$, $\Delta p(t) := p(t) - \hat{p}(t)$ and $\Delta u(t) := u(t) - \hat{u}(t)$. Then by the solution formula of the Cauchy problems for Δx and Δp we get

$$\|\Delta x\|_{1,\infty} \leq c_1(\|\xi\|_\infty + \|\Delta u\|_\infty), \quad \|\Delta p\|_{1,\infty} \leq c_2(\|\xi\|_\infty + \|\pi\|_\infty + \|\Delta u\|_\infty + |\nu|), \tag{4.21}$$

which shows that there is a compact set D_0 such that $(x,p,u) \in \widetilde{\mathcal{X}}_{D_0}$. Therefore $F^{-1}(B_{\widetilde{\mathcal{Y}}}(0;b)) \subseteq \widetilde{\mathcal{X}}_{D_0} \subseteq \widetilde{\mathcal{X}}_D$ for every D containing D_0 which implies that $F_D : \widetilde{\mathcal{X}}_D \rightrightarrows \mathcal{Y}$ is strongly bi-metrically regular relative to $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$ at $\hat{z} \in \widetilde{\mathcal{X}}$ for $0 \in \widetilde{\mathcal{Y}}$. Q.E.D.

Below we prove a stability result in the same spirit as [27, Theorem 4.1], which concerns Mayer's problems. We mention that there is a gap in the proof of [27, Theorem 4.1], but it can be easily corrected by using Theorem 4.2 instead of [27, Theorem 2.1]. This is done in the next theorem which, in addition, extends [27, Theorem 4.1] to Bolza problems.

Theorem 4.7. *Let assumption (A1') be fulfilled, let $\hat{z} = (\hat{x}, \hat{p}, \hat{u})$ be a solution of the non-perturbed system* (PMP)*, and let Assumption (B) be fulfilled. Let $D \subset \mathbb{R}^n$ be a compact set such that $f(\widetilde{\mathcal{X}}_D) \subset \widetilde{\mathcal{Y}}$ and the mapping F_D is strongly bi-metrically regular relative to $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$ at $\hat{z} \in \widetilde{\mathcal{X}}_D$ for $0 \in \widetilde{\mathcal{Y}}$ (see Lemma 4.6). Then there exist positive real numbers ε_0, δ and c with the following property.*

For any positive number $\varepsilon \le \varepsilon_0$ let $\widetilde{a}, \widetilde{B}, \widetilde{g}, \widetilde{w}, \widetilde{s}$ be any functions satisfying the assumptions given above in this section and such that

- *the functions $\widetilde{a}$, $\widetilde{B}$, $\widetilde{s}$, $\widetilde{a}_x$, $\widetilde{B}_x$, $\widetilde{w}_x$, $\widetilde{s}_x$, $\widetilde{B}_t$, $\widetilde{s}_t$ are all bounded by ε on $D \times [0,T]$;*
- *the functions $\widetilde{a}$, $\widetilde{B}$, $\widetilde{s}$, $\widetilde{a}_x$, $\widetilde{B}_x$, $\widetilde{w}_x$, $\widetilde{s}_x$ are Lipschitz continuous in x with Lipschitz constant ε;*
- *the function $\nabla\widetilde{g}$ is bounded by ε and Lipschitz continuous on D with Lipschitz constant ε.*

Then

(i) the perturbed system (4.19) *has a unique solution $z^* = (x^*, p^*, u^*)$ in the δ-neighborhood of $\hat{z}$ in $\widetilde{\mathcal{X}}_D$ and*

$$\widetilde{d}_{\mathcal{X}}(z^* - \hat{z}) \le c\varepsilon.$$

(ii) the mapping $f + F : \widetilde{\mathcal{X}}_D \rightrightarrows \mathcal{Y}$ is strongly bi-metrically regular at z^* for 0 relative to $\widetilde{\mathcal{Y}} \subset \mathcal{Y}$.

Proof. We want to apply Theorem 4.2 for the mappings $\Phi = F$ and $\varphi = f$ at the point $(\hat{z}, \hat{y})$, where $\hat{y} := f(\hat{x}, \hat{p}, \hat{u})$. Let ς, a, b be the numbers in the definition of strong bi-metric regularity of F at $\hat{z}$ for 0, and let μ, ς', a', b', γ be arbitrary numbers such that the conditions (4.2) are fulfilled.

Since $\widetilde{a}$, $\widetilde{B}$, $\widetilde{s}$, $\widetilde{a}_x$, $\widetilde{B}_x$, $\widetilde{w}_x$, $\widetilde{s}_x$, $\widetilde{B}_t$, $\widetilde{s}_t$, $\nabla\widetilde{g}$ are all bounded by ε and $\dot{\hat{x}}$, $\hat{p}$, $\dot{\hat{p}}$ are bounded by $|D| := \sup_{x\in D}|x|$ and $|\hat{u}| \le \sqrt{m}$ we have that

$$\begin{aligned} d_{\widetilde{\mathcal{Y}}}(\hat{y}, 0) &= \| - \widetilde{a}(\hat{x},t) - \widetilde{B}(\hat{x},t)\hat{u}\|_\infty + \|\big(\widetilde{a}_x(\hat{x},t) + (\widetilde{B}(\hat{x},t)\hat{u})_x\big)^\top \hat{p} + \widetilde{w}_x(\hat{x},t)^\top + \widetilde{s}_x(\hat{x},t)^\top \hat{u}\|_\infty + \\ &\quad + \|\widetilde{B}(\hat{x},t)^\top \hat{p} + \widetilde{s}(\hat{x},t)\|_\infty + \| \Big(\widetilde{B}(\hat{x},t)^\top \hat{p}\Big)_x \dot{\hat{x}} + \widetilde{B}_t(\hat{x},t)^\top \hat{p} + \widetilde{B}(\hat{x},t)^\top \dot{\hat{p}}\|_\infty + |\nabla\widetilde{g}(\hat{x}(T))| \\ &\le (1 + \sqrt{m})\varepsilon + (2 + |D| + \sqrt{m})\varepsilon + (|D| + 1)\varepsilon + (|D|^2 + 2|D|)\varepsilon + \varepsilon \\ &\le C_1\varepsilon, \end{aligned}$$

for some constant C_1 only depending on $|D|$. Similarly for $z \in B_{\widetilde{\mathcal{X}}_D}(\hat{z}; a')$ we have

$$\widetilde{d}_{\mathcal{Y}}(0, f(z)) \le C_1\varepsilon,$$

which gives

$$\widetilde{d}_{\mathcal{Y}}(\hat{y}, f(z)) \le \widetilde{d}_{\mathcal{Y}}(\hat{y}, 0) + \widetilde{d}_{\mathcal{Y}}(0, f(z)) \le 2C_1\varepsilon. \tag{4.22}$$

Next since $\widetilde{B}$, $\widetilde{a}_x$, $\widetilde{B}_x$, $\widetilde{s}_x$ are bounded by ε and $\widetilde{a}$, $\widetilde{B}$, $\widetilde{s}$, $\widetilde{a}_x$, $\widetilde{B}_x$, $\widetilde{w}_x$, $\widetilde{s}_x$, $\nabla\widetilde{g}$ are Lipschitz continuous with Lipschitz constant ε, p is bounded by $|D|$ and $|u| \leq \sqrt{m}$ we have that for any $z, z' \in B_{\widetilde{\mathcal{X}}_D}(\hat{z}; a')$

$$\begin{array}{rcl} d_{\mathcal{Y}}(f(z), f(z')) & = & \| - \widetilde{a}(x,t) - \widetilde{B}(x,t)u + \widetilde{a}(x',t) + \widetilde{B}(x',t)u'\|_1 \\ & + & \|\big(\widetilde{a}_x(x,t) + (\widetilde{B}(x,t)u)_x\big)^\top p + + \widetilde{w}_x(x,t)^\top + \widetilde{s}_x(x,t)^\top u \\ & & \quad -\big(\widetilde{a}_x(x',t) + (\widetilde{B}(x',t)u')_x\big)^\top p' - \widetilde{w}_x(x',t)^\top - \widetilde{s}_x(x',t)^\top u'\|_1 \\ & + & \|\widetilde{B}(x,t)^\top p + \widetilde{s}(x,t) - \widetilde{B}(x',t)^\top p' - \widetilde{s}(x',t)\|_\infty + |\nabla\widetilde{g}(x(T)) - \nabla\widetilde{g}(x'(T))| \\ & \leq & \varepsilon\big[(\|x - x'\|_1 + \sqrt{m}|D|\,\|x - x'\|_1 + \|u - u'\|_1)\|) \\ & + & (|D|\,\|x - x'\|_1 + \|p - p'\|_1 + \sqrt{m}|D|\,\|x - x'\|_1 + |D|\,\|u - u'\|_1 \\ & & \quad +\sqrt{m}\|p - p'\|_1 + \|x - x'\|_1 + \sqrt{m}\|x - x'\|_1 + \|u - u'\|_1) \\ & + & (\|x - x'\|_\infty + |D|\,\|x - x'\|_\infty + \|p - p'\|_\infty) \ + |x(T) - x'(T)|\big] \\ & \leq & C_2\varepsilon\|z - z'\|_{\widetilde{\mathcal{X}}_D} \end{array} \tag{4.23}$$

for some constant C_2 only depending on $|D|$.

Hence, if we choose ε_0, δ and c such that

$$2C_1\varepsilon_0 \leq \gamma,\ C_2\varepsilon_0 \leq \mu,\ C_1\varepsilon_0 < b',\ \delta = a',\ c = \varsigma' C_1,\ c\varepsilon_0 < a',$$

then we can apply Theorem 4.2 to see that $f + F$ is strongly bi-metrically regular at $\hat{z}$ for $\hat{y}$ with constants ς', a' and b'. Therefore, there is a unique $z^* \in B_{\widetilde{\mathcal{X}}_D}(\hat{z}; a')$ such that

$$0 \in f(z^*) + F(z^*).$$

and we have

$$\widetilde{d}_{\mathcal{X}}(z^* - \hat{z}) \leq \varsigma' d_{\mathcal{Y}}(0, \hat{y}) \leq \varsigma' C_1\varepsilon = c\varepsilon,$$

which proves (i). Moreover since $(z^*, 0) \in int(B_{\widetilde{\mathcal{X}}_D}(\hat{z}; a') \times B_{\widetilde{\mathcal{Y}}}(\hat{y}; b'))$, the map $f + F$ is also strongly bi-metrically regular at z^* for 0. This proves (ii). Q.E.D.

We mention that the issue of stability with respect to linearization of the strong bi-metric regularity property (in the spirit of Robinson's theorem [28]) is more complicated and will be a subject of a separate investigation, together with further applications of this property.

5 A Newton-type method for bang-bang optimal control problems

In this section we investigate the convergence of a Newton-type method for solving affine optimal control problems under conditions which guarantee that the (strengthened) sub-regularity property in Theorem 3.3 holds for the linearized problem along the optimal solution. For this, we first present an abstract result which is similar to, but stronger than [9, Theorem 6.1], since it is based on the stronger version of sub-regularity in Theorem 3.3.

Theorem 5.1. *Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces. Let the mapping $\varphi : X \to Y$ be Fréchet differentiable ($D\varphi$ denotes the derivative) and let $\Phi : X \rightrightarrows Y$ be a set-valued mapping. Let $\hat{x}$ be a solution of the inclusion*

$$\varphi(x) + \Phi(x) \ni 0.$$

Assume that there are positive constants R, L and c such that

$$\|D\varphi(x) - D\varphi(\hat{x})\| \leq L\|x - \hat{x}\|_X \quad \forall x \in B_X(\hat{x}, R) \tag{5.1}$$

and

$$\|x - \hat{x}\|_X \leq c\|y\|_Y \tag{5.2}$$

for every $x \in X$ and $y \in \varphi(\hat{x}) + D\varphi(\hat{x})(x - \hat{x}) + \Phi(x)$.

Then for $x \in B_X(\hat{x}, r)$, where $r = \min\{R, \frac{2}{5cL}\}$, and for every solution $z \in X$ of the Newton inclusion

$$\varphi(x) + D\varphi(x)(z - x) + \Phi(z) \ni 0, \tag{5.3}$$

it holds that $z \in B_X(\hat{x}, r)$ and

$$\|z - \hat{x}\|_X \leq \frac{1}{r}\|x - \hat{x}\|_X^2. \tag{5.4}$$

Before proving the theorem we mention that condition (5.2) is a strengthened form of the metric sub-regularity of the partial linearization $x \to \varphi(\hat{x}) + D\varphi(\hat{x})(x - \hat{x}) + \Phi(x)$ of the mapping $\varphi + \Phi$. The inclusion $z \in B_X(\hat{x}, r)$ implies that any finite or infinite sequence generated by the Newton inclusion (5.3) and starting from $B_X(\hat{x}, r)$ (if such exists) stays in $B_X(\hat{x}, r)$. Inequality (5.4) claims quadratic convergence of any such sequence which starts in the interior of $B_X(\hat{x}, r)$.

Proof. For any $x \in B_X(\hat{x}, r)$, let $z \in X$ be an arbitrary solution of (5.3) (if any). Then,

$$\varphi(\hat{x}) + D\varphi(\hat{x})(z - \hat{x}) + \Phi(z) \ni \varphi(\hat{x}) - \varphi(x) + D\varphi(\hat{x})(z - \hat{x}) - D\varphi(x)(z - x).$$

This means that z solves (5.3) with perturbation y given by the right-hand side of the inclusion above. Therefore, (5.2) yields that

$$\|z - \hat{x}\|_X \leq c\|\varphi(\hat{x}) - \varphi(x) + D\varphi(\hat{x})(z - \hat{x}) - D\varphi(x)(z - x)\|_Y.$$

Now using (5.1) we get

$$\begin{aligned}
\|z - \hat{x}\|_X &\leq c(\|\varphi(\hat{x}) - \varphi(x) + D\varphi(\hat{x})(x - \hat{x})\|_Y + \|(D\varphi(\hat{x}) - D\varphi(x))(z - x)\|_Y) \\
&\leq \frac{cL}{2}\|x - \hat{x}\|_X^2 + cL\|x - \hat{x}\|_X \|z - x\|_X \\
&\leq \frac{cL}{2}\|x - \hat{x}\|_X^2 + cL\|x - \hat{x}\|_X (\|z - \hat{x}\|_X + \|x - \hat{x}\|_X).
\end{aligned}$$

Hence,

$$(1 - cL\|x - \hat{x}\|_X)\|z - \hat{x}\|_X \leq \frac{3cL}{2}\|x - \hat{x}\|_X^2.$$

Since $1 - cL\|x - \hat{x}\|_X \geq (1 - cLr) \geq \frac{3}{5}$ we obtain (5.4), which implies that $z \in B_X(\hat{x}, r)$. Q.E.D.

Remark 5.2. A similar convergence result of the Newton's method can be found in [7] for variational inequalities and nonlinear programming. In that paper, the author introduces the conditions of hemi-stability hemi-regularity in order to ensure the convergence of the Newton's method. The assumptions in Theorem 5.1 are weaker, but existence of a Newton sequence is not claimed, similarly as to [9, Theorem 6.1]. Existence will follow in the analysis of optimal control problems that follow.

Now, we shall use Theorem 5.1 to investigate the convergence of the Newton method for the following affine optimal control problem:

$$\begin{array}{ll} \text{minimize} & C(x,u) \\ \text{subject to} & \dot{x}(t) = a(x(t),t) + B(x(t),t)u(t), \quad t \in [0,T], \\ & u(t) \in U := [-1,1]^m, \\ & x(0) = x_0, \end{array} \tag{5.5}$$

where

$$C(x,u) := g(x(T)) + \int_0^T \left[w(x(t),t) + \langle s(x(t),t), u(t) \rangle \right] dt.$$

Here the functions $a : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, $B : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n \times m}$, $w : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, $s : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}$ are given. Further, we use the following assumptions.

Assumption (A1"). The functions a, B, w, s are twice differentiable in x, and all these functions and derivatives of first and second order are continuous in t and locally Lipschitz in x, uniformly in t. g is twice continuously differentiable with Lipschitz derivate. The problem (5.5) has a solution, $(\hat{x}, \hat{u})$.

Remark 5.3. The optimality can be understood as local, since it is only important that the Pontryagin maximum principle is fulfilled for $(\hat{x}, \hat{u})$. Due to the linearity of the problem with respect to the control and the compactness and convexity of the control constraints, existence of an optimal solution is granted if the differential equation in (5.5) has a solution on $[0,T]$ for every $u \in \mathcal{U}$.

By the Pontryagin minimum principle, there exists an absolutely continuous function $\hat{p}$ such that the triple $(\hat{x}, \hat{p}, \hat{u})$ solves for a.e. $t \in [0,T]$ the system

$$\begin{array}{l} 0 = \dot{x}(t) - a(x(t),t) - B(x(t),t)u(t), \\ 0 = \dot{p}(t) + \big(a_x(x(t),t) + (B(x(t),t)u(t))_x\big)^\top p(t) + w_x(x(t),t)^\top + s_x(x(t),t)^\top u(t), \\ 0 \in B(x(t),t)^\top p(t) + s(x(t),t) + N_U(u(t)), \\ 0 = p(T) - \nabla g(x(T)), \end{array} \tag{5.6}$$

where the subscript x (as in a_x) means differentiation with respect to x.

We rewrite system (5.6) as the following generalized equation

$$0 \in f(x,p,u) + G(x,p,u), \tag{5.7}$$

where $f : \mathcal{X} \to \mathcal{Y}$ is given by

$$f(x,p,u)(t) := \begin{pmatrix} \dot{x} - a(x,t) - B(x,t)u \\ \dot{p} + \big(a_x(x,t) + (B(x,t)u)_x\big)^\top p + w_x(x,t)^\top + s_x(x,t)^\top u \\ B(x,t)^\top p + s(x,t) \\ p(T) - \nabla g(x(T)) \end{pmatrix}, \tag{5.8}$$

$G : \mathcal{X} \rightrightarrows \mathcal{Y}$ is given by

$$G(x,p,u) = \begin{pmatrix} 0 \\ 0 \\ N_{\mathcal{U}}(u) \\ 0 \end{pmatrix}, \tag{5.9}$$

and $\mathcal{X}$ and $\mathcal{Y}$ are the spaces defined in Section 2, namely $\mathcal{X} = W^{1,1}_{x_0} \times W^{1,1} \times L^1$, $\mathcal{Y} = L^1 \times L^1 \times L^\infty \times \mathbb{R}^n$.

Following [13, Chapter 6.3], we define the Newton-type method for solving problem (5.7) as follows, where $z^k := (x^k, p^k, u^k)$ denotes the obtained iterate at step $k = 0, 1, \ldots$.

Newton's method:

1. Choose $z^0 \in \mathcal{X}$.
2. Given z^k, obtain z^{k+1} as a solution of the generalized equation

$$f(z^k) + Df(z^k)(z^{k+1} - z^k) + G(z^{k+1}) \ni 0. \tag{5.10}$$

Here, $Df(z)$ is the Jacobian of f at z.

We mention that if z^k satisfies (5.10) then u^k is an admissible control, because $N_{\mathcal{U}}(u) = \emptyset$ whenever $u \notin \mathcal{U}$.

For any $\bar{z} \in \mathcal{X}$ the inclusion $f(\bar{z}) + Df(\bar{z})(z - \bar{z}) + G(z) \ni 0$ represents the Pontryagin system of necessary optimality conditions for a linear-quadratic problem which can be recast as (P) by introducing an additional state variable, similarly in the proof of Theorem 4.5. We denote this problem by $LP(\bar{z})$ (we skip its explicit formulation, which can be found for instance in [11, Section 5]). For the next theorem it is important to ensure that the claim in Theorem 3.3 holds for the particular problem $LP(\hat{z})$ corresponding to $\bar{z} = \hat{z}$, which obviously has the solution $\hat{z}$ – the solution of the non-linearized problem (5.5). Therefore, we make the following assumptions, related to Assumption (A2) and (A3) in Section 2.

Assumption (A2"). The objective functional in problem $LP(\hat{z})$ is convex on the set of all admissible pairs $\mathcal{F}$.

Assumption (A3"). The switching function $\hat{\sigma}(t)$ in problem $LP(\hat{z})$, which is

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

satisfies Assumption (A3) with $\kappa = 1$.

The next theorem claims that on the assumptions made, Newton's method generates a sequence quadratically converging to the optimal solution of (5.5).

Theorem 5.4. *Let Assumption (A1") be fulfilled and let $\hat{z} := (\hat{x}, \hat{p}, \hat{u})$ be a solution of problem (5.6). Let, in addition, Assumptions (A2") at (A3") be fulfilled for $\hat{z}$. Then there exists a neighborhood $O \subset \mathcal{X}$ of $\hat{z}$ such that for any starting point $z^0 \in O$ there is a sequence $\{z^k\}_{k=1}^\infty = \{(x^k, p^k, u^k)\}_{k=1}^\infty$ (not necessarily unique) generated by the Newton method (5.10) and any such sequence is quadratically convergent to $\hat{z}$, i.e. there is a constant $c > 0$ such that*

$$\|x^{k+1} - \hat{x}\|_{1,1} + \|p^{k+1} - \hat{p}\|_{1,1} + \|u^{k+1} - \hat{u}\|_1 \le c \left(\|x^k - \hat{x}\|_{1,1} + \|p^k - \hat{p}\|_{1,1} + \|u^k - \hat{u}\|_1 \right)^2.$$

Proof. Since problem $LP(z^k)$ has a solution and the generalized equation (5.10) represents the Pontryagin necessary optimality conditions for this problem, the iterate z^k exists for every k.

We will apply Theorem 5.1 with spaces $\mathcal{X}$ and $\mathcal{Y}$ (for X and Y) and mappings f and G (for φ and Φ).

An easy but cumbersome calculation (which we skip) shows that Assumption (A1") implies that the mapping $f : \mathcal{X} \to \mathcal{Y}$ is Fréchet differentiable with locally Lipschitz derivative. Thus condition (5.1) in Theorem 5.1 is satisfied with $\varphi = f$ and some constants R and L. Moreover, thanks to Assumptions (A1")–(A3"), Problem $LP(\hat{z})$ fulfills Assumptions (A1)-(A3) in Theorem 3.3. This implies (see Remark 3.4 and Footnote 1) that condition (5.2) in Theorem 5.1 is also fulfilled with some constant c. Then the convergence claimed in the present theorem follows from Theorem 5.1 with the neighborhood O defined as the open ball in $\mathcal{X}$ centered at $\hat{z}$ and with radius r, where r is defined in Theorem 5.1. Q.E.D.

Conclusion

This paper contributes to the regularity theory for Bolza-type optimal control problems with linear dynamics, quadratic in the state and linear in the control objective integrand, and a non-linear terminal term. Conditions for Lipschitz/Hölder sub-regularity and bi-metric regularity are obtained and the results are utilized for obtaining a convergence result for the Newton method applied to non-linear problems that are affine with respect to the control. One of this conditions, which is particularly restrictive, requires that the optimal control is of pure bang-bang type. Extensions of the regularity results and the Newton method to control-affine optimal control problems with singular arcs is an important open area.

References

[1] Adly S., Cibulka R., Ngai H. V.: Newton's method for solving inclusions using set-valued approximations, SIAM J. Optim. 25 (1), 159–184 (2015)

[2] Alt W., Baier R., Gerdts M., Lempio F.: Approximation of Linear Control Problems with Bang-Bang Solutions. Optimization. 62(1), 9–32 (2013)

[3] Alt W., Schneider C., Seydenschwanz M.: Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang-bang solutions. Appl. Math. and Comp. 287-288, 104–124 (2016)

[4] Aragon Artacho F. J., Mordukhovich B. S.: Enhanced metric regularity and Lipschitzian properties of variational systems, J. Global Optim. 50 (1) 145–167 (2011)

[5] Aronna M. S., Bonnans J. F., Martinon P.: A shooting algorithm for optimal control problems with singular arcs. J. Optim. Theory Appl. 158, 419–459 (2013)

[6] Bressan A., Piccoli B.: Introduction to the Mathematical Theory of Control. American Institute of Mathematical Sciences, 2007.

[7] Bonnans F. J. : Local analysis of Newton-type methods for variational inequalities and nonlinear programming, Appl. Math. Optim., 29, pp. 161–186 (1994)

[8] Cannarsa P., Sinestrari C.: Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Boston, MA: Birkhäuser Boston Inc., 2004

[9] Cibulka R., Dontchev A.L., Kruger A.Y.: Strong metric subregularity of mappings in variational analysis and optimization. J. Math. Anal. Appl. 457, 1247–1282 (2017)

[10] Dontchev A.L., Hager W.W.: Lipschitzian Stability in Nonlinear Control and Optimization. SIAM J. Control Optim., 31(3), 569–603 (1993)

[11] Dontchev A.L., Hager W.W., Veliov V.M.: Uniform convergence and mesh independence of the Newton method in optimal control. SIAM J. Control and Optim., 39(3), 961–980 (2000)

[12] Dontchev A.L., Malanowski K.: A Characterization of Lipschitzian Stability in Optimal Control. Calculus of variations and optimal control 411, 62–76 (2000)

[13] Dontchev A.L., Rockafellar R.T.: Implicit Functions and Solution Mappings: A View from Variational Analysis. Second edition. Springer, New York (2014)

[14] Ekeland. I.: On the variational principle. J. Math. Anal. and Appl., 47, 324–353 (1974)

[15] Felgenhauer U.: A Newton-type method and optimality test for problems with bang-singular-bang optimal control. Pure Appl. Funct. Anal., 1 (2), 197–215 (2016)

[16] Felgenhauer U.: On Stability of Bang-Bang Type Controls. SIAM J. Control Optim., 41(6), 1843–1867 (2003)

[17] Felgenhauer U.: Stability analysis of variational inequalities for bang-singular-bang controls. Control and Cybernetics, 42(3), 557–592 (2013)

[18] Felgenhauer U., Poggiolini L., Stefani G.: Optimality and stability result for bang-bang optimal controls with simple and double switch behaviour. Control and Cybernetics, 38, 1305–1325 (2009)

[19] Haunschmied J.L., Pietrus A., Veliov V.M.: The Euler Method for Linear Control Systems Revisited. Proceedings of the 9th International Conference on Large-Scale Scientific Computing, 90–97 (2013)

[20] Ioffe A.D.: Metric Regularity. Theory and Applications - a survey. arXiv:1505.07920. `https://arxiv.org/abs/1505.07920`, 24 Oct 2015.

[21] Ledzewicz U., Marriott J., Maurer H., Schättler H.: Realizable protocols for optimal administration of drugs in mathematical models for anti-angiogenic treatment. Mathematical Medicine and Biology: A Journal of the IMA, 27(2), 157–179 (2010)

[22] Ledzewicz U., Schättler H.: Optimal Bang-Bang Controls for a Two-Compartment Model in Cancer Chemotherapy. Journal of Optimization Theory and Applications, 114(3), 609–637 (2002)

[23] Ledzewicz U., Schättler H.: Geometric Optimal Control: Theory, Methods and Examples. Springer, New York (2012)

[24] Maurer H. and Osmolovskii N.P.: Second Order Sufficient Conditions for Time-Optimal Bang-Bang Control SIAM J. Control Optim., 42(6), 2239–2263 (2004)

[25] Maurer H. and Osmolovskii N.P.: Applications to Regular and Bang-Bang Control: Second-Order Necessary and Sufficient Optimality Conditions in Calculus of Variations and Optimal Control. SIAM Advances in Design and Control, vol. 24 (2012)

[26] Pietrus A., Scarinci T., Veliov V.M.: High Order Discrete Approximations to Mayer's Problems for Linear Systems. SIAM J. Control Optim. 56 (1), 102–119 (2018)

[27] Quincampoix M., Veliov V.M.: Metric Regularity and Stability of Optimal Control Problems for Linear Systems. SIAM J. Control Optim., 51(5), 4118–4137 (2013)

[28] Robinson S.M.: Strongly Regular Generalized Equations. Mathematics of Operations Research, 5(1), 43–62 (1980)

[29] Schneider C., Wachsmuth G.: Regularization and discretization error estimates for optimal control of ODEs with group sparsity. ESAIM: Control, Optimisation and Calculus of Variations (2017), doi:10.1051/cocv/2017049

[30] Seydenschwanz M.: Convergence results for the discrete regularization of linear-quadratic control problems with bang-bang solutions. Comput. Optim. Appl., 61(3), 731–760 (2015)

[31] Veliov V.M.: On the Convexity of Integrals of Multivalued Mappings: Applications in Control Theory. Journal of Optimization Theory and Applications, 54(3), 541–563 (1987)

[32] Veliov V.M.: Error analysis of discrete approximations to bang-bang optimal control problems: the linear case. Control and Cybernetics, 34(3), 967–982 (2005)

On the Convergence of the Gradient Projection Method for Convex Optimal Control Problems with Bang-Bang Solutions.*

J. Preininger† and P. T. Vuong‡

January 9, 2018

Abstract

We revisit the gradient projection method in the framework of nonlinear optimal control problems with bang-bang solutions. We obtain the strong convergence of the iterative sequence of controls and the corresponding trajectories. Moreover, we establish a convergence rate, depending on a constant appearing in the corresponding switching function and prove that this convergence rate estimate is sharp. Some numerical illustrations are reported confirming the theoretical results.

Keywords: Gradient projection method, Strong convergence, Convergence rate, Optimal control, Bang-bang control.

Mathematics Subject Classification (2010). 47J20, 49J15, 49M05, 90C25, 90C30.

1 Introduction

Numerical solution methods for various optimal control problems have been investigated during the last decades [9, 8, 10, 11, 6]. However, in most of the literature, the optimal controls are assumed to be at least Lipschitz continuous. This assumption is rather strong, as whenever the control appears linearly in the problem, the lack of coercivity typically leads to discontinuities of the optimal controls. Recently, optimal control problems with bang-bang solutions attract more attention. Stability and error analysis of bang-bang controls can be found in [14, 32, 26]. Euler discretizations for linear-quadratic optimal control problems with bang-bang solutions were studied in [1, 2, 29, 5]. Higher order schemes for linear and linear-quadratic optimal control problems with bang-bang solutions were developed in [24, 27].

On the other hand, among many traditional solution methods in optimization, projection-type methods are widely applied because of their simplicity and efficiency [13, 15, 31].

*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, jakob.preininger@tuwien.ac.at.

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

Recently, the gradient projection method has been reconsidered for solving general optimal control problems [22, 28]. Under some suitable conditions, it was proved that the control sequence converges weakly to an optimal control and the corresponding trajectory sequence converges strongly to an optimal trajectory. However, no convergence rate result has been established.

In this paper, we study the gradient projection method for optimal control problems with bang-bang solutions. In particular we consider the following problem

$$\text{minimize } \psi(x,u) := g(x(T)) + \int_0^T h(t,x(t),u(t))dt \tag{1.1}$$

subject to

$$\dot{x}(t) = f(t,x(t),u(t)) \text{ for a.e. } t \in [0,T], \quad x(0) = x_0, \tag{1.2}$$

and

$$u(t) \in U := [-1,1]^m \text{ for a.e. } t \in [0,T]. \tag{1.3}$$

Here $[0,T]$ is a fixed time horizon, admissible controls are all measurable functions $u : [0,T] \to U$, while $x(t) \in \mathbb{R}^n$ denotes the state of the system at time $t \in [0,T]$ and the functions $f : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are given.

Further we assume (see the next section for precise formulations) that the data are smooth enough, that the problem (1.1)-(1.3) is convex and that for the (unique) optimal control u^* the objective function fulfills a certain growth condition. In particular we show that this condition is satisfied in the bang-bang case if each component of the associated switching function satisfies a growth condition as given in [29, 25].

Under these assumptions, we prove that the control sequence actually converges strongly to the solution. Moreover, the convergence rates for both controls and states are provided, depending on the constant appearing in the growth condition for the switching function. An example is analysed showing that the estimation for these convergence rates is sharp.

The paper is organized as follows: In Section 2, we specify the assumptions we use and recall some facts which will be useful in the sequel. Section 3 discusses the convergence properties of the gradient projection method. Some numerical examples of linear-quadratic type are reported in Section 4 illustrating the results in the previous section. Some final remarks are given in the last section.

2 Preliminaries

In this section, we will clarify the assumptions used and recall some important facts which are necessary to establish our result.

By $\mathcal{U} := L^2([0,T],U)$ we denote the set of all admissible controls and if not stated otherwise $\|\cdot\|$ denotes the L^2-norm. The first two assumptions guarantee that the problem (1.1)-(1.3) is meaningful.

Assumption (A1). For any given control $u \in \mathcal{U}$ there is a unique solution $x = x(u)$ of (1.2) on $[0, T]$.

Assumption (A2). The problem (1.1)-(1.3) has a solution (x^*, u^*).

Now recall the Hamiltonian of (1.1)-(1.3) as

$$H(t, x, u, p) = \langle p, f(t, x, u)\rangle + h(t, x, u).$$

Then by the Pontryagin maximum principle there is an absolutely continuous function p^* such that (x^*, u^*, p^*) solves the adjoint equation

$$\begin{array}{rcl} \dot{p}(t) & = & -H_x(t, x(t), u(t), p(t)) = -f_x(t, x(t), u(t))^\top p(t) - h_x(t, x(t), u(t))^\top \text{ for a.e. } t \in [0, T] \\ p(T) & = & \nabla g(x(T)), \end{array} \tag{2.1}$$

and for every $u \in U$

$$\langle H_u(t, x^*(t), u^*(t), p^*(t)), u - u^*(t)\rangle \geq 0 \text{ for a.e. } t \in [0, T].$$

We define $J : \mathcal{U} \to \mathbb{R}$ via $J(u) := \psi(x(u), u)$, where $x(u)$ is the solution (1.2). Then we have the following useful formula for the gradient of J (see, e.g. [31, 22]).

$$\nabla J(u)(t) = H_u(t, x(t), u(t), p(t)) = f_u(t, x(t), u(t))^\top p(t) + h_u(t, x(t), u(t))^\top, \tag{2.2}$$

where x and p are the unique solution of (1.2) and (2.1) depending on $u \in \mathcal{U}$.

Assumption (A3). The objective function J is continuously differentiable on $\mathcal{U}$ with Lipschitz derivative.

We denote by L the Lipschitz modulus of the gradient ∇J of J and write $J^* := J(u^*)$ for its optimal value. The following result is well known (see e.g. [23, Lemma 1.30]).

Lemma 2.1. *Suppose that (A3) is fulfilled. Then for every $u, v \in \mathcal{U}$ the following estimation holds*

$$J(v) - J(u) - \langle \nabla J(u), v - u\rangle \leq \frac{L}{2}\|v - u\|^2.$$

Assumptions (A1)-(A3) are common in optimal control. For example the following two assumptions (B1)-(B2) imply (A1)-(A3) (cf. [22])

Assumption (B1). The functions f and h are of the form $f(t, x, u) = f_0(x) + f_1(x)u$ and $h(t, x, u) = h_0(x) + \langle h_1(x), u\rangle$ respectively, where $f_0 : \mathbb{R}^n \to \mathbb{R}^n$, $f_1 : \mathbb{R}^n \to \mathbb{R}^{n\times m}$, $h_0 : \mathbb{R}^n \to \mathbb{R}$ and $h_1 : \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable.

Assumption (B2). There exists $c \geq 0$ such that for every $x \in \mathbb{R}^n$ and $u \in U$:

$$\langle x, f(t, x, u)\rangle \leq c(1 + |x|^2).$$

Additionally we assume the following.

Assumption (A4). The objective function J is convex.

Note that if the set $\mathcal{F}$ of admissible pairs is convex this assumption is equivalent to the statement that the function ψ is convex on $\mathcal{F}$. In particular this is the case if f is affine (i.e. f is of the form $f(t,x,u) = A(t)x + B(t)u + d(t)$) as in [29, 25].

Further we will assume a growth condition for J that is similar to (4.7) in [3].

Assumption (A5). For a solution u^* of (1.1)-(1.3) there are constants $\beta > 0$ and $\theta \geq 0$ such that for every $u \in \mathcal{U}$ we have

$$J(u) - J(u^*) \geq \beta \|u - u^*\|^{2\theta+2}.$$

Note that in particular (A5) implies that the solution u^* is unique.

Remark 2.2. For coercive optimal control problems (in the sense of [12]) Assumptions (A1)-(A4) are fulfilled as well as (A5) for $\theta = 0$. In these problems the objective function J however is even strongly convex and therefore one can apply known results (e.g. [21, Theorem 2.1.15]) directly to show linear convergence of the gradient projection method in this case.

In the following we will show that Assumption (A5) is fulfilled for bang-bang controls with no singular arcs. We recall that in the case of bang-bang controls the function $\sigma^* := H_u(\cdot, x^*, u^*, p^*)$ is called *switching function* corresponding to the triple (x^*, u^*, p^*). For every $j \in \{1, \dots, m\}$ denote by σ_j^* its j-th component. The following assumption says that the switching function σ^* satisfies a growth condition around the switching points, which implies that u^* is strictly bang-bang.

Assumption (B3). There exist real numbers $\theta, \alpha, \tau > 0$ such that for all $j \in \{1, \dots, m\}$ and $s \in [0,T]$ with $\sigma_j^*(s) = 0$ we have

$$|\sigma_j^*(t)| \geq \alpha |t - s|^\theta \quad \forall t \in [s - \tau, s + \tau] \cap [0, T].$$

Assumption (B3) plays the main role in the study of regularity, stability and error analysis of discretization techniques for optimal control problems with bang-bang solutions. Many variations of this assumption are used in the literature about bang-bang controls. To our knowledge the first assumption of this type was introduced by Felgenhauer [14] for continuously differentiable switching functions with $\theta = 1$ to study the stability of bang-bang controls. Alt et. al. [1, 2, 4] used a slightly stronger version of (B3) with $\theta = 1$, that additionally excludes the endpoints 0 and T as zeros of the switching function, to investigate the error bound for Euler approximation of linear-quadratic optimal control problems with bang-bang solutions. Quincampoix and Veliov [26] used a rank condition which implies (B3) (including cases where $\theta \neq 1$) to obtain the metric regularity and stability of Mayer problems for linear systems. Seydenschwanz [29], Preininger et. al. [25], Pietrus, Scarinci and Veliov [24, 27] used this assumption in the study of metric (sub)-regularity, stability and error estimate for discretized schemes of linear-quadratic optimal control problems with bang-bang solutions.

To prove that (B3) implies (A5) we need the following lemma, which is a simplified version of [29, Lemma 1.3] (see also, [1, Lemma 4.1]).

Lemma 2.3. *Let Assumptions (A1)-(A2) be fulfilled and let u^* be a solution of* (1.1)-(1.3) *such that (B3) is fulfilled for some $\theta > 0$. Then there exists constants $\beta > 0$ such that for any feasible $u \in \mathcal{U}$ it holds*

$$\int_0^T \sigma^*(t)^T \left(u(t) - u^*(t)\right) dt \geq \beta \|u - u^*\|_1^{\theta+1},$$

where $\|\cdot\|_1$ is the L^1-norm.

Proposition 2.4. *Let Assumptions (A1)-(A2) and (A4) be fulfilled and let u^* be a solution of* (1.1)-(1.3) *such that (B3) is fulfilled. Then (A5) holds.*

Proof. From Assumption (A4) and (2.2) we obtain

$$J(u) - J(u^*) \geq \langle \nabla J(u^*), u - u^* \rangle = \int_0^T \sigma^*(t)^T \left(u(t) - u^*(t)\right) dt. \tag{2.3}$$

Since $\|\cdot\|^2 \leq C\|\cdot\|_1$ on $\mathcal{U}$ for some constant $C > 0$, from Lemma 2.3 there exists $\beta > 0$ such that

$$\int_0^T \sigma^*(t)^T \left(u(t) - u^*(t)\right) dt \geq \beta \|u - u^*\|_1^{\theta+1} \geq \frac{\beta}{C^{\theta+1}} \|u - u^*\|^{2\theta+2}. \tag{2.4}$$

Combining (2.3) and (2.4) we obtain (A5). Q.E.D.

To define the gradient projection method in the next chapter we will need the following notion of a projection. For each $u \in \mathcal{U}$, there exists a unique point in $\mathcal{U}$ (see [17, p. 8]), denoted by $P_{\mathcal{U}}(u)$, such that

$$\|u - P_{\mathcal{U}}(u)\| \leq \|u - v\| \quad \forall v \in \mathcal{U}.$$

It is well known [17, Theorem 2.3] that the projection operator can be characterized by

$$\langle u - P_{\mathcal{U}}(u), v - P_{\mathcal{U}}(u) \rangle \leq 0 \quad \forall v \in \mathcal{U}. \tag{2.5}$$

Further to establish the convergence rate of the gradient projection method, we will need the following lemmas.

Lemma 2.5. *[18, Lemma 7.1] Let $\alpha > 0$ and let $\{\delta_k\}_{k=0}^{\infty}$ and $\{s_k\}_{k=0}^{\infty}$ be two sequences of positive numbers satisfying the conditions*

$$s_{k+1}(\delta_k s_{k+1}^{\alpha} + 1) \leq s_k \quad \forall k \in \mathbb{N}.$$

Then there is a number $\gamma > 0$ such that

$$s_k \leq \left(s_0^{-\alpha} + \gamma \sum_{i=0}^{k-1} \min\{\delta_i, \delta_i^{\frac{\alpha}{\alpha+1}}\} \right)^{-\frac{1}{\alpha}} \quad \forall k \in \mathbb{N}.$$

In particular, we have $\lim_{k\to\infty} s_k = 0$ whenever $\sum_{k=0}^{\infty} \delta_k = \infty$.

Lemma 2.6. *[7, Lemma 3.2] Let* $\{\alpha_k\}, \{s_k\}$ *be sequences in* $\mathbb{R}_+$ *satisfying*

$$\sum_{i=0}^{\infty} \alpha_k s_k < \infty,$$

the sequence $\{\alpha_k\}$ *is non-summable and the sequence* $\{s_k\}$ *is decreasing. Then*

$$s_k = o\left(\frac{1}{\sum_{i=0}^{k} \alpha_i}\right),$$

where the o-notation means that $s_k = o(1/t_k)$ *if and only if* $\lim_{k\to\infty} s_k t_k = 0$.

3 Convergence Analysis

We consider the following Gradient Projection Method (GPM):

Algorithm GPM.

Step 0: Choose a sequence $\{\lambda_k\}$ of positive real numbers and an initial control $u_0 \in \mathcal{U}$. Set $k = 0$.

Step 1: Compute the gradient $\nabla J(u_k)(t) := f_u(t, x_k(t), u_k(t))^\top p_k(t) + h_u(t, x_k(t), u_k(t))^\top$ by solving the following differential equations

$$\begin{aligned} \dot{x}_k(t) &= f(t, x_k(t), u_k(t)), \quad x_k(0) = x_0; \\ \dot{p}_k(t) &= -f_x(t, x_k(t), u_k(t))^\top p_k(t) - h_x(t, x_k(t), u_k(t))^\top, \quad p_k(T) = \nabla g(x_k(T)). \end{aligned} \tag{3.1}$$

Step 2: Compute

$$u_{k+1} = P_{\mathcal{U}}(u_k - \lambda_k \nabla J(u_k)). \tag{3.2}$$

Step 3: If $u_{k+1} = u_k$ then Stop. Otherwise replace k by $k+1$ and go to **Step 1**.

It is known (see e.g. [21, Theorem 2.1.14]) that for J continuously differentiable with Lipschitz derivative the gradient (projection) method has the convergence rate $O(\frac{1}{k})$ in terms of the objective value. I.e. that

$$J(u_k) - J^* = O(\frac{1}{k}). \tag{3.3}$$

For the strongly convex objective function, it is known that the iterative sequence $\{u_k\}$ converges linearly to the unique solution. However, it is not possible to show convergence for the iterative sequence $\{u_k\}$ for the general convex case. Here, thanks to Assumptions (A1)-(A5), we are able to prove that the iterative sequence $\{u_k\}$ generated by the GPM converges strongly to an optimal control. Moreover, the convergence rate is established, depending on the constants θ appearing in Assumption (A5).

The following estimate will be used repeatedly in our convergence analysis.

Proposition 3.1. *Let Assumption (A1)-(A4) be satisfied, let u^* be a solution of (1.1)-(1.3) such that Assumption (A5) is fulfilled with some $\theta > 0$ and $\beta > 0$. Then for all $k \in \mathbb{N}$, the following estimate holds*

$$\|u_{k+1} - u^*\|^2 \leq \|u_k - u^*\|^2 - (1 - \lambda_k L) \|u_{k+1} - u_k\|^2 - 2\lambda_k \beta \|u_{k+1} - u^*\|^{2\theta+2}. \tag{3.4}$$

Proof. Since $u_{k+1} = P_{\mathcal{U}}(u_k - \lambda_k \nabla J(u_k))$, it follows from (2.5) that

$$\langle u_k - \lambda_k \nabla J(u_k) - u_{k+1}, u - u_{k+1} \rangle \leq 0 \quad \forall u \in \mathcal{U}. \tag{3.5}$$

Substituting $u = u^* \in \mathcal{U}$ into the latter inequality yields

$$\langle u_k - \lambda_k \nabla J(u_k) - u_{k+1}, u^* - u_{k+1} \rangle \leq 0,$$

or equivalently

$$\langle u_k - u_{k+1}, u^* - u_{k+1} \rangle \leq \lambda_k \langle \nabla J(u_k), u^* - u_{k+1} \rangle.$$

This implies that

$$\begin{aligned}
\|u_{k+1} - u^*\|^2 &= \|u_k - u^*\|^2 + 2 \langle u_k - u^*, u_{k+1} - u_k \rangle + \|u_{k+1} - u_k\|^2 \\
&= \|u_k - u^*\|^2 + 2 \langle u_{k+1} - u^*, u_{k+1} - u_k \rangle - \|u_{k+1} - u_k\|^2 \\
&\leq \|u_k - u^*\|^2 + 2\lambda_k \langle \nabla J(u_k), u^* - u_{k+1} \rangle - \|u_{k+1} - u_k\|^2 \\
&= \|u_k - u^*\|^2 \\
&\quad -2\lambda_k \left[\langle \nabla J(u_k), u_{k+1} - u^* \rangle + \frac{L}{2} \|u_{k+1} - u_k\|^2 + \left(\frac{1}{2\lambda_k} - \frac{L}{2} \right) \|u_{k+1} - u_k\|^2 \right] \\
&= \|u_k - u^*\|^2 - (1 - \lambda_k L) \|u_{k+1} - u_k\|^2 \\
&\quad -2\lambda_k \left[\langle \nabla J(u_k), u_k - u^* \rangle + \langle \nabla J(u_k), u_{k+1} - u_k \rangle + \frac{L}{2} \|u_{k+1} - u_k\|^2 \right].
\end{aligned} \tag{3.6}$$

Since J has Lipschitz derivative, we have from Lemma 2.1 that

$$J(v) - J(u) - \langle \nabla J(u), v - u \rangle \leq \frac{L}{2} \|v - u\|^2 \quad \forall u, v \in \mathcal{U}.$$

Substituting $u = u_k$ and $v = u_{k+1}$ into the last inequality yields

$$- \langle \nabla J(u_k), u_{k+1} - u_k \rangle - \frac{L}{2} \|u_{k+1} - u_k\|^2 \leq J(u_k) - J(u_{k+1}). \tag{3.7}$$

Moreover, since J is convex, we obtain

$$- \langle \nabla J(u_k), u_k - u^* \rangle \leq J(u^*) - J(u_k) \tag{3.8}$$

Combining (3.6), (3.7) and (3.8) gives

$$\|u_{k+1} - u^*\|^2 \leq \|u_k - u^*\|^2 - (1 - \lambda_k L) \|u_{k+1} - u_k\|^2 - 2\lambda_k \left(J(u_{k+1}) - J(u^*) \right). \tag{3.9}$$

Using Assumption (A5) we obtain

$$\|u_{k+1} - u^*\|^2 \leq \|u_k - u^*\|^2 - (1 - \lambda_k L)\,\|u_{k+1} - u_k\|^2 - 2\lambda_k \beta \|u_{k+1} - u^*\|^{2\theta+2},$$

which is (3.4). Q.E.D.

We are now in the position to establish the strong convergence and the convergence rate of $\{u_k\}$ to a solution.

Theorem 3.2. *Let Assumptions (A1)-(A4) be satisfied, let u^* be a solution of (1.1)-(1.3) such that Assumption (A5) is fulfilled with some $\theta > 0$. Let the sequence $\{\lambda_k\}$ be chosen such that*

$$0 < \lambda_{\min} \leq \lambda_k \leq \frac{1}{L} \quad \forall k \in \mathbb{N}.$$

Then we have

(i) $\|u_k - u^*\|^2 \leq \eta k^{-\frac{1}{\theta}}$, for all k, where $\eta > 0$ is a constant;

(ii) The sequence $\{J(u_k)\}$ is monotonically decreasing. Moreover $\sum_{k=0}^{\infty} (J(u_k) - J(u^*)) < +\infty$.

Proof. We first prove that $\{u_k\}$ converges strongly to u^*. From (3.4) and $0 < \lambda_{\min} \leq \lambda_k \leq \frac{1}{L}$, the sequence $\{\|u_k - u^*\|\}$ is decreasing and bounded from below by 0, and therefore it converges. Moreover, since

$$2\lambda_{\min}\beta \|u_{k+1} - u^*\|^{2\theta+2} \leq \|u_k - u^*\|^2 - \|u_{k+1} - u^*\|^2 \tag{3.10}$$

we conclude that $\{\|u_k - u^*\|\}$ converges to 0, which means $\{u_k\}$ converges strongly to u^*.

Now we can apply Lemma 2.5 for $s_k = \|u_k - u^*\|^2$, $\alpha = \theta$ and $\delta_k = 2\lambda_{\min}\beta$ to obtain the convergence rate (i) for $\{\|u_k - u^*\|\}$.

Substituting $u = u_k$ in (3.5) implies

$$\lambda_k \langle \nabla J(u_k), u_k - u_{k+1} \rangle \geq \|u_{k+1} - u_k\|^2. \tag{3.11}$$

Combining (3.7) and (3.11) we get

$$J(u_{k+1}) - J(u_k) \leq \left(\frac{L}{2} - \frac{1}{\lambda_k}\right) \|u_{k+1} - u_k\|^2 \leq 0. \tag{3.12}$$

Hence the sequence $\{J(u_k)\}$ is monotonically decreasing. Now from (3.9) and $0 < \lambda_{\min} \leq \lambda_k \leq \frac{1}{L}$ we have

$$2\lambda_{\min} (J(u_k) - J(u^*)) \leq \|u_{k-1} - u^*\|^2 - \|u_k - u^*\|^2 \quad \forall k \in \mathbb{N}.$$

Summing this inequality from 0 to $i - 1$ we obtain

$$\sum_{k=0}^{i-1} (J(u_k) - J(u^*)) \leq \frac{1}{2\lambda_{\min}} \left(\|u_0 - u^*\|^2 - \|u_i - u^*\|^2\right).$$

Finally, taking the limit as $i \to \infty$, we obtain (ii). Q.E.D.

Remark 3.3. From (ii) in Theorem 3.2, we can conclude that $J(u_k) - J(u^*) = o(\frac{1}{k})$, which significantly improves the error estimate $J(u_k) - J(u^*) = O(\frac{1}{k})$ in (3.3).

The following example illustrates that the estimation (i) in Theorem 3.2 cannot be improved when λ_k is bounded from below by a constant $\lambda_{\min}$.

Example 3.4. Consider the following optimal control problem

$$\begin{array}{ll} \text{minimize} & \int_0^T \sigma(t)u(t)dt \\ \text{subject to} & u(t) \in U := [-1,1]^m, \end{array} \tag{3.13}$$

where σ is any continuous function fulfilling Assumption (B3). Then $\nabla J(u)(t) = \sigma(t)$ is independent of u and the optimal control is given by $u^*(t) = -sgn(\sigma(t))$. Starting the GPM with $u_0 \equiv 0$ and $\lambda_k = \lambda$ for some $\lambda \in \mathbb{R}^+$ we get

$$u_k(t) = \begin{cases} 1, & \text{if} \quad -k\lambda\sigma(t) > 1, \\ -k\lambda\sigma(t), & \text{if} \quad -1 \leq -k\lambda\sigma(t) \leq 1, \\ -1, & \text{if} \quad -k\lambda\sigma(t) < -1. \end{cases}$$

In the special case $\sigma(t) = t^\theta$, we therefore have $u_k(t) = \max\{-1, -k\lambda t^\theta\}$. This implies that for $k > \frac{1}{\lambda T^\theta}$, we have

$$\|u_k(t) - u^*(t)\|^2 = \int_0^{(k\lambda)^{-\frac{1}{\theta}}} (1 - k\lambda t^\theta)^2 dt = (k\lambda)^{-\frac{1}{\theta}}(1 - \frac{2}{\theta+1} + \frac{1}{2\theta+1}) = Ck^{-\frac{1}{\theta}}.$$

For the objective value we get

$$J(u_k) - J(u^*) = \left(\frac{1}{\theta+1} - \frac{1}{2\theta+1}\right)(k\lambda)^{-1-\frac{1}{\theta}}, \tag{3.14}$$

which is stronger than (ii). It remains unknown whether in the general case the estimation (ii) can be improved to an estimation similar to (3.14).

Using the stronger Assumptions (B1)-(B2) the convergence rate of the corresponding trajectories can be obtained as a corollary of Theorem 3.2 and [22, Lemma 2].

Corollary 3.5. *Let Assumptions (B1)-(B2) and (A4) be satisfied and let* (x^*, u^*) *be a solution of* (1.1)-(1.3) *such that assumption (A5) is fulfilled with some* $\theta > 0$. *Further suppose that* $\lambda_k \in [\lambda_{\min}, 1/L] \subset (0, 1/L]$. *Then the sequence* $\{x_k(t)\}$ *of trajectories converges strongly to the solution* x^*. *Moreover, there exists a positive constant* C *such that for all* k *it holds,*

$$\|x_k - \hat{x}\|_c \leq Ck^{-\frac{1}{2\theta}},$$

where $\|x(\cdot)\|_c = \max_{t\in[0,T]} |x(t)|$.

When the Lipschitz modulus L is difficult to estimate, one can consider the non-summable diminishing stepsizes as follow.

Theorem 3.6. *Let assumption (A1)-(A4) be satisfied, let u^* be a solution of (1.1)-(1.3) such that assumption (A5) is fulfilled with some $\theta > 0$. Let the sequence $\{\lambda_k\}$ be chosen such that*

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

Then the sequence $\{u_k\}$ converges strongly to u^*. Moreover there exists $N > 0$ such that for all $k \geq N$, it holds

(i) $\|u_k - u^*\|^2 \leq C\mu_k^{-\frac{1}{\theta}}$

(ii) $J(u_k) - J(u^*) = o\left(\frac{1}{\mu_k}\right)$,

where $\mu_k := \sum_{i=N}^{k-1} \lambda_i$ and C is a constant.

Proof. Let $\beta > 0$ be as in Proposition 3.1. Since $\lim_{k\to\infty} \lambda_k = 0$, there exists $N > 0$ such that for all $k \geq N$ we have $1 - \lambda_k L > 0$ and $2\lambda_k\beta < 1$. From (3.4) we have that $\{\|u_k - u^*\|\}$ is decreasing, therefore it converges. Moreover

$$2\lambda_k\beta\|u_{k+1} - u^*\|^{2\theta+2} \leq \|u_k - u^*\|^2 - \|u_{k+1} - u^*\|^2 \quad \forall k \geq N.$$

Using Lemma 2.5 with $s_k = \|u_{k+N} - u^*\|^2$, $\alpha = \theta$ and $\delta_k := 2\lambda_{k+N}\beta$ we get that there exists $\gamma > 0$ such that

$$\|u_k - u^*\|^2 \leq \left(\|u_N - u^*\|^{-2\theta} + \gamma \sum_{i=N}^{k-1} \lambda_i\right)^{-\frac{1}{\theta}} \quad \forall k \geq N,$$

which shows (i).

From (3.9), we have

$$2\lambda_k \left(J(u_{k+1}) - J(u^*)\right) \leq \|u_k - u^*\|^2 - \|u_{k+1} - u^*\|^2 \quad \forall k \geq N.$$

leading to

$$\sum_{k=N}^{\infty} \lambda_k \left(J(u_{k+1}) - J(u^*)\right) < \infty.$$

Applying Lemma 2.6 with $\alpha_k = \lambda_{N+k}$ and $s_k = J(u_{N+k}) - J(u^*)$ we obtain (ii). Q.E.D.

Using the same example as above we can again show that the estimation (i) cannot be improved.

Example 3.7. Consider the problem (3.13) with $\sigma(t) := t^\theta$ again. As before we use GPM with $u_0 \equiv 0$ but now with non-constant λ_k. Denoting $\mu_k := \sum_{i=0}^{k-1} \lambda_i$ we get $u_k(t) = \max\{-1, -\mu_k t^\theta\}$. Hence for k big enough such that $\mu_k > \frac{1}{T^\theta}$ we have

$$\|u_k(t) - u^*(t)\|^2 = \int_0^{\mu_k^{-\frac{1}{\theta}}} (1 - \mu_k t^\theta)^2 dt = \mu_k^{-\frac{1}{\theta}} (1 - \frac{2}{\theta+1} + \frac{1}{2\theta+1}) = C\mu_k^{-\frac{1}{\theta}}$$

and

$$J(u_k) - J(u^*) = \left(\frac{1}{\theta+1} - \frac{1}{2\theta+1}\right) \mu_k^{-1-\frac{1}{\theta}}.$$

Similar to Corollary 3.5 we obtain

Corollary 3.8. *Let Assumptions (B1)-(B2) and (A4) be satisfied and let (x^*, u^*) be a solution of (1.1)-(1.3) such that assumption (A5) is fulfilled with some $\theta > 0$. Further let the sequence $\{\lambda_k\}$ be chosen such that*

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

Then the sequence $\{x_k(t)\}$ of trajectories converges strongly to the solution x^*. Moreover, there exists a positive constant C such that for all k it holds,

$$\|x_k - \hat{x}\|_c \leq C\mu_k^{-\frac{1}{2\theta}}.$$

4 Numerical Illustrations

In this section, we present some numerical experiments for a class of optimal control problems with bang-bang solutions namely linear-quadratic problem, described as follow.

$$\begin{array}{ll} \text{minimize} & \psi(x,u) \\ \text{subject to} & \dot{x}(t) = A(t)x(t) + B(t)u(t) + d(t), \quad t \in [0,T], \\ & u(t) \in U := [-1,1]^m, \\ & x(0) = x_0, \end{array} \tag{4.1}$$

where

$$\psi(x,u) := \frac{1}{2}x(T)Qx(T) + q^\top x(T) + \int_0^T \left(\frac{1}{2}x(t)^\top W(t)x(t) + x(t)^\top S(t)u(t)\right) dt.$$

Here we use the classical Euler discretization where the error estimate can be found in [1, 2, 5]. We choose a natural number N and define the *mesh size* $h := T/N$. Since the optimal control is assumed to be bang-bang, we identify the discretized control $u^N := (u_0, u_1, \ldots, u_{N-1})$ with its piece-wise constant extension:

$$u^N(t) = u_i \text{ for } t \in [t_i, t_{i+1}),\ i = 0, 1, \ldots, N-1.$$

Moreover, we identify the discretized state $x^N := (x_0, x_1, \ldots, x_N)$ and costate $p^N := (p_0, p_1, \ldots, p_N)$ with its piece-wise linear interpolations

$$x^N(t) = x_i + \frac{t - t_i}{h}(x_{i+1} - x_i), \text{ for } t \in [t_i, t_{i+1}),\ i = 0, 1, \ldots, N-1$$

and

$$p^N(t) = p_i + \frac{t_i - t}{h}(p_{i-1} - p_i), \text{ for } t \in (t_{i-1}, t_i],\ i = N, N-1, \ldots, 1.$$

The Euler discretization of (1.1) is given by

$$\begin{array}{ll} \text{minimize} & \psi_N(x^N, u^N) \\ \text{subject to} & x_{i+1}^N = x_i^N + h\left[A(t_i)x_i^N + B(t_i)u_i^N + d(t_i)\right], \\ & x^N(0) = x_0, \\ & u_i^N \in U, \end{array} \tag{P_N}$$

where ψ_N is the cost function defined by

$$\psi_N(x^N, u^N) := \frac{1}{2}x_N^\top Q x_N + q^\top x_N + h\sum_{i=0}^{N-1}\left[\frac{1}{2}x_i^T W(t_i)x_i + x_i^T S(t_i)u_i\right].$$

Observe that (P_N) is a quadratic optimization problem over a polyhedral convex set, where the gradient projection method converges linearly, see e.g., [30]. This means that for each N, there exists $\rho_N \in (0,1)$ such that

$$\|u_{k+1}^N - u^{N*}\| \leq \rho_N \|u_k^N - u^{N*}\|, \quad \forall k \in \mathbb{N}.$$

In the following examples, we will consider various values of N which suggest that

$$\lim_{N\to\infty} \rho_N = 1.$$

This will confirm the sublinear rate obtained in Theorem 3.2. The codes are implemented in Matlab. We perform all computations on a windows desktop with an Intel(R) Core(TM) i7-2600 CPU at 3.4GHz and 8.00 GB of memory. Since ∇J is linear in u, one can roughly estimate its Lipschitz constant by $L = \|\nabla J(u_0)\|/\|u_0\|$. We choose starting control $u_0(t) = 1\,\forall t \in [0,T]$ and stepsize $\lambda = 1/L$. The stopping condition is $\|u_k^N - u_{k-1}^N\| \leq \epsilon$, where $\epsilon = 10^{-10}$.
The following example is taken from [27].

Example 4.1.

$$\begin{array}{ll} \text{minimize} & -by(1) + \int_0^1 \frac{1}{2}\left(x(t)\right)^2 dt \\ \text{subject to} & \dot{x}(t) = y(t), \quad x(0) = a \\ & \dot{y}(t) = u(t), \quad y(0) = 1. \\ & u(t) \in [-1,1]. \end{array} \tag{4.2}$$

Here, with appropriate values of a and b, there is a unique optimal solution 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 [27], we choose $a = 1, b = 0.1$, then $\tau = 0.492487520$ is a simple zero of the switching function. Therefore, $\theta = 1$ and the exact optimal control is

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

The convergence results for Example 4.1 with some different values of N are reported in Table 4.1. We can see that when N increases, ρ_N also increases and approaches 1. This means that we can only guarantee the sublinear convergence for the continuous problem. Figure 4.1 displays the optimal control and the optimal states when the discretized size $N = 50$.

The following second example is taken from [1, Example 6.1]

Table 4.1: Convergence rates for Example 4.1

N	10	20	50	100	200	500
ρ_N	0.7701	0.9181	0.9839	0.9902	0.9964	0.9976

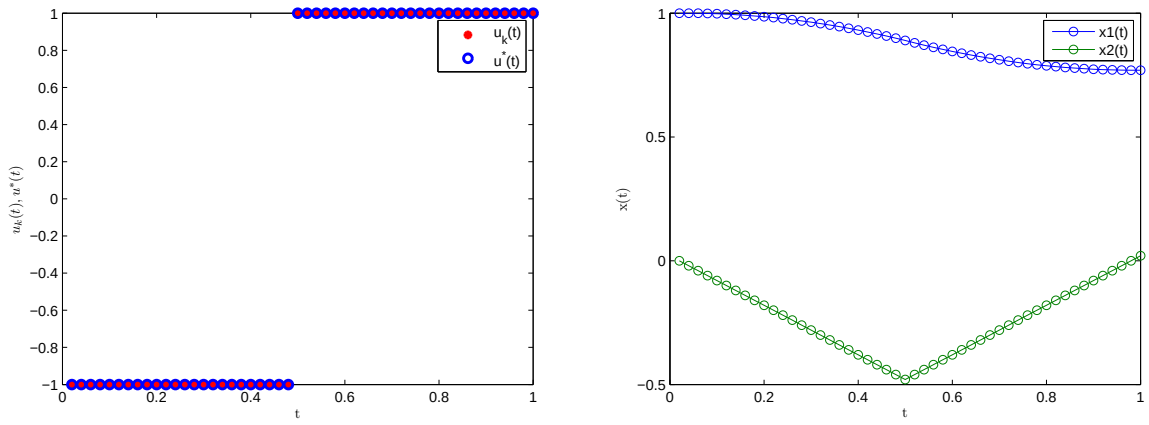


Figure 4.1: Optimal control (left) and optimal states (right) for $N = 50$.

Example 4.2.

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\left((x_1(5))^2 + (x_2(5))^2\right) \\ &\text{subject to} && \dot{x_1}(t) = x_2(t), \\ & && \dot{x_2}(t) = u(t), \quad \forall t \in [0,5]. \\ & && x_1(0) = 6, \quad x_2(0) = 1, \\ & && u(t) \in [-1,1]. \end{aligned} \tag{4.3}$$

The exact optimal control is given by

$$u^*(t) = \begin{cases} 1 & \text{if } t \in (\tau, 5] \\ -1 & \text{if } t \in (0, \tau], \end{cases}$$

where $\tau = 3.5174292$.

The convergence results for Example 4.2 with some different values of N are reported in Table 4.2. Again, we see that when N increases, ρ_N also increases and approaches 1. Figure 4.2 displays

Table 4.2: Convergence rates for Example 4.2

N	10	20	50	100	200	500
ρ_N	0.9625	0.9724	0.9905	0.9937	0.9943	0.9944

the optimal control and the optimal states for $N = 50$.

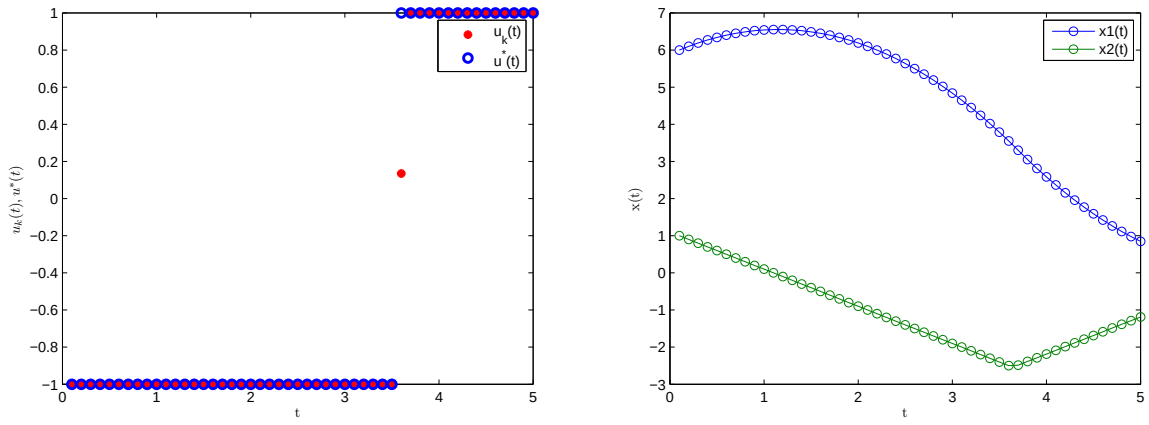


Figure 4.2: Optimal control (left) and optimal states (right) for Example 4.2 when $N = 50$.

In the next example, we consider a problem in which assumption (A5) is satisfied for $\theta \neq 1$ (see also [27, 29]).

Example 4.3. Here we present experiments with a family of problems satisfying assumption (A5) with various values of θ, given in [29]. Below, the time-interval is $[0, 1]$, the dimension of the state is $n = \theta + 1$ and the dynamics system depends on parameters s_j:

$$\begin{aligned} &\text{minimize} && x_1(1) \\ &\text{subject to} && \dot{x_j}(t) = s_j x_{j+1}(t) + u(t), \quad j = 1, ..., \theta \\ &&& \dot{x}_{\theta+1}(t) = u(t), \\ &&& x(0) = 0, \\ &&& u(t) \in [-1, 1]. \end{aligned} \tag{4.4}$$

For any natural number θ, the values of the parameters s_j are chosen as

$$s_j := -2(\theta - j + 1) \quad j = 1, ..., \theta.$$

Then assumption (A5) is satisfied with the constant θ [29] and exact optimal control is given by

$$u^*(t) = \begin{cases} 1 & \text{if } t \in [0, 1/2] \\ -1 & \text{if } t \in (1/2, 1] \end{cases}$$

if θ is odd, and $u^*(t) = -1$ if θ is even. The convergence results for Example 4.3 when $\theta = 2, 3$ with some different values of N are reported in Table 4.3. Figure 4.3 displays the approximate optimal controls after 1000 iterations for $N = 500$. It seems like the optimal control has θ switching points. This is to be expected since σ^* has a zero of order θ at $1/2$.

5 Concluding remarks

Note that the main results in Theorem 3.2 and Theorem 3.6 use Assumption (A5) which is more general than just the bang-bang case. For example Assumption (A5) is also satisfied in the strongly

Table 4.3: Convergence rates for Example 4.3

N	10	20	50	100	200	500
$\theta = 2$						
ρ_N	0.9418	0.9686	0.9865	0.9962	0.9953	0.9947
$\theta = 3$						
ρ_N	0.9245	0.9781	0.9936	0.9922	0.9968	0.9986

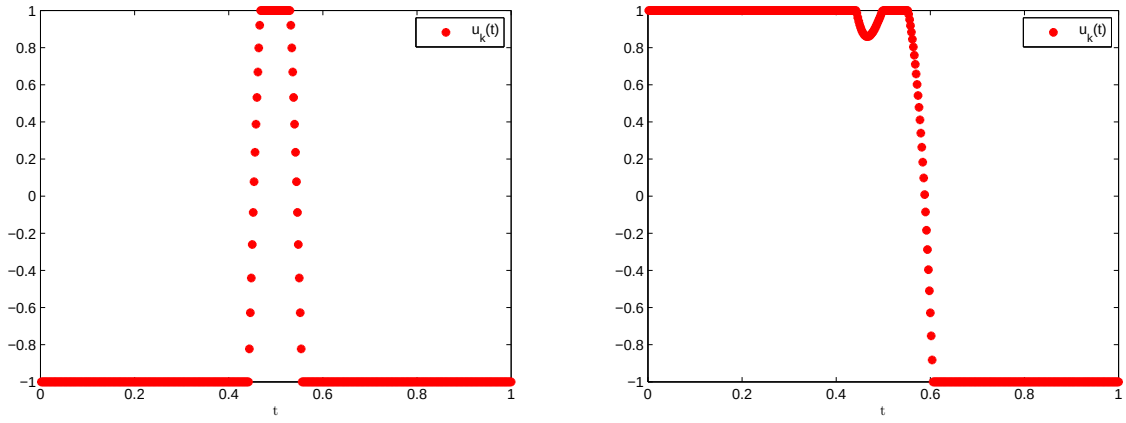


Figure 4.3: Approximate optimal controls after 1000 iterations when $\theta = 2$(left) and $\theta = 3$ (right) for Example 4.3 with $N = 500$.

convex case, where even better convergence results are known. Further it would be interesting to see under what assumptions our results still apply in the case of singular arcs. This is challenging due to the fact that currently there is no condition similar to the bang-bang Assumption (B3) that ensures Assumption (A5) and therefore remains as a topic for future research.

Acknowledgement

The authors thank Vladimir Veliov for introducing them to the topic and for fruitful discussions. They are also thankful to Ursula Felgenhauer and the two anonymous referees for constructive comments which helped improving the presentation of the paper significantly.

References

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

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

[3] Alt W., Felgenhauer U., Seydenschwanz M.: 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] Alt W., Schneider C., Seydenschwanz M.: An implicit discretization scheme for linear-quadratic control problems with bangbang solutions. Optimization Methods and Software, 29(3), 535–560 (2014)

[5] Alt W., Schneider C., Seydenschwanz M.: Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang-bang solutions. Appl. Math. Comp., 287-288, 104–124 (2016)

[6] Bonnans J.F., Festa A.: Error estimates for the Euler discretization of an optimal control problem with first-order state constraints. SIAM J. Numer. Anal., 55(2), 445–471 (2017)

[7] Dong Y.: Comments on the proximal point algorithm revisited. J. Optim. Theory Appl., 166(1), 343-349 (2015)

[8] Dontchev A.L.: An a priori estimate for discrete approximations in nonlinear optimal control. SIAM J. Control Optim., 34(4), 1315–1328 (1996)

[9] Dontchev A.L., Hager W.W.: Lipschitzian stability in nonlinear control and optimization. SIAM J. Control Optim., 31(3), 569–603 (1993)

[10] Dontchev A.L., Hager W.W., Malanowski K.: Error bounds for Euler approximation of a state and control constrained optimal control problem. Numerical Functional Analysis and Optimization, 21(5-6), 653–682 (2000)

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

[12] Dontchev A.L., Hager W.W., Veliov V.M.: Uniform convergence and mesh independence of Newton's method for discretized variational problems. SIAM J. Control Optim., 39(3), 961–980 (2000)

[13] Dunn J.C.: Global and asymptotic convergence rate estimates for a class of projected gradient processes. SIAM J. Control Optim., 19(3), 368–400 (1981)

[14] Felgenhauer U.: On stability of bang-bang type controls. SIAM J. Control Optim., 41(6), 1843–1867 (2003)

[15] Kelley C.T. and Sachs E.W.: Mesh Independence of the Gradient Projection Method for Optimal Control Problems. SIAM J. Control Optim., 30(2), 477–493 (1991)

[16] Khoroshilova E.V.: Extragradient-type method for optimal control problem with linear constraints and convex objective function. Optim. Lett., 7, 1193–1214 (2012)

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

[18] Li G. and Mordukhovich B.S.: Hölder metric subregularity with applications to proximal point method. SIAM J. Optim., 22(4), 1655–1684 (2012)

[19] Luo Z.Q., Tseng P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Annals of Operations Research, 46(1), 157–178 (1993)

[20] Luo Z.Q., Tseng P.: Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem. SIAM J. Optim., 2(1), 43–54 (1992)

[21] Nesterov Y.: Introductory Lectures on Convex Optimization. Springer (2004)

[22] Nikol'skii M.S.: Convergence of the gradient projection method in optimal control problems. Comp. Math. Model., 18(2), 148–156 (2007)

[23] Peypouquet J.: Convex Optimization in Normed Spaces: Theory, Methods and Examples. Springer (2015)

[24] Pietrus A., Scarinci T., Veliov V.M.: High order discrete approximations to Mayer's problems for linear systems. SIAM J. Control Optim., 56(1) 102-119 (2018)

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

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

[27] Scarinci T., Veliov V.M.: Higher-order numerical schemes for linear quadratic problems with bang-bang controls. Computational Optimization and Applications, 1-20 (2017), doi 10.1007/s10589-017-9948-z

[28] E. Scheiber.: On the gradient method applied to optimal control problem. Bulletin of the Transilvania University of Brasov, 7(56), 139–148 (2014)

[29] Seydenschwanz M.: Convergence results for the discrete regularization of linear-quadratic control problems with bang-bang solutions. Comput. Optim. Appl., 61(3), 731–760 (2015)

[30] Tuan H.N.: Linear convergence of a type of iterative sequences in nonconvex quadratic programming. J. Math. Anal. Appl., 423(2), 1311–1319 (2015)

[31] Vasil'ev F.P.: Optimization Methods [in Russian]. Factorial Press, Moscow (2002)

[32] Veliov V.M.: Error analysis of discrete approximation to bang-bang optimal control problems: the linear case. Control and Cybernetics, 34(3), 967–982 (2005)

Curriculum Vitae of Jakob Preininger

Personal data:

Name: Jakob Preininger MSc
Date of birth: June 20, 1988
Citizenship: Austria
E-mail: preininger.jakob@gmx.at

Academic position:

2014 - 2018: Project assistant at Vienna University of Technology for the FWF Project P 26640-N25 “Regularity, stability and computation of equilibria“

Education:

Since 2014: PhD in mathematics (optimal control theory)
at the Vienna University of Technology
supervisor: Prof. Vladimir Veliov

2011 - 2013: MSc in mathematics (algebra, number theory and combinatorics)
with distinction at the Faculty of Mathematics
at the University of Vienna

2007 - 2011: BSc in mathematics with distinction
at the Faculty of Mathematics at the University of Vienna

Teaching activities:

2010 - 2012: Tutor at the Faculty of Mathematics at the University of Vienna for Hilfsmittel aus der EDV (“Introduction to LATeX and Mathematica“, 2011 – 2012),
Lineare Algebra für PhysikerInnen (“Linear algebra for physicists“, 2010 – 2011) and Analysis (“Calculus“, 2009 – 2010)

Scholarships:

Performance scholarships awarded by the University of Vienna
for the academic years 2007-2008, 2008-2009 and 2010-2011

Extracurricular activities:

2017: Participation at the11th International Conference on Large-Scale Scientific Computations in Sozopol (Bulgaria)

2015: In the organizing team of the 13th Viennese Workshop on Optimal Control and Dynamic Games in Vienna (Austria)

2012: Participation at the seminar “Algorithms for Complex Multiplication over Finite Fields“ in Oberwolfach (Germany)

2008 - 2011: Annual participation at the Vojtech Jarnik International Mathematical Competition (VJIMC) in Ostrava (Czech Republic)

2008 - 2010: Participation at the International Mathematics Competition (IMC) in Blagoevgrad (Bulgaria, 2008 and 2010) and Budapest (Hungary, 2009)

2004 - 2006: Participation at the International Mathematical Olympiad (IMO) in Athens (Greece, 2004), Mérida (Mexico, 2005) and Ljubljana (Slovenia, 2006)

2004 - 2006: Participation at the Austrian Mathematical Olympiad (ÖMO) achieved 4th, 3rd and 2nd place

Publications:

- J. Preininger, Phan Tu Vuong: On the Convergence of the Gradient Projection Method for Optimal Control Problems with Bang-Bang Solutions. To appear in Computational Optimization and Applications (2018)
- J. Preininger, T. Scarinci, V.M. Veliov: Metric regularity properties in bang-bang type linear-quadratic optimal control problems. Preprint. (See http://orcos.tuwien.ac.at/fileadmin/t/orcos/ResearchReports/2017-04.pdf)
- J. Preininger, T. Scarinci, V.M. Veliov: On the Regularity of Linear-Quadratic Optimal Control Problems with Bang-Bang Solutions. In: Lirkov I., Margenov S. (eds) Large-Scale Scientific Computing. LSSC 2017. Lecture Notes in Computer Science, vol 10665. Springer, Cham (2018)
- R. Cibulka, A. L. Dontchev, J. Preininger, T. Roubal and V. Veliov: Kantorovich-type Theorems for Generalized Equations. Journal of Convex Analysis 25(2), 2018
- R. Cibulka, J. Preininger, T. Roubal: On uniform regularity and strong regularity. Preprint. (See https://orcos.tuwien.ac.at/fileadmin/t/orcos/Research_Reports/2018-01.pdf)

All of these are available online at https://orcos.tuwien.ac.at/research/research_reports/