Mathematical Programming (2021) 187:459–486
https://doi.org/10.1007/s10107-020-01488-z

FULL LENGTH PAPER

Series A

A sequential homotopy method for mathematical programming problems

Andreas Potschka[1] · Hans Georg Bock[1]

Received: 10 May 2019 / Accepted: 4 March 2020 / Published online: 23 March 2020

Abstract

We propose a sequential homotopy method for the solution of mathematical programming problems formulated in abstract Hilbert spaces under the Guignard constraint qualification. The method is equivalent to performing projected backward Euler timestepping on a projected gradient/antigradient flow of the augmented Lagrangian. The projected backward Euler equations can be interpreted as the necessary optimality conditions of a primal-dual proximal regularization of the original problem. The regularized problems are always feasible, satisfy a strong constraint qualification guaranteeing uniqueness of Lagrange multipliers, yield unique primal solutions provided that the stepsize is sufficiently small, and can be solved by a continuation in the stepsize. We show that equilibria of the projected gradient/antigradient flow and critical points of the optimization problem are identical, provide sufficient conditions for the existence of global flow solutions, and show that critical points with emanating descent curves cannot be asymptotically stable equilibria of the projected gradient/antigradient flow, practically eradicating convergence to saddle points and maxima. The sequential homotopy method can be used to globalize any locally convergent optimization method that can be used in a homotopy framework. We demonstrate its efficiency for a class of highly nonlinear and badly conditioned control constrained elliptic optimal control problems with a semismooth Newton approach for the regularized subproblems.

Keywords Mathematical programming · Hilbert space · Globalization · Projected gradient flow · Homotopy methods

We gratefully acknowledge support by the German Federal Ministry of Education and Research under Grants MOPhaPro (05M16VHA) and MOReNet (05M18VHA). HGB also gratefully acknowledges support by the German Research Foundation within the priority program DFG-SPP 1962. We thank Paul Manns for his helpful comments on an earlier version of the manuscript.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10107-020-01488-z.

Extended author information available on the last page of the article

Mathematics Subject Classification 49M05 · 49M37 · 58C15 · 65K05 · 65K15 · 90C30

1 Introduction

Let X and Y be real Hilbert spaces and $C \subseteq X$ a nonempty closed convex set. Let the nonlinear objective function $\phi : X \to \mathbb{R}$ and the nonlinear constraint function $c : X \to Y$ be twice continuously Fréchet differentiable. We consider the mathematical programming problem

$$\min \phi(x) \quad \text{over } x \in C \quad \text{subject to } c(x) = 0. \tag{1}$$

This formulation is equivalent to a more prevalent formulation that allows $c(x) \in C_c$ for some nonempty closed convex set C_c [by the use of slack variables $s \in Y$ via $c(x) - s = 0$ and $(x, s) \in C \times C_c$]. Further restrictions on the overall setting are stated in Sect. 1.5 after we settle the notation in Sect. 1.4.

This setting naturally comprises finite dimensional problems (also known as Nonlinear Programming Problems, NLPs) of the form

$$\min_{x \in \mathbb{R}^n} \phi(x) \quad \text{subject to } x^{\mathrm{l}} \leq x \leq x^{\mathrm{u}} \text{ and } c(x) = 0$$

with $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and $C = \{x \in \mathbb{R}^n \mid x^{\mathrm{l}} \leq x \leq x^{\mathrm{u}}\}$, where some components of x^{u} and x^{l} may take on values of $\pm\infty$.

Another popular example is partial differential equation (PDE) constrained optimization, where $X = U \times Q$ is a product of the state and control space, C encodes pointwise constraints on the controls, and $c(x) = c((u, q)) = 0$ is the PDE constraint, where we often assume that the state $u \in U$ is locally uniquely determined by the control $q \in Q$ as an implicit function $u(q)$ via $c((u(q), q)) = 0$.

1.1 Structure of the article

We give a concise overview of the results of this article in Sect. 1.2. We outline our contributions and connections to existing methods in Sect. 1.3. In the remainder of Sect. 1, we settle our notation, state the general assumptions, and provide the statements of important classical results. We give a short proof of the necessary optimality conditions we use and discuss two central constraint qualifications in Sect. 2. Our main results on projected gradient/antigradient flows for (1) follow in Sect. 3. The application of a projected backward Euler method on the projected gradient/antigradient flow results in a sequential homotopy method, which we describe in Sect. 4. We present numerical results for a local semismooth Newton method globalized by the sequential homotopy approach for a class of highly nonlinear and badly conditioned elliptic PDE-constrained optimal control problems with control constraints in Sect. 5.

1.2 Overview: a novel solution approach based on a sequence of homotopies

We propose the following general solution approach in this paper: We construct and analyze existence and uniqueness of a primal-dual projected gradient/antigradient flow for an augmented Lagrangian. The equilibria of the flow are critical points of (1) and vice versa. Under reasonable assumptions, we prove that critical points that are not local minima cannot be asymptotically stable. Small perturbations will make the flow escape these unwanted critical points. We then apply a projected version of backward Euler timestepping. We provide an interpretation of the backward Euler equations as the optimality conditions of a primal-dual proximally regularized counterpart to (1), which satisfies a strong constraint qualification, even though (1) might only satisfy the Guignard constraint qualification [26], the weakest of all constraint qualifications. This gives rise to a sequential homotopy method, in which a sequence of proximally regularized subproblems needs to be solved by (possibly inexact) fast numerical methods that are only required to converge locally.

We invite the reader to read the supplementary material, in which we sketch without proofs the salient features of our approach with an illustrative example in finite dimensions without inequalities.

1.3 Related work and contributions

We advance and bridge several fields of optimization with this paper.

The field of globalized Newton methods based on differential equation methods applied to the Newton flow started in the early 1950s with Davidenko [20] and continues to raise scientific interest over the decades [9,15,21,23,24,34,42,51,52], predominantly due to the affine invariance properties of the Newton flow [22]. By trading the affine invariance of the Newton flow for the stability properties of the gradient flow, we obtain from a dynamical systems point of view the advantage of being repelled from maxima or saddle points when solving nonlinear optimization problems.

The Newton method, which is equivalent to forward Euler timestepping on the Newton flow with stepsize $\Delta t = 1$, has the prominent property of quadratic local convergence. Backward Euler timestepping on the gradient/antigradient flow can attain superlinear local convergence if the solution is sufficiently regular so that we can take the stepsize Δt to infinity or, equivalently, drive the proximal coefficient λ to zero, provided that we use a local solver in the numerical homotopy method with at least superlinear local convergence. Driving λ to zero is usually possible if the solution satisfies certain second order sufficient optimality conditions.

Three methods in the field of convex optimization are closely related to our approach. The first method is the proximal point algorithm for closed proper convex functions, which can be interpreted as a backward Euler timestepping on the gradient flow of the objective function, while the gradient descent method amounts to forward Euler timestepping on the gradient flow (see, e.g., [48, Sect. 4.1] and references therein). We extend this approach to nonconvex optimization problems with explicit handling of nonlinear equality constraints, as they appear for instance in optimal control. To this end, we extend a second method, the primal-dual projected gradient/antigradient flow of [8, Chap. 6, 7], from the finite-dimensional convex to the infinite-dimensional nonconvex setting with the help of an augmented Lagrangian technique in the framework of projected differential equations in Hilbert space [19]. The third method we extend is the closely related Arrow–Hurwicz gradient method [8, Chap. 10], which amounts to projected forward Euler timestepping on the projected gradient/antigradient flow of the Lagrangian without augmentation ($\rho = 0$). Our sequential homotopy method is equivalent to projected backward Euler timestepping. Hence, it bears the same connection with the Arrow–Hurwicz gradient method as the proximal point algorithm with gradient descent.

From a Sequential Quadratic Programming (SQP) perspective (see, e.g., [46]), our approach resolves all the numerical difficulties on the nonlinear level such as subproblem infeasibility, degeneracy, and nonconvexity due to indefinite subproblem Hessians. Existing approaches often pass these difficulties on to the level of the quadratic subproblem solvers, which may fail to resolve these issues in a way that guarantees convergence of the overall nonlinear iteration. Our method can thus be used as a black-box globalization framework for any locally convergent optimization method that can be used within a continuation framework, e.g., methods of structure-exploiting inexact Sequential Quadratic Programming (SQP) [28,33,49,50,53] or semismooth Newton methods [31–33,35,43,54,56,57]. The local methods are even allowed to converge to maxima and saddle points. These issues are taken care of by our sequential homotopy method. For the application of local SQP methods, we can guarantee that the quadratic subproblems are always feasible and that they satisfy a strong constraint qualification that implies unique subproblem Lagrange multipliers. In addition, they are convex if the augmentation parameter ρ is sufficiently large and the stepsize Δt is sufficiently small when we are still far away from a solution.

Our approach uses the theory of projected differential equations due to Cojocaru and Jonker [19], which have a tight connection to differential inclusions [10] and evolutionary/differential variational inequalities [18,47]. We are mainly interested in their equilibrium points, which satisfy a variational inequality (VI). Other methods to compute solutions to VIs have been described in the literature (see, e.g., [13,45]), which are based on semismooth iterations on reformulations using special Nonlinear Complementarity Problem (NCP) functions.

Projected gradient flows for constrained optimization problems in finite dimensions have also been considered with techniques from Riemannian geometry (see, e.g., [29, 30,37,55] and references therein), but the resulting methods produce only feasible iterates. It is often computationally wasteful to satisfy all constraints for iterates far away from an optimum and to force the iterates to follow a feasible manifold with possibly high curvature.

For an introduction to augmented Lagrangian approaches in Hilbert spaces we refer to [36] and references therein. We point out that our approach relies on the augmented Lagrangian mainly to remove negative curvature of the Lagrangian in the kernel of the constraints. In contrast to classical augmented Lagrangian methods, we do not alternate between updates of the primal and dual variables but rather update primal and dual variables simultaneously as in augmented Lagrangian-SQP methods [36, Chap. 6].

1.4 Notation

We abbreviate the nonnegative real numbers with $\mathbb{R}_{\geq 0}$. By $(x_k) \subset X$ we denote a sequence $x_0, x_1, \ldots$ of elements in X. By X^* we denote the topological dual of X, by $(.,.)_X : X \times X \to \mathbb{R}$ the inner product, by $\|.\|_X : X \to \mathbb{R}_{\geq 0}$ the norm, and by $\langle ., . \rangle_{X^*,X} : X^* \times X \to \mathbb{R}$ the duality pairing. By $R_X : X^* \to X$ we denote the Riesz isomorphism (see, e.g., [58, Sect. III.6]), which satisfies the identity

$$\left(R_X x^*, x\right)_X = \left\langle x^*, x \right\rangle_{X^*,X} \quad \text{for all } x^* \in X^*, x \in X$$

and likewise for Y. As usual, $\mathcal{L}(X, Y)$ denotes the Banach-space of all continuous linear operators from X to Y. For $A \in \mathcal{L}(X, Y)$, the (Banach space) dual operator $A^* \in \mathcal{L}(Y^*, X^*)$ and the (Hilbert space) adjoint operator $A^\star \in \mathcal{L}(Y, X)$ are defined by

$$\begin{aligned} \left\langle A^* y^*, x \right\rangle_{X^*,X} &= \left\langle y^*, Ax \right\rangle_{Y^*,Y} && \text{for all } x \in X, y^* \in Y^*, \\ \left(A^\star y, x\right)_X &= (y, Ax)_Y && \text{for all } x \in X, y \in Y, \end{aligned}$$

which implies $A^\star R_Y = R_X A^*$. We denote the Fréchet-derivative of $c(x)$ with $c'(x) \in \mathcal{L}(X, Y)$. We denote the objective gradient by $\nabla\phi(x) = R_X \phi'(x) \in X$ and the adjoint of the constraint derivative by $\nabla c(x) = \left(c'(x)\right)^\star \in \mathcal{L}(Y, X)$. For a linear operator $A \in \mathcal{L}(X, Y)$, we denote its kernel by $\ker(A) = \{x \in X \mid Ax = 0\}$ and its range by $\operatorname{ran}(A) = \{y \in Y \mid \exists x \in X : y = Ax\}$. For an open set $\Omega \in \mathbb{R}^n$, we denote with $L^2(\Omega)$ the standard Hilbert space of square Lebesgue-integrable functions on Ω, with $H_0^1(\Omega)$ the Sobolev-space of functions with square Lebesgue-integrable derivatives and zero trace at the boundary, and with $H^{-1}(\Omega)$ its dual space. We denote the feasible set of (1) with $\mathcal{F} = \{x \in C \mid c(x) = 0\}$.

1.5 General assumptions

A central role in this article is played by the augmented objective and augmented Lagrangian

$$\phi^\rho(x) = \phi(x) + \frac{\rho}{2}\|c(x)\|_Y^2, \qquad L^\rho(x, y) = \phi^\rho(x) + (y, c(x))_Y, \tag{2}$$

defined for some fixed $\rho \in \mathbb{R}_{\geq 0}$ and arbitrary $x \in C$ and $y \in Y$. Throughout this article, we make the following assumptions:

Assumption 1 *For all $x \in \mathcal{F}$, $\operatorname{ran}(c'(x))$ is closed in Y.*

Assumption 2 *For some fixed $\rho \in \mathbb{R}_{\geq 0}$ we have the coercivity condition*

$$\phi_{\text{low}}^\rho = \inf_{x \in C} \phi^\rho(x) > -\infty \quad \textit{and} \quad \lim_{\|x\|_X \to \infty} \phi^\rho(x) = \infty.$$

Assumption 3 *The functions $c(x)$, $L^\rho(x, y)$ and the gradient $\nabla L^\rho(x, y)$ are locally Lipschitz continuous.*

1.6 Well-known results

Let us recall the following well-known definitions.

Definition 1 (*Tangent cone*) For $\bar{x} \in X$ and a nonempty set $M \subseteq X$, we call

$$T(M, \bar{x}) = \{d \in X \mid \text{ there exist sequences } (x_k) \subset M, (\lambda_k) \subset \mathbb{R}_{\geq 0} \\ \text{with } x_k \to \bar{x} \text{ and } \lambda_k(x_k - \bar{x}) \to d \text{ as } k \to \infty\}$$

the *tangent cone* to M at $\bar{x}$.

Definition 2 (*Projection*) For a nonempty closed convex set $K \subseteq X$, we denote by $P_K : X \to K$ the *projection operator* of X onto K, which is uniquely defined by

$$\|P_K(x) - x\|_X = \inf_{\tilde{x} \in K} \|\tilde{x} - x\|_X \quad \text{for all } x \in X.$$

For properties of projection operators, we refer the reader to [59].

Definition 3 (*Polar cone*) For a cone $K \subseteq X$, we call

$$K^- = \{d \in X \mid (d, x)_X \leq 0 \text{ for all } x \in K\}$$

the *polar cone* of K.

Remark 1 If $K \subseteq X$ is a linear subspace, then $x \in K$ implies $-x \in K$ and thus equality holds in the definition of $K^- = \{d \in X \mid (d, x)_X = 0 \text{ for all } x \in K\} = K^\perp$.

We shall make use of the following classical results from convex analysis.

Lemma 1 (Moreau decomposition) *If $K \subseteq X$ is a nonempty closed convex cone, then every $x \in X$ has a unique decomposition $x = P_K(x) + P_{K^-}(x) =: x^+ + x^-$, where $\left(x^-, x^+\right)_X = 0$. A simple consequence is the identity*

$$(x, P_K(x))_X = \left(x^+ + x^-, x^+\right)_X = \|P_K(x)\|_X^2 .$$

Proof See [44] according to [59, Lemma 2.2 and Corollary 2]. □

Lemma 2 *Let $K \subseteq X$ be a nonempty closed convex set and let $\bar{x} \in K$. If $x \in T^-(K, \bar{x}) + \bar{x}$, then $P_K(x) = \bar{x}$.*

Proof Choose any $y \in K$. Then, $y - \bar{x} \in T(K, \bar{x})$, e.g., with $\lambda_k = k + 1$ and $x_k = (1 - \lambda_k^{-1})\bar{x} + \lambda_k^{-1} y \in K$. Because $x - \bar{x} \in T^-(K, \bar{x})$, we obtain $(x - \bar{x}, y - \bar{x})_X \leq 0$. The result follows from [59, Lemma 1.1], because $y \in K$ was chosen arbitrarily. □

2 Necessary optimality conditions

The basis for the sequential homotopy method we propose in Sect. 4 is a necessary optimality condition due to Guignard [26]. Because the separation of nonlinearities $c(x) = 0$ and inequalities $x \in C$ in (1) allow for a much shorter proof, we state it here for the sake of convenience.

Lemma 3 *If $\bar{x} \in \mathcal{F}$ is a local optimum of* (1)*, then $-\nabla\phi(\bar{x}) \in T^{-}(\mathcal{F}, \bar{x})$.*

Proof Let $d \in T(\mathcal{F}, \bar{x})$ with corresponding sequences $(x_k) \subset X$ and $(\lambda_k) \subset \mathbb{R}_{\geq 0}$. Using the shorthand $d_k = \lambda_k(x_k - \bar{x})$, we obtain the assertion from letting $k \to \infty$ in

$$0 \leq \lambda_k \left[\phi(x_k) - \phi(\bar{x})\right] = \left\langle \phi'(\bar{x}), d_k \right\rangle_{X^*, X} + \|d_k\|_X \frac{o\left(\|x_k - \bar{x}\|_X\right)}{\|x_k - \bar{x}\|_X} \to \left(\nabla\phi(\bar{x}), d\right)_X .$$

□

Definition 4 (*GCQ*) We say that the *Guignard Constraint Qualification (GCQ)* holds at $\bar{x} \in \mathcal{F}$ if

$$\ker^{\perp}(c'(\bar{x})) + T^{-}(C, \bar{x}) = T^{-}(\mathcal{F}, \bar{x}).$$

Theorem 1 (Necessary optimality conditions) *If $\bar{x} \in \mathcal{F}$ is a local optimum of* (1) *that satisfies GCQ, then there exists a multiplier $\bar{y} \in Y$ such that*

$$-\nabla\phi(\bar{x}) - \nabla c(\bar{x})\bar{y} \in T^{-}(C, \bar{x}). \tag{3}$$

Proof The proof is based on the Closed Range Theorem (see, e.g., [58, Sect. VII.5] with premultiplication by the Riesz isomorphism R_X to obtain the Hilbert space version), which states that Assumption 1 is equivalent to

$$\ker^{\perp}(c'(\bar{x})) = \operatorname{ran}(\nabla c(\bar{x})).$$

Together with Lemma 3 and GCQ we obtain

$$-\nabla\phi(\bar{x}) \in T^{-}(\mathcal{F}, \bar{x}) = \ker^{\perp}(c'(\bar{x})) + T^{-}(C, \bar{x}) = \operatorname{ran}(\nabla c(\bar{x})) + T^{-}(C, \bar{x}).$$

Thus, there exists a $\bar{y} \in Y$ such that $-\nabla\phi(\bar{x}) - \nabla c(\bar{x})\bar{y} \in T^{-}(C, \bar{x})$. □

Definition 5 (*Critical point*) We call $(\bar{x}, \bar{y}) \in \mathcal{F} \times Y$ a *critical point* if (3) holds.

The method we propose below enjoys the benefit that its subproblems lift the original problem into a larger space with additional structural properties in X, C, and c, which result in satisfaction of a constraint qualification that is much stronger than GCQ, even though problem (1) only satisfies GCQ.

Lemma 4 *Let $X = U \times Q$, equipped with the canonical inner product derived from the Hilbert spaces U and Q, and let $C = U \times C_Q$ for some nonempty closed convex set $C_Q \subseteq Q$. Furthermore, assume there exists a continuously Fréchet-differentiable mapping $S : C_Q \to U$ such that for all $x = (u, q) \in C$*

$$\text{(a)}\;\; c((u,q)) = 0 \;\; \textit{iff} \;\; u = S(q), \quad \text{(b)}\;\; \operatorname{ran} c'_u(x) = Y, \quad \text{(c)}\;\; \operatorname{ran} \nabla_u c(x) = U.$$

Then, $\mathcal{F}$ is nonempty, every $\bar{x} \in \mathcal{F}$ satisfies GCQ, and the Lagrange multiplier $\bar{y}$ in (3) is uniquely determined.

Proof The feasible set $\mathcal{F} = \{(S(q), q) \mid q \in C_Q\}$ is nonempty because C_Q is nonempty. Let $\bar{x} = (S(\bar{q}), \bar{q}) \in \mathcal{F}$ and choose some $d \in T(C_Q, \bar{q})$. By definition, there exist sequences $(q_k) \subset C_Q$ and $(\lambda_k) \subset \mathbb{R}_{\geq 0}$ such that $\lambda_k(q_k - \bar{q}) \to d$. Using (a), we choose a sequence $(x_k) \subset \mathcal{F}$ according to $x_k = (S(q_k), q_k)$ to guarantee $x_k \to \bar{x}$ and

$$\begin{aligned} \lambda_k(x_k - \bar{x}) &= \lambda_k(S(q_k) - S(\bar{q}), q_k - \bar{q}) \\ &= \lambda_k(S'(\bar{q})(q_k - \bar{q}) + o(\|q_k - \bar{q}\|_Q), q_k - \bar{q}) \to (S'(\bar{q})d, d), \end{aligned} \tag{4}$$

which shows that $T(\mathcal{F}, \bar{x}) \supseteq \{(S'(\bar{q})d, d) \mid d \in T(C_Q, \bar{q})\}$. In order to show that equality holds between the two sets, we notice that if $(e, d) \in T(\mathcal{F}, \bar{x})$ then $d \in T(C_Q, \bar{q})$ and (4) implies $e = S'(\bar{q})d$. Hence, we obtain

$$T(\mathcal{F}, \bar{x}) = \{(S'(\bar{q})d, d) \mid d \in T(C_Q, \bar{q})\}.$$

In order to compute its polar cone, let $x = (u, \tilde{q}) \in X$ such that

$$0 \geq \big(u, S'(\bar{q})d\big)_U + (\tilde{q}, d)_Q = (\nabla S(\bar{q})u + \tilde{q}, d)_Q \quad \text{for all } d \in T(C_Q, \bar{q}).$$

We choose $q = \nabla S(\bar{q})u + \tilde{q}$ in order to obtain

$$\begin{aligned} T^-(\mathcal{F}, \bar{x}) &= \big\{(u, \tilde{q}) \in X \mid (u, e)_U + (\tilde{q}, d)_Q \leq 0 \text{ for all } (e, d) \in T(\mathcal{F}, \bar{x})\big\} \\ &= \big\{(u, \tilde{q}) \in X \mid (\nabla S(\bar{q})u + \tilde{q}, d)_Q \leq 0 \text{ for all } d \in T(C_Q, \bar{q})\big\} \\ &= \big\{(u, q - \nabla S(\bar{q})u) \mid u \in U, q \in T^-(C_Q, \bar{q})\big\}. \end{aligned} \tag{5}$$

For the other polar cone in the definition of GCQ, we get

$$T^-(C, \bar{x}) = T^-(U \times C_Q, (\bar{u}, \bar{q})) = \big(U \times T(C_Q, \bar{q})\big)^- = \{0\} \times T^-(C_Q, \bar{q}). \tag{6}$$

Taking the derivative of $c(S(q), q) = 0$ with respect to q in direction $d \in Q$ yields

$$c'_u(\bar{x})S'(\bar{q})d + c'_q(\bar{x})d = 0.$$

As a consequence of the Closed Range Theorem [58, Sect. VII.5, Corollary 1], (c) is equivalent to the existence of a continuous inverse of $c_u'(\bar{x})$, from which we see that

$$\ker c'(\bar{x}) = \left\{(e, d) \in X \mid c_u'(\bar{x})e + c_q'(\bar{x})d = 0\right\} = \left\{(S'(\bar{q})d, d) \mid d \in Q\right\}.$$

Thus, its orthogonal complement amounts to

$$\begin{aligned}\ker^{\perp} c'(\bar{x}) &= \left\{(u, q) \in X \mid \left(u, S'(\bar{q})d\right)_U + (q, d)_Q = 0 \text{ for all } d \in Q\right\}\\ &= \left\{(u, q) \in X \mid (\nabla S(\bar{q})u + q, d)_Q = 0 \text{ for all } d \in Q\right\}\\ &= \{(u, -\nabla S(\bar{q})u) \mid u \in U\}.\end{aligned} \tag{7}$$

Hence, it follows from (6), (7), and (5) that

$$T^{-}(C, \bar{x}) + \ker^{\perp} c'(\bar{x}) = \left\{(u, q - \nabla S(\bar{q})u \mid u \in U, q \in T^{-}(C_Q, \bar{q})\right\} = T^{-}(\mathcal{F}, \bar{x}),$$

which shows that GCQ holds at $\bar{x}$. Regarding multiplier uniqueness, we take the U-components of (3) and (6) to deduce

$$\nabla_u \phi(\bar{x}) + \nabla_u c(\bar{x})\bar{y} = 0,$$

from which the uniqueness of $\bar{y}$ follows from the the existence of a continuous inverse of $\nabla_u c(\bar{x})$ by virtue of (b) and [58, Sect. VII.5, Corollary 1]. □

3 Projected gradient/antigradient flow

We study a primal-dual gradient/anti-gradient flow (from now on simply called *gradient flow*) of the augmented Lagrangian L^{ρ}, defined in (2), projected on the closed convex set C in the framework of projected differential equations in Hilbert space [19] according to

$$\dot{x}(t) = P_{T(C, x(t))}\left(-\nabla_x L^{\rho}(x(t), y(t))\right), \qquad \dot{y}(t) = \nabla_y L^{\rho}(x(t), y(t)), \tag{8}$$

where the gradients with respect to x and y evaluate to

$$\nabla_x L^{\rho}(x, y) = \nabla \phi(x) + \nabla c(x)\left[y + \rho c(x)\right], \qquad \nabla_y L^{\rho}(x, y) = c(x).$$

The following existence theorem uses L^{ρ} and $\frac{1}{2}\|c(.)\|_Y^2$ as Lyapunov-type functions. Due to Lemma 1, the t-derivative of L^{ρ} along the flow is given by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} L^{\rho}(x(t), y(t)) &= \left(\nabla_x L^{\rho}(x(t), y(t)), \dot{x}(t)\right)_X + \left(\nabla_y L^{\rho}(x(t), y(t)), \dot{y}(t)\right)_Y\\ &= -\left\| P_{T(C, x(t))}\left(-\nabla_x L^{\rho}(x(t), y(t))\right)\right\|_X^2 + \|c(x(t))\|_Y^2.\end{aligned} \tag{9}$$

The positive sign in front of the last term in (9) reflects the saddle point nature of the Lagrangian approach and complicates the use of Lyapunov arguments in comparison to the unconstrained case. We pursue the basic idea that by increasing ρ, we can make the negative term overpower the ρ-independent positive term. That this is not always possible will be discussed after the following theorem.

Theorem 2 (Unique existence of solutions) *Let Assumptions 2 and 3 be satisfied. Then, there exists an interval* $[0, t_{\text{final}}]$ *and a uniquely determined pair of absolutely continuous functions* $(x, y) : [0, t_{\text{final}}] \to C \times Y$ *that satisfy the projected gradient flow Eq.* (8) *and* $(x(0), y(0)) = (x_0, y_0)$*. The final time* t_{final} *can be extended as long as the condition*

$$\frac{\mathrm{d}}{\mathrm{d}t} L^{\rho}(x(t), y(t)) \leq 0 \tag{10}$$

holds almost everywhere on $[0, t_{\text{final}}]$*. In addition, if for some* $\gamma_1, \gamma_2 \in (0, 1)$ *the conditions* (10) *and*

$$\gamma_1 \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\|c(x(t))\|_Y^2\right) \leq -\frac{\mathrm{d}}{\mathrm{d}t} L^{\rho}(x(t), y(t)) - \gamma_2 \|c(x(t))\|_Y^2 , \tag{11}$$

hold almost everywhere in $\mathbb{R}_{\geq 0}$*, we have*

$$\int_0^{\infty} \left\| P_{T(C,x(t))}\left(-\nabla_x L^{\rho}(x(t), y(t))\right)\right\|_X^2 \mathrm{d}t < \infty \quad \textit{and} \quad \int_0^{\infty} \|c(x(t))\|_Y^2 \,\mathrm{d}t < \infty. \tag{12}$$

Furthermore, if there is a set $M \subseteq X \times Y$ *such that* ∇L^{ρ} *is (globally) Lipschitz continuous on* M *and* $(x(t), y(t)) \in M$ *for all* $t \in [0, \infty)$*, we obtain*

$$P_{T(C,x(t))}\left(-\nabla_x L^{\rho}(x(t), y(t))\right) \to 0 \quad \textit{and} \quad c(x(t)) \to 0 \quad \textit{for } t \to \infty.$$

Proof By Assumption 3, $\nabla L^{\rho}(x, y)$ is Lipschitz continuous in a neighborhood of (x_0, y_0) with some Lipschitz constant $b < \infty$. By virtue of [19, Theorem 3.1], there exists an $l > 0$ and a uniquely determined pair of absolutely continuous functions $(x, y) : [0, l] \to C \times Y$ that satisfy (8) for almost all $t \in [0, l]$ and $x(0) = x_0$, $y(0) = y_0$. Without loss of generality, (10) is satisfied on $[0, l]$ and we can repeatedly extend the local solution by the above arguments until (10) or (11) is violated for some $t_{\text{final}} > 0$. As long as (10) is satisfied, no blowup is possible in finite time. To see this, we first observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\|y(t)\|_Y^2\right) = (y(t), c(x(t)))_Y , \tag{13}$$

which implies in combination with (10) and Assumption 2 that

$$\begin{aligned}\frac{1}{2}\left\|y(t)\right\|_Y^2 &= \frac{1}{2}\left\|y_0\right\|_Y^2 + \int_0^t (y(\tau), c(x(\tau)))_Y \,\mathrm{d}\tau \\ &= \frac{1}{2}\left\|y_0\right\|_Y^2 + \int_0^t \left[L^\rho(x(\tau), y(\tau)) - \phi^\rho(x(\tau))\right] \mathrm{d}\tau \\ &\leq \frac{1}{2}\left\|y_0\right\|_Y^2 + t\left[L^\rho(x_0, y_0) - \phi^\rho_{\mathrm{low}}\right].\end{aligned}$$

This establishes that there can be no blowup of y in finite time. In addition, x cannot blow up in finite time because then $L^\rho(x(t), y(t))$ would tend to infinity by virtue of Assumption 2.

Hence, we can extend the local solutions to global solutions on the whole interval $\mathbb{R}_{\geq 0}$ if the condition (10) holds almost everywhere. In this case, Eqs. (13), (2), and Assumption 2 imply that for $t > 0$

$$\begin{aligned}\frac{1}{t}\int_0^t L^\rho(x(\tau), y(\tau))\,\mathrm{d}\tau &= \frac{1}{t}\int_0^t \phi^\rho(x(\tau))\,\mathrm{d}\tau + \frac{1}{t}\left[\frac{1}{2}\left\|y(t)\right\|_Y^2 - \frac{1}{2}\left\|y_0\right\|_Y^2\right] \\ &\geq \phi^\rho_{\mathrm{low}} - \frac{1}{2t}\left\|y_0\right\|_Y^2. \end{aligned} \tag{14}$$

Using the monotonicity $L^\rho(x(\tau), y(\tau)) \leq L^\rho(x(s), y(s))$ for $0 < s \leq \tau$ implied by (10), we obtain for $s \leq t$ that

$$\frac{1}{t}\int_0^t L^\rho(x(\tau), y(\tau))\,\mathrm{d}\tau \leq \frac{1}{t}\int_0^s L^\rho(x(\tau), y(\tau))\,\mathrm{d}\tau + \frac{t-s}{t} L^\rho\left(x(s), y(s)\right). \tag{15}$$

We concatenate (14) and (15) and let $t \to \infty$, which yields

$$L^\rho(x(s), y(s)) \geq \phi^\rho_{\mathrm{low}} \quad \text{for all } s \in \mathbb{R}_{\geq 0}.$$

Hence, we obtain

$$0 \geq \int_0^t \frac{\mathrm{d}}{\mathrm{d}\tau} L^\rho(x(\tau), y(\tau))\,\mathrm{d}\tau = L^\rho(x(t), y(t)) - L^\rho(x_0, y_0) \geq \phi^\rho_{\mathrm{low}} - L^\rho(x_0, y_0). \tag{16}$$

If condition (11) holds additionally, the boundedness of the integral in (16) implies with integration of assumption (11) that

$$\begin{aligned}&\gamma_2 \int_0^t \left\|c(x(\tau))\right\|_Y^2 \,\mathrm{d}\tau \\ &\quad\leq -\int_0^t \frac{\mathrm{d}}{\mathrm{d}t} L^\rho(x(t), y(t))\,\mathrm{d}\tau - \gamma_1 \int_0^t \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\left\|c(x(t))\right\|_Y^2\right)\mathrm{d}\tau\end{aligned}$$

$$
\begin{aligned}
&\le L^\rho(x_0, y_0) - L^\rho(x(t), y(t)) - \gamma_1 \left[\frac{1}{2} \|c(x(t))\|_Y^2 - \frac{1}{2} \|c(x_0)\|_Y^2 \right] \\
&\le L^\rho(x_0, y_0) - \phi_{\text{low}}^\rho + \gamma_1 \frac{1}{2} \|c(x_0)\|_Y^2 . \qquad (17)
\end{aligned}
$$

Hence, $\int_0^\infty \|c(x(t))\|_Y^2 \, \mathrm{d}t < \infty$ and we can establish (12) by way of (16) and the representation (9).

If now there is a set $M \subseteq X \times Y$ such that ∇L^ρ is Lipschitz continuous on M and $(x(t), y(t)) \in M$ for all $t \in [0, \infty)$, then the integrand in (17) is absolutely continuous (as a concatenation of an absolutely continuous function with Lipschitz continuous functions). This implies uniform continuity of the integrand and we can deduce that $\|c(x(t))\|_Y^2 \to 0$ for $t \to \infty$. In combination with (16) and the representation (9), this implies that

$$
\frac{\mathrm{d}}{\mathrm{d}t} L^\rho(x(t), y(t)) = - \left\| P_{T(C,x(t))} \left(-\nabla_x L^\rho(x(t), y(t)) \right) \right\|_X^2 + \|c(x(t))\|_Y^2 \to 0
$$

and finally $P_{T(C,x(t))} \left(-\nabla_x L^\rho(x(t), y(t)) \right) \to 0$ for $t \to \infty$. □

Discussion of Theorem 2 If we do not obtain a solution up to $t_{\text{final}} = \infty$, it must be due to violation of (10) or (11). In this case, we may try to increase ρ in order for the negative term in (9) to overpower the positive one. To understand the behavior for $\rho \to \infty$, we let $\beta = 1/(1 + \rho) \in [0, 1]$ and consider a reparametrization of the flow Eq. (8) via $x_\beta(t) = x(\beta t)$, $y_\beta(t) = y(\beta t)$, which leads to

$$
\begin{aligned}
\dot{x}_\beta(t) &= P_{T(C,x_\beta(t))} \Big(-\beta \nabla_x L^0(x_\beta(t), y_\beta(t)) - (1 - \beta) \nabla c(x_\beta(t)) c(x_\beta(t)) \Big), \\
\dot{y}_\beta(t) &= \beta \nabla_y L^\rho(x_\beta(t), y_\beta(t)).
\end{aligned}
$$

For $\beta = 0$, these flow equations reduce to the projected gradient flow for minimizing the constraint violation $\|c(x)\|_Y^2$ over $x \in C$ according to

$$
\dot{x}_\beta(t) = P_{T(C,x_\beta(t))} \left(-\nabla c(x_\beta(t)) c(x_\beta(t)) \right), \qquad \dot{y}_\beta(t) = 0.
$$

Hence, violation of (10) or (11) for large ρ can only occur if for $\beta = 1$ we get stuck in a locally infeasible point $\tilde{x}$ of problem (1), which means

$$
P_{T(C,\tilde{x})} \left(-[\nabla c(\tilde{x})] c(\tilde{x}) \right) = 0 \quad \text{but} \quad c(\tilde{x}) \neq 0.
$$

This case must arise for instance if $\mathcal{F} = \varnothing$ and it is reassuring that the theory provides room for this pathological case and that we at least obtain a point of (locally) minimal constraint violation.

We also remark that boundedness of $y(t)$ can for instance be ensured by the sufficient condition that for some $\gamma_3 > 0$ we have (omitting t-arguments)

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\,\|c(x)\|_Y^2\right) = \Big(P_{T(C,x)}\left(-\nabla_x L^0(x,y) - \rho\nabla c(x)c(x)\right), \nabla c(x)c(x)\Big)_X \\ \leq -\gamma_3\,\|c(x)\|_Y^2\,. \tag{18}$$

In this case, Grönwall's inequality (see, e.g., [5]) implies $\|c(x(t))\|_Y \leq \|c(x_0)\|_Y\, e^{-\gamma_3 t}$ and consequently

$$\|y(t) - y_0\|_Y \leq \int_0^\infty \|c(x(t))\|_Y\,\mathrm{d}t \leq \gamma_3^{-1}\,\|c(x_0)\|_Y\,.$$

Assumption (18) is obviously too restrictive for the case of a feasible initial guess $c(x_0) = 0$, which would imply $y(t) \equiv y_0$. Hence, we prefer the weaker assumption (11) in Theorem 2.

We next characterize equilibrium points of (8) assuming they exist.

Lemma 5 (Equilibria are critical) *Equilibrium points* $(\bar{x}, \bar{y}) \in C \times Y$ *of* (8) *are critical points of* (1) *and vice versa.*

Proof Let $(\bar{x}, \bar{y}) \in C \times Y$ be an equilibrium point of (8), implying $0 = \nabla_y L^\rho(\bar{x}, \bar{y}) = c(\bar{x})$ and consequently $\bar{x} \in \mathcal{F}$. From $0 = P_{T(C,\bar{x})}(-\nabla_x L^\rho(\bar{x}, \bar{y}))$, we can derive with Lemma 1 that $-\nabla_x L^\rho(\bar{x}, \bar{y}) = P_{T^-(C,\bar{x})}(-\nabla_x L^\rho(\bar{x}, \bar{y})) \in T^-(C, \bar{x})$. Because $c(\bar{x}) = 0$, we have $\nabla_x L^\rho(\bar{x}, \bar{y}) = \nabla\phi(\bar{x}) + \nabla c(\bar{x})\bar{y}$. Hence, $(\bar{x}, \bar{y})$ is a critical point.

Let now $(\bar{x}, \bar{y}) \in \mathcal{F} \times Y$ be a critical point of (1). Because $\bar{x} \in \mathcal{F}$, the antigradient vanishes due to $\nabla_y L^\rho(\bar{x}, \bar{y}) = c(\bar{x}) = 0$. By definition, we also have that

$$-\nabla_x L^\rho(\bar{x}, \bar{y}) = -\nabla\phi(\bar{x}) - \nabla c(\bar{x})\,[\bar{y} + \rho c(\bar{x})] = -\nabla\phi(\bar{x}) - \nabla c(\bar{x})\bar{y} \in T^-(C, \bar{x}).$$

Moreau decomposition of $-\nabla_x L^\rho(\bar{x}, \bar{y})$ then yields that $P_{T(C,\bar{x})}(-\nabla_x L^\rho(\bar{x}, \bar{y})) = 0$. This shows that both right-hand sides of (8) vanish and that $(\bar{x}, \bar{y})$ is an equilibrium point. □

Among the critical points we are apparently only interested in those that are minima of (1). For the finite-dimensional unconstrained case, we recall that asymptotically stable equilibria of the gradient flow are strict local minima of the objective function and that the converse is true if the objective is analytic in a neighborhood of the minimum [1]. This is of high practical relevance, because the gradient flow will be attracted to strict local minima and, conversely, small perturbations (for instance due to numerical round-off) will usually make the flow escape unwanted critical points such as saddle points or maxima.

For the constrained case, the situation is more complicated because the intrinsic saddle point structure of the Lagrangian requires a gradient/antigradient flow, for which to our knowledge no results on asymptotic stability exist so far. We show that critical points that admit an emanating feasible curve of descent are not asymptotically stable (under reasonable conditions). This implies that the projected gradient/antigradient flow will not be attracted to these undesired critical points. To prove this result, we need the following three definitions.

Definition 6 (*Descent curve*) We call a continuous function $\bar{x} : [0, 1] \to \mathcal{F}$ a *descent curve* of (1), if $\phi(\bar{x}(t_2)) < \phi(\bar{x}(t_1))$ for all $0 \leq t_1 < t_2 \leq 1$.

Definition 7 (*Stability*) An equilibrium $(\bar{x}, \bar{y}) \in C \times Y$ of the projected gradient flow (8) is *stable* if for every neighborhood $U \times V \subset X \times Y$ of $(\bar{x}, \bar{y})$ there exists a smaller neighborhood $U_1 \times V_1$ of $(\bar{x}, \bar{y})$ such that solutions $(x, y) : [0, \infty) \to (U \cap C) \times V$ of (8) exist for all initial values $(x_0, y_0) \in (U_1 \cap C) \times V_1$. If, in addition, it holds for all these solutions that $\lim_{t\to\infty}(x(t), y(t)) = (\bar{x}, \bar{y})$, then $(\bar{x}, \bar{y})$ is *asymptotically stable*.

Definition 8 (*Flow ribbon*) For a continuous function $(\bar{x}, \bar{y}) : [0, 1] \to C \times Y$ we denote by $\mathcal{R}(\bar{x}, \bar{y}) \subseteq C \times Y$ the *flow ribbon emanating from the curve* $(\bar{x}, \bar{y})$, which we define as the union of the images of all curves $(x, y) : \mathbb{R}_{\geq 0} \to C \times Y$ satisfying (8) with initial values $(x(0), y(0)) = (\bar{x}(l), \bar{y}(l))$ for some $l \in [0, 1]$.

We can think of a flow ribbon as the trajectory of a curve under the gradient/antigradient flow (8), just as if the curve at $t = 0$ is the first thread and we weave together the threads into a fabric while moving along the flow. This somewhat unusual definition is required to keep the set of points (x, y) small on which assumption (19) in the following theorem must hold (compare also Example 1 below).

Theorem 3 *Let $\bar{x} : [0, 1] \to \mathcal{F}$ be a descent curve and $\bar{y}(t) \equiv \bar{y} \in Y$ such that $(\bar{x}(0), \bar{y})$ is a critical point of* (1) *and let there exist a neighborhood $U \times V \subset X \times Y$ of $(\bar{x}(0), \bar{y})$ such that for all $(x, y) \in \mathcal{R}(\bar{x}, \bar{y}) \cap (U \times V)$ with $L^\rho(x, y) < L^\rho(\bar{x}(0), \bar{y})$ it holds that*

$$\|c(x)\|_Y^2 \leq \left\| P_{T(C,x)} \left(-\nabla_x L^\rho(x, y)\right) \right\|_X^2 . \tag{19}$$

Then $(\bar{x}(0), \bar{y})$ is not asymptotically stable.

Proof by contradiction Assume $(\bar{x}(0), \bar{y})$ is asymptotically stable. By Definition 7, there exists a neighborhood $U_1 \times V_1 \subset U \times V$ of $(\bar{x}(0), \bar{y})$, which admits for each element as initial value a global solution to (8). We choose $l \in (0, 1]$ such that $(x_0, y_0) := (\bar{x}(l), \bar{y}) \in U_1 \times V_1$. Because $\bar{x}$ is a descent curve, we have that $c(x_0) = 0$ and

$$L^\rho(\bar{x}(0), \bar{y}) - L^\rho(x_0, y_0) = \phi(\bar{x}(0)) - \phi(x_0) =: \varepsilon > 0. \tag{20}$$

By Definition 7, a solution $(x, y) : [0, \infty) \to (U \cap C) \times V$ of (8) with $x(0) = x_0$ and $y(0) = y_0$ exists and converges to $(\bar{x}(0), \bar{y})$. Using assumption (19) and Eqs. (9) and (20), we observe that

$$\begin{aligned} L^\rho(x(t), y(t)) &= L^\rho(x_0, y_0) + \int_0^t \frac{\mathrm{d}}{\mathrm{d}t} L^\rho(x(\tau), y(\tau))\, \mathrm{d}\tau \\ &\leq L^\rho(x_0, y_0) = L^\rho(\bar{x}(0), \bar{y}) - \varepsilon \end{aligned}$$

for all $t \in \mathbb{R}_{\geq 0}$, which implies that (x, y) cannot converge to $(\bar{x}(0), \bar{y})$. Hence, $(\bar{x}(0), \bar{y})$ is not asymptotically stable. □

In order to validate that assumption (19) does not reduce the assertion of Theorem 3 to one about the empty set, we provide a simple example.

Example 1 *(Simple nonconvex quadratic program)* We consider the problem

$$\min \tfrac{1}{2}\left(x_1^2 - x_2^2\right) \quad \text{over } x \in \mathbb{R} \times \mathbb{R}_{\geq 0} =: C \quad \text{subject to } x_1 = 0.$$

It is easy to verify that $(x_1, x_2, y) = (0, 0, 0)$ is a critical point and the objective is unbounded for the feasible points $x_1 = 0$, $x_2 \to \infty$. The augmented Lagrangian amounts to

$$L^\rho(x, y) = \tfrac{1}{2}(x_1^2 - x_2^2) + x_1 y + \tfrac{\rho}{2}x_1^2 = (1 + \rho)\tfrac{1}{2}x_1^2 - \tfrac{1}{2}x_2^2 + x_1 y.$$

The projected gradient flow equations then read (omitting (t)-arguments)

$$\dot{x} = P_{T(C,x)}(-\nabla_x L^\rho(x, y)) = \begin{pmatrix} -(1 + \rho)x_1 - y \\ \max(0, x_2) \end{pmatrix}, \qquad \dot{y} = \nabla_y L^\rho(x, y) = x_1.$$

For the descent curve $\bar{x}(l) = (0, l, 0)^T$, $l \in [0, 1]$, with corresponding $\bar{y} \equiv 0$, we can easily solve the flow equations and obtain the flow ribbon

$$\mathcal{R}(\bar{x}, \bar{y}) = \left\{(0, le^t, 0)^T \mid t \in \mathbb{R}, l \in [0, 1]\right\} = \{0\} \times \mathbb{R}_{\geq 0} \times \{0\}.$$

Hence, we see that for $(x, y) \in \mathcal{R}(\bar{x}, \bar{y})$ it holds that

$$L^\rho(x, y) = -\tfrac{1}{2}x_2^2 \leq 0 = L^\rho(\bar{x}(0), \bar{y})$$

with strict inequality for $x_2 > 0$. Inequality (19) holds for all $(x, y) \in \mathcal{R}(\bar{x}, \bar{y})$ by virtue of

$$\|c(x)\|_2^2 = x_1^2 = 0 \leq \max(0, x_2)^2 = \left\| P_{T(C,x)}(-\nabla_x L^\rho(x, y))\right\|_2^2.$$

Hence, this example satisfies all assumptions of Theorem 3.

4 Projected backward Euler: a sequential homotopy method

It is well-known that the projection in (8) is actually the derivative of the projection of the primal variable onto C in direction of the negative primal gradient:

Lemma 6 *For a nonempty closed convex set $K \subseteq X$, the Gâteaux derivative of the projection of $x \in X$ onto K in the direction $\delta x \in X$ is the projection of δx onto the tangent cone $T(K, x)$, i.e.,*

$$\lim_{h \to 0^+} h^{-1}\left(P_K(x + h\delta x) - x\right) = P_{T(K,x)}(\delta x).$$

Proof See [59, Lemma 4.5]. □

This motivates following the flow defined by (8) from $(\hat{x}, \hat{y}) \in C \times Y$ to $(x, y) \in C \times Y$ with a projected backward Euler step of stepsize $\Delta t > 0$ by solving

$$x - P_C\left(\hat{x} - \Delta t \nabla_x L^\rho(x, y)\right) = 0, \qquad y - \hat{y} - \Delta t c(x) = 0, \tag{21}$$

because Lemma 6 ensures consistency by virtue of

$$\lim_{\Delta t \to 0} \frac{x - \hat{x}}{\Delta t} = \lim_{\Delta t \to 0} \frac{P_C\left(\hat{x} - \Delta t \nabla_x L^\rho(x, y)\right) - \hat{x}}{\Delta t} = P_{T(C,\hat{x})}\left(-\nabla_x L^\rho(\hat{x}, \hat{y})\right).$$

From a computational point of view, the projected backward Euler system (21) is an ideal candidate for the application of local (possibly inexact) semismooth Newton methods (see, e.g., [43,54,57]), which we will investigate in more detail in Sect. 5.

In addition, the projected backward Euler system (21) can be interpreted as necessary optimality conditions of a primal-dual proximally regularized version of the augmented form of (1). With $\lambda = 1/\Delta t$, it reads

$$\begin{aligned} &\min\ \phi^\rho(x) + \lambda\left[\tfrac{1}{2}\left\|x - \hat{x}\right\|_X^2 + \tfrac{1}{2}\left\|w - \hat{y}\right\|_Y^2\right] \text{ over } w \in Y, x \in C \\ &\text{subject to } c(x) + \lambda w = 0. \end{aligned} \tag{22}$$

Uniqueness of solutions to (22) can be guaranteed for sufficiently large λ.

Theorem 4 *The regularized problem (22) has the following properties for $\lambda > 0$:*

1. *It satisfies the strong constraint qualification of Lemma 4.*
2. *Its primal-dual solutions $(\bar{w}, \bar{x}, \bar{y}) \in Y \times C \times Y$ satisfy (21) and $\bar{w} = -\Delta t c(\bar{x})$.*
3. *For $\lambda \to \infty$, i.e., $\Delta t \to 0$, its unique primal-dual solution $(\bar{w}, \bar{x}, \bar{y})$ tends to $(0, \hat{x}, \hat{y})$ provided that ∇L^ρ is globally Lipschitz continuous.*

For $\lambda = 0$, i.e., $\Delta t = \infty$, its primal-dual solutions $\bar{x}$ and $\bar{y}$ coincide with those of problem (1) and arbitrary $\bar{w} \in Y$.

Proof With $U = Y$, $Q = X$, we can apply Lemma 4 with the solution mapping $S(x) = -\Delta t c(x)$ and $\operatorname{ran} \lambda \mathrm{I}_Y = \operatorname{ran} \lambda \mathrm{I}_Y^\star = Y$. This shows assertion 1. We call the Lagrangian of (22) homotopy Lagrangian or proximal Lagrangian and denote it by

$$L^{\lambda,\rho}(w, x, y) = L^\rho(x, y) + \lambda\left[\tfrac{1}{2}\left\|x - \hat{x}\right\|_X^2 + \tfrac{1}{2}\left\|w - \hat{y}\right\|_Y^2 + (y, w)_Y\right].$$

By Lemma 4, GCQ holds at all feasible points and Theorem 1 yields that

$$\left(-\nabla_w L^{\lambda,\rho}(\bar{w}, \bar{x}, \bar{y}), -\nabla_x L^{\lambda,\rho}(\bar{w}, \bar{x}, \bar{y})\right) \in \{0\} \times T^-(C, \bar{x}), \tag{23}$$

from which we can deduce that $\bar{w} = \hat{y} - \bar{y}$ because of

$$0 = \nabla_w L^{\lambda,\rho}(\bar{w}, \bar{x}, \bar{y}) = \lambda(\bar{w} - \hat{y} + \bar{y}). \tag{24}$$

Hence, the feasibility of $(\bar{w}, \bar{x})$ implies that

$$c(\bar{x}) + \lambda[\hat{y} - \bar{y}] = 0.$$

Multiplication with Δt yields the second equation of (21). For the x-part of (23), we observe that

$$-\Delta t \nabla_x L^{\lambda,\rho}(\bar{w}, \bar{x}, \bar{y}) = -\bar{x} + \hat{x} - \Delta t \nabla_x L^{\rho}(\bar{x}, \bar{y}) \in T^{-}(C, \bar{x}),$$

implying $\hat{x} - \Delta t \nabla_x L^{\rho}(\bar{x}, \bar{y}) \in T^{-}(C, \bar{x}) + \bar{x}$ and by Lemma 2 that therefore $P_C(\hat{x} - \Delta t \nabla_x L^{\rho}(\bar{x}, \bar{y})) = \bar{x}$, which coincides with the first equation of (21). This shows assertion 2.

We can now use (21) to define a fixed point iteration $z^{k+1} = \Phi(z^k)$ on $C \times Y$ via

$$\Phi(z) = (P_C(\hat{x} - \Delta t \nabla_x L^{\rho}(z)), \hat{y} + \Delta t \nabla_y L^{\rho}(z)).$$

Let ω denote the Lipschitz constant of ∇L^{ρ}. For $\Delta t < \frac{1}{\omega}$, the mapping Φ is a contraction because P_C is Lipschitz continuous with modulus 1:

$$\begin{aligned}\|\Phi(z) - \Phi(\tilde{z})\|_{X \times Y}^2 &= \left\| P_C(\hat{x} - \Delta t \nabla_x L^{\rho}(z)) - P_C(\hat{x} - \Delta t \nabla_x L^{\rho}(\tilde{z})) \right\|_X^2 \\ &\quad + \Delta t^2 \left\| \nabla_y L^{\rho}(z) - \nabla_y L^{\rho}(\tilde{z}) \right\|_Y^2 \leq (\omega \Delta t)^2 \|z - \tilde{z}\|_{X \times Y}^2 .\end{aligned}$$

The Banach fixed point theorem yields uniqueness and existence of a fixed point $(\bar{x}, \bar{y})$, which together with $\bar{w} = \hat{y} - \bar{y}$ is the unique solution of (22). For $\Delta t = 0$, the fixed point and thus the solution is obviously $(0, \hat{x}, \hat{y})$. In order to prove convergence of $(\bar{w}, \bar{x}, \bar{y})$ to $(0, \hat{x}, \hat{y})$ for $\Delta t \to 0$, we observe that

$$\begin{aligned}\left\| \bar{z} - \hat{z} \right\|_{X \times Y} &= \left\| \Phi(\bar{z}) - \Phi(\hat{z}) + \Phi(\hat{z}) - \hat{z} \right\|_{X \times Y} \\ &\leq \omega \Delta t \left\| \bar{z} - \hat{z} \right\|_{X \times Y} + \left\| \Phi(\hat{z}) - \hat{z} \right\|_{X \times Y} ,\end{aligned}$$

which implies (recall that $\Phi(\hat{z})$ depends continuously on Δt)

$$\left\| \bar{z} - \hat{z} \right\|_{X \times Y} \leq \frac{1}{1 - \omega \Delta t} \left\| \Phi(\hat{z}) - \hat{z} \right\|_{X \times Y} \to 0 \quad \text{for } \Delta t \to 0.$$

This finally proves assertion 3. □

The artificial introduction of the variable w in (22) allows a lifting of the dual regularization term $y - \hat{y}$ in the backward Euler system (21) onto primal variables. From a linear algebra perspective, this can be understood as a Schur complement approach, as we see in the following example.

Example 2 *(A quadratic program)* Let $C = X$, $\rho = 0$, $\Delta t > 0$, $c(x) = Ax - b$, and $\phi(x) = \frac{1}{2}(x, Hx)_X - (g, x)_X$ for $A \in \mathcal{L}(X, Y)$, $H = H^{\star} \in \mathcal{L}(X, X)$, $g \in X$, and

$b \in Y$. The necessary optimality conditions of the homotopy problem (22) are then equivalent to the linear system

$$\begin{pmatrix} \lambda \mathrm{I}_Y & 0 & \lambda \mathrm{I}_Y \\ 0 & H + \lambda \mathrm{I}_X & A^\star \\ \lambda \mathrm{I}_Y & A & 0 \end{pmatrix} \begin{pmatrix} \bar{w} \\ \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} \lambda \hat{y} \\ \lambda \hat{x} + g \\ b \end{pmatrix}.$$

If we eliminate $\bar{w}$ with a Schur complement approach, we obtain the backward Euler system (21) as a primal-dual regularization of the original saddle point system for (1) according to

$$\left[\begin{pmatrix} H & A^\star \\ A & 0 \end{pmatrix} + \lambda \begin{pmatrix} \mathrm{I}_X & 0 \\ 0 & -\mathrm{I}_Y \end{pmatrix} \right] \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} g + \lambda \hat{x} \\ b - \lambda \hat{y} \end{pmatrix}.$$

We can derive two interesting equivalent reformulations of (22). The first reformulation substitutes $v = \sqrt{\lambda} w$, from which we obtain

$$\begin{aligned} &\min\ \phi^\rho(x) + \frac{\lambda}{2} \left\| x - \hat{x} \right\|_X^2 + \frac{1}{2} \left\| v - \sqrt{\lambda} \hat{y} \right\|_Y^2 \quad \text{over } v \in Y, x \in C \\ &\text{subject to } c(x) + \sqrt{\lambda} v = 0. \end{aligned} \tag{25}$$

The advantage of (25) over (22) is that the optimal v is also uniquely determined for $\lambda = 0$. The second reformulation completely eliminates $w = -\Delta t c(x)$. This leads to the problem

$$\min\ \phi^\rho(x) + \frac{\lambda}{2} \left\| x - \hat{x} \right\|_X^2 + \frac{1}{2} \left\| \sqrt{\lambda} \hat{y} + \sqrt{1/\lambda} c(x) \right\|_Y^2 \quad \text{over } x \in C,$$

which has no equality constraint and might allow for the application of projected Newton/gradient methods similar to, e.g., [14,16,38].

The homotopy problem (22) and Theorem 4 provide a complementary interpretation of using projected backward Euler steps (21) for the gradient flow Eq. (8): we trace the solutions of (22) from some primal-dual starting point $(0, \hat{x}, \hat{y})$ as a continuation in λ until the homotopy breaks down. The result yields an update for $(\hat{x}, \hat{y})$ and we can repeat the procedure. If, at one point, we are able to drive λ to zero, we can solve the original problem (1) with superlinear local convergence rate by the means of a locally superlinearly convergent method for the homotopy problem (22), e.g., a semismooth Newton method. If it is never possible to drive λ to zero, we at least follow the gradient flow (8) with a projected backward Euler method with stepsize $1/\lambda$. If we fix λ to some positive value, we obtain a locally linear convergence rate provided that the gradient flow converges exponentially.

5 Numerical case study in PDE constrained optimization

We apply the proposed method to the following benchmark problem adapted from [42]: Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary and let constants $a, b, \gamma >$

0, control bounds $q_{\mathrm{l}}, q_{\mathrm{u}} \in L^r(\Omega)$, $r \in (2, \infty]$, and a target function $u_{\mathrm{d}} \in L^2(\Omega)$ be given. We solve the control-constrained quasilinear elliptic optimal control problem

$$\begin{aligned}
&\min \ \frac{1}{2}\int_\Omega |u - u_{\mathrm{d}}|^2 + \frac{\gamma}{2}\int_\Omega |q|^2 \quad \text{over } u \in H_0^1(\Omega), q \in L^2(\Omega)\\
&\text{subject to } \nabla \cdot \left(\left[a + b\,|u|^2\right]\nabla u\right) = q,\\
&q_{\mathrm{l}} \le q \le q_{\mathrm{u}}.
\end{aligned} \tag{26}$$

In addition to [42], we include pointwise control bounds. For smaller values of a and γ, problem (26) becomes more and more ill-conditioned, while the effects of nonlinearity become more challenging for larger values of b.

To transform problem (26) into the form (1), we use the variables $x = (u, q) \in X = U \times Q = H_0^1(\Omega) \times L^2(\Omega)$, $y \in Y = U^* = H^{-1}(\Omega)$ and define the closed convex set

$$C = U \times \{q \in Q \mid q_{\mathrm{l}} \le q \le q_{\mathrm{u}}\} =: U \times C_Q \subset U \times Q = X$$

and the functions $\phi : X \to \mathbb{R}$ and $c : X \to Y$ via

$$\begin{aligned}
\phi((u, q)) &= \frac{1}{2}\int_\Omega |u - u_{\mathrm{d}}|^2 + \frac{\gamma}{2}\int_\Omega |q|^2\\
\langle c((u, q)), \varphi\rangle_{U^*,U} &= \int_\Omega \nabla\varphi \cdot \left[a + b\,|u|^2\right]\nabla u - \int_\Omega \varphi q \quad \text{for all } \varphi \in U,
\end{aligned}$$

where c is the weak form of the PDE in (26). The problem has a continuously Fréchet-differentiable solution operator $S : Q \to U$ in the sense of Lemma 4 [17].

5.1 Implementation aspects

From an implementation point of view, the projected backward Euler system (21) with all its required derivatives can be conveniently generated by the use of the Unified Form Language [2,4] in combination with Algorithmic Differentiation [25], as it is implemented in the DOLFIN/FEniCS project [3,39–41].

When evaluating the augmented objective $\phi^\rho(x) = \phi(x) + \frac{\rho}{2}\|c(x)\|_Y^2$, the inner product $(y, c(x))_Y$, or the dual proximal term in (22), we face the problem of computing norms and inner products in $Y = H^{-1}(\Omega)$, which we can facilitate computationally with the use of the Riesz isomorphism $\|y\|_Y = \|R_U y\|_U$. If we choose the norm $\|u\|_U = \|\nabla u\|_{L^2(\Omega)^2}$ on U, the evaluation of $R_U y$ boils down to one solution of a Poisson problem with right-hand side y and homogeneous Dirichlet boundary conditions. The difficulty from a computational vantage point is that R_U is a large dense matrix in contrast to its inverse R_U^{-1}, which is a sparse finite element stiffness matrix. For practical purposes, we always work with the Riesz representation of the dual variable $y_R = R_U y$ directly, eliminating the need for evaluating the Riesz isomorphism for the dual variables.

From a linear algebra point of view, it is important to exploit the special structure of the augmentation term $\frac{\rho}{2}\|c(x)\|_Y^2$. We extend a well-known argument for the special case of $\lambda = 0$ (see, e.g., [36, p. 158f]) to the case $\lambda \geq 0$: For fixed (x, y), let us denote the gradients and the second derivative of the augmented Lagrangian $L^\rho(x, y)$ by

$$\nabla_x L^\rho(x, y) = \nabla_x L^0(x, y + \rho c(x)) =: F_1, \quad \nabla_y L^\rho(x, y) = c(x) =: F_2,$$
$$\nabla_{xx} L^\rho(x, y) = \nabla_{xx} L^0(x, y + \rho c(x)) + \rho \nabla c(x) c'(x) =: H + \rho A^\star A.$$

Disregarding inequalities for a moment, each Newton step for the (appropriately scaled) backward Euler equation (21) requires us to solve the linear system

$$\begin{pmatrix} \lambda \mathrm{I}_X + H + \rho A^\star A & A^\star \\ A & -\lambda \mathrm{I}_Y \end{pmatrix} \begin{pmatrix} \delta x \\ \delta y \end{pmatrix} = - \begin{pmatrix} F_1 + \lambda(x - \hat{x}) \\ F_2 - \lambda(y - \hat{y}) \end{pmatrix}. \tag{27}$$

The problem here is that $A^\star A = R_X A^* R_Y^{-1} A = R_X A^* R_U A$ becomes a dense matrix after discretization by finite elements due to R_U. Hence, we must avoid the formation of $A^\star A$. Instead of (27) we solve the equivalent system

$$\begin{pmatrix} \lambda \mathrm{I}_X + H & A^\star \\ A & -(1 + \rho\lambda)^{-1} \lambda \mathrm{I}_Y \end{pmatrix} \begin{pmatrix} \delta x \\ \delta \tilde{y} \end{pmatrix} = - \begin{pmatrix} F_1 + \lambda(x - \hat{x}) \\ (1 + \rho\lambda)^{-1} \left(F_2 - \lambda(y - \hat{y})\right) \end{pmatrix} \tag{28}$$

with the reconstruction $\delta y = (1 + \rho\lambda)^{-1}(\delta\tilde{y} + \rho F_2)$. The equivalence can easily be checked. Because we work with $y_R = R_U y$ directly, we need to compute the Riesz representation $c_R = R_U c(x)$ first, evaluate the Lagrangian derivatives at $(x, y_R + \rho c_R)$, solve the unaugmented Newton system (28) (reformulated for y_R instead of y) for $(\delta x, \delta \tilde{y}_R)$, and finally reconstruct $\delta y_R = (1 + \rho\lambda)^{-1}(\delta \tilde{y}_R + \rho c_R)$.

The enforcement of the projection onto C in (21) can be easily implemented on top of (28): Let us consider the block row corresponding to the gradient with respect to u in (21) scaled by λ, which reads

$$\begin{aligned} 0 &= \lambda q - \lambda P_{C_Q}\left(\hat{q} - \Delta t \nabla_q L^\rho((u, q), y)\right) \\ &= \lambda q - \lambda P_{C_Q}\left(\hat{q} - \Delta t \left[\gamma q - R_U(y + \rho c(x))\right]\right). \end{aligned}$$

This nonsmooth equation together with the remaining smooth block rows of (21) scaled by λ can be solved efficiently with a semismooth Newton method. To this end, we need to address a norm gap for the pointwise defined projector

$$P_{C_Q}(q)(\xi) = \max(q_{\mathrm{l}}(\xi), \min(q(\xi), q_{\mathrm{u}}(\xi))) \quad \text{for } \xi \in \Omega,$$

which is known to be semismooth only if it maps from $L^r(\Omega) \subsetneq Q$ to $Q = L^2(\Omega)$ (see, e.g., [57, Sect. 3.3] or [31, Theorem 4.2]). Indeed, this higher regularity holds here if the initial guess satisfies $q_0 \in L^r(\Omega)$: For problem (26), the Q part of the projected backward Euler equation (21) simplify to $q = P_{C_Q}\left(\hat{q} - \Delta t \left[\gamma q - R_U(y + \rho c(x))\right]\right)$. By induction, we can assume that $q, \hat{q} \in L^r(\Omega)$. Then, the argument of the projection

Algorithm 1: Sequential homotopy method

```
Data: $z_0 = (x_0, y_0) \in C \times Y = Z$, $\Theta \in (0, 1)$, $\lambda_{\text{term}} > 0$, $\lambda_{\text{inc}} > 1$, $\text{TOL} > 0$
1  Initialize $z = (x, y) \leftarrow z_0$, $\lambda \leftarrow 1$
2  Loop (over homotopies)
3      Initialize $\hat{z} = (\hat{x}, \hat{y}) \leftarrow z$
4      Loop (to trace single homotopy leg)
5          Compute $z^+$ by one semismooth Newton step for (21) starting from $z$
6          Compute $z^{++}$ by one simplified semismooth Newton step for (21) starting from $z^+$
7          if $\|z^{++} - z^+\|_Z \leq \Theta \|z^+ - z\|_Z$ then
8              Accept iterate $z \leftarrow z^{++}$
9              if $\lambda \leq \lambda_{\text{term}}$ and $\|z - \hat{z}\|_Z \leq \text{TOL}$ then return solution $z$ Update homotopy parameter
               $\lambda$ (e.g., by a proportional-integral controller)
10             break inner loop
11         else increase homotopy parameter $\lambda \leftarrow \lambda_{\text{inc}}\lambda$
```

operator P_{C_Q} also lies in $L^r(\Omega)$, because $R_U[y + \rho c(x)] \in H_0^1(\Omega)$, which is continuously embedded in $L^r(\Omega)$. Because $q_l, q_u \in L^r(\Omega)$, we obtain $q = P_{C_Q}(.) \in L^r(\Omega)$, which completes the induction step.

5.2 Solution algorithm

We provide in Algorithm 1 pseudocode for a prototypical implementation of the sequential homotopy method with a classical continuation approach. It consists of an outer loop over the subsequent homotopies. In the inner loop, the reference point $\hat{z} = (\hat{x}, \hat{y})$ is fixed and we trace the solution of (21) with one semismooth Newton step followed by one inexact semismooth Newton step.

The computationally heavy part is the computation of z^+ in line 5 by one local semismooth Newton step at z and of z^{++} in line 6 by one local simplified semismooth Newton step at z^+. Here, *simplified* means that the system matrix of the previous semismooth Newton system is reused, subject to modifications concerning the current active set guess derived from the residual evaluated at z^+. We accept an iterate for the current value of λ if the following natural monotonicity test is satisfied in line 7: We require that the simplified semismooth Newton increment is smaller in norm than a contraction factor $\Theta \in (0, 1)$ times the semismooth Newton increment.

If the monotonicity test fails, we enlarge λ by a constant factor to drive the solution of (21) closer to $\hat{z}$ in order to eventually enter the region of local superlinear convergence of the semismooth Newton method.

If the monotonicity test is satisfied, we accept z^{++} as the new iterate. If λ and the norm of the outer loop increment $z - \hat{z}$ are small enough, then we terminate with the solution z, otherwise we predict a new stepsize which should eventually drive λ close to zero. We then commence the next outer iteration.

There are many possibilities to predict the next λ after acceptance of the current iterate. For the numerical results below, we use a heuristic motivated by a discrete proportional-integral (PI) controller: We try to choose λ such that the contraction factor $\theta = \|z^{++} - z^+\|_Z / \|z^+ - z\|_Z$ is close to a given reference $\theta_{\text{ref}} \in (0, 1)$. We choose to predict $\lambda \leftarrow \lambda/\lambda_{\text{mod}}$, where $\log \lambda_{\text{mod}}$ is the manipulated variable. To this end, let $e = \log \theta_{\text{ref}} - \log \theta$ and let I denote the sum of all previous values of e over the last successful outer loops. We then set with some constants K_P and K_I

$$\log \lambda_{\text{mod}} \leftarrow K_P e + K_I I.$$

In each accepted iteration, we have the simple update $I \leftarrow I + e$. In case the monotonicity test fails, we possibly reset the integral term $I \leftarrow \min(I, 0)$. We can also clip λ at a lower bound $\lambda_{\min}$. For a related concept in the stepsize control of one-step methods for ordinary differential equations we refer to [27, p. 28ff].

It is also possible to keep all iterates inside C with an additional projection in the local semismooth Newton step (see, e.g., [57]). We found the method to require fewer iterations on (26) without projection steps, even though we are aware that if $z \notin C$, we might run into problems with the monotonicity test in line 7 of Algorithm 1 because $\left\| z^+ - z \right\|_Z$ might not tend to 0 for $\lambda \to \infty$.

Alternatively to Algorithm 1, it is conceivable to update the reference point $\hat{z}$ less frequently and to trace each homotopy leg until it nearly breaks down in a singularity. In our experience, this approach of long homotopy legs leads to a more complicated algorithm and requires the solution of more and worse conditioned linear systems. We prefer the sequential homotopy method with short homotopy legs in the form of Algorithm 1.

5.3 Numerical results

We apply Algorithm 1 to problem (26) on $\Omega = (0, 1)^2$ with the target state $u_{\text{d}}(\xi) = 12(1 - \xi_1)\xi_1(1 - \xi_2)\xi_2$ from [42] and control bounds

$$q_{\text{l}}(\xi) = -50, \qquad q_{\text{u}}(\xi) = \min\left(50, 800 \max\left(\left(\xi_1 - \tfrac{1}{2}\right)^2, \left(\xi_2 - \tfrac{1}{2}\right)^2\right)\right)$$

for the parameters $a = 10^{-p}$, $b = 10^p$ for $p = 0, \ldots, 5$ with continuous piecewise linear (P1) finite elements on regular triangular grids with $N = 64, 128, 256, 512$ elements along each side of the unit square.

We perform Algorithm 1 with the initial guess $z_0 = 0$ and the parameters $\Theta = 0.9$, $\lambda_{\text{term}} = 10^{-8}$, $\lambda_{\text{inc}} = 2$, and $\text{TOL} = 10^{-8}$. We fix the choice of the penalty parameter to $\rho = 0.1$. For the stepsize PI controller, we set $\theta_{\text{ref}} = 0.5$, $K_P = 0.2$, $K_I = 0.005$, and $\lambda_{\min} = 10^{-12}$. Figures 1 and 2 depict the resulting optimal controls and states.

We compare the sequential homotopy method of Algorithm 1 with a nonlinear VI solver described in [13,45] and implemented in the production quality software package PETSc [11,12]. For better comparison, we use the direct solver MUMPS [6,7] for the solution of the linear systems in both approaches. The use of inexact linear algebra solvers is no conceptual problem, as long as they yield a locally convergent nonlinear iteration. The efficiency of iterative linear algebra methods, however, depends crucially on the use of suitable structure-exploiting preconditioners. This topic exceeds the scope of this paper and is the subject of future research.

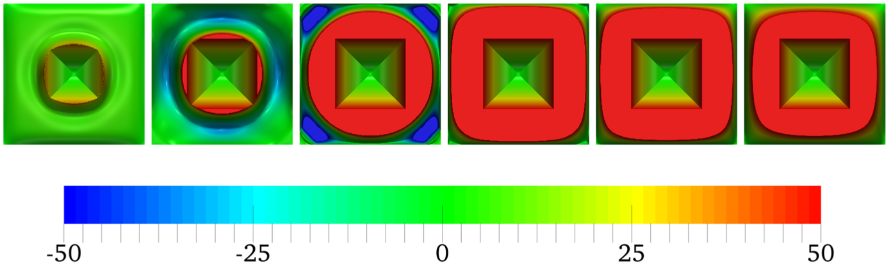

Fig. 1 Optimal controls for problem (26) with $a = 10^{-p}$ and $b = 10^{p}$ for $p = 0, 1, \dots, 5$ from left to right. The lower bounds at -50 are only active for $p = -2$ (deep blue) (color figure online)

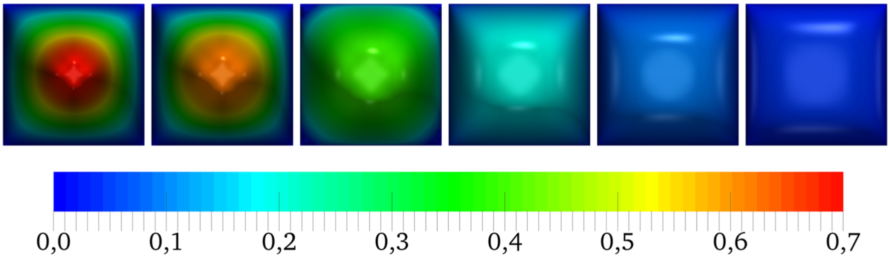

Fig. 2 Optimal states for problem (26) with $a = 10^{-p}$ and $b = 10^{p}$ for $p = 0, 1, \dots, 5$ from left to right (color figure online)

For the VI solver, we consider two implemented globalization strategies, a backtracking line-search (bt) and an error-oriented monotonicity test (nleqerr). As it turns out, the VI solver did not solve any of the problem instances when started at the initial guess $z_0 = 0$, failing either by raising an error or reaching the limit of 5.000 residual evaluations, even for a reduced termination tolerance of 10^{-5} on the l^∞-norm of the residuals. Some problem instances could be solved successfully after dropping the lower control bound, which is only active for $a = 10^{-2}$, $b = 10^{2}$. In some of these instances the residual norm stalled between 10^{-5} and 10^{-8}.

We compare in Table 1 the sequential homotopy method of Algorithm 1 (with a sharper termination tolerance of 10^{-8} on the Z-norm of the homotopy increment and upper and lower bounds) to the VI approach with reduced termination tolerance as above and only upper bounds. We can observe that the sequential homotopy method solves all problem instances with mesh-independent convergence (subject to some fluctuation for the worse conditioned problems). The VI approach with backtracking is faster for the less demanding but fails for the more demanding instances. The VI approach with error-oriented monotonicity test solves at least two of the more demanding instances successfully, although only one with an efficiency comparable to the sequential homotopy method.

In Fig. 3, we see that even though slightly different numbers of iterations (depicted with markers) are performed on different meshes for the case $a = 10^{-2}$, $b = 10^{2}$, roughly the same flow time of 10^{11} has to be traversed to reach the required tolerance of $\mathrm{TOL} = 10^{-8}$. We also see that the stepsizes Δt eventually become very large

Table 1 Comparison of the sequential homotopy method of Algorithm 1 with a nonlinear VI solver with backtracking (bt) and error-oriented monotonicity test (nleqerr) for different instances of problem (26) and varying discretizations (N)

Problem parameters			Solution	Sequential homotopy			VI (bt)		VI (nleqerr)	
$\log_{10} a$	$\log_{10} b$	N	#act	#disc	#mat	#res	#mat	#res	#mat	#res
0	0	64	637	0	20	40	9	26	↗	
0	0	128	2545	0	21	42	9	26	↗	
0	0	256	10101	0	20	40	9	26	40	123
0	0	512	40193	0	20	40	9	26	12	33
− 1	1	64	1121	0	32	64	16	67	↗	
− 1	1	128	4405	0	31	62	13	50	↗	
− 1	1	256	17525	0	32	64	12	42	↗	
− 1	1	512	69857	0	32	64	11	37	24	69
− 2	2	64	2897	5	55	115	↗		↗	
− 2	2	128	11533	15	75	165	⤩		↗	
− 2	2	256	45649	4	60	124	⤩		↗	
− 2	2	512	182293	5	58	121	↗		↗	
− 3	3	64	3505	1	46	93	↗		110	378
− 3	3	128	13997	1	47	95	↗		↗	
− 3	3	256	55709	4	55	114	↗		↗	
− 3	3	512	222385	3	54	111	↗		↗	
− 4	4	64	3405	4	59	122	↗		∞	
− 4	4	128	13477	4	56	116	↗		54	137
− 4	4	256	53609	5	60	125	↗		↗	
− 4	4	512	214009	4	63	130	↗		↗	
− 5	5	64	2933	11	73	157	↗		↗	
− 5	5	128	11609	10	78	166	↗		↗	
− 5	5	256	46265	14	82	178	↗		↗	
− 5	5	512	184657	14	83	180	↗		↗	

The cardinality of the discrete optimal active set is given in the #act column. The column #disc shows the number of discarded steps, which are reasonably low, hinting at the efficiency of the PI control stepsize prediction. The columns #mat and #res show the number of required matrix and residual evaluations. The sequential homotopy method solves all instances and exhibits mesh-independent convergence (subject to some fluctuations for the worse conditioned problems). The symbol ⤩ denotes an error in the line-search, the symbol ∞ an error after exceeding 5.000 residual evaluations, and the symbol ↗ an error after not more than 10^{-8} relative reduction of a criticality measure over 100 system matrix evaluations

and lead to superlinear convergence. This is the typical numerical behavior of the sequential homotopy method on all considered instances. For $N = 128$ some extra steps are carried out around $t = 10^5$ and $t = 10^7$.

6 Summary

We provided sufficient conditions for the existence of global solutions to the projected gradient/antigradient flow (8) and showed that critical points with emanating

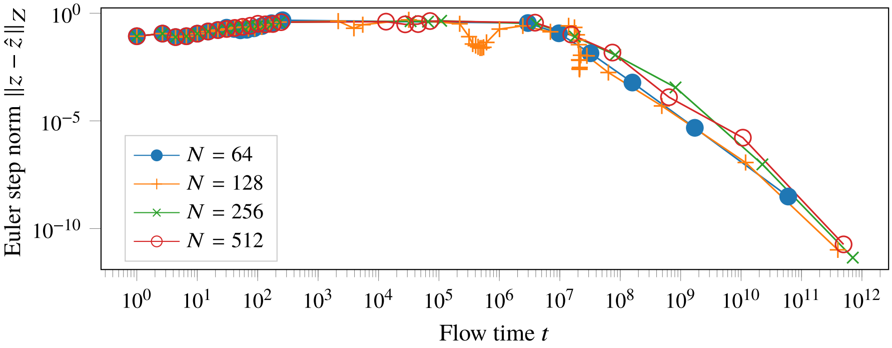

Fig. 3 The projected backward Euler step norms $\|z - \hat{z}\|_Z$ for (26) with $a = 10^{-2}$, $b = 10^{2}$ on different meshes plotted with respect to the flow time t, which is the sum of all accepted step sizes $\Delta t = 1/\lambda$ (color figure online)

descent curves cannot be asymptotically stable and are thus not attracting for the flow. We applied projected backward Euler timestepping to derive the necessary optimality conditions of a primal-dual proximally regularized counterpart (22) of (1). The regularized problem can be solved by a homotopy method, giving rise to a sequence of homotopy problems. The sequential homotopy method can be used to globalize any locally convergent optimization method that can be employed efficiently in a homotopy framework. The sequential homotopy method with a local semismooth Newton solver outperforms state-of-the-art VI solvers for a challenging class of PDE-constrained optimization problem with control constraints.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Absil, P.A., Kurdyka, K.: On the stable equilibrium points of gradient systems. Syst. Control Lett. **55**(7), 573–577 (2006)
2. Alnæs, M.S.: UFL: a finite element form language. In: Logg, A., Mardal, K.A., Wells, G.N. (eds.) Automated Solution of Differential Equations by the Finite Element Method. Lecture Notes in Computational Science and Engineering, Chap. 17, vol. 84. Springer, Berlin, Heidelberg (2012)
3. Alnæs, M.S., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M.E., Wells, G.N.: The FEniCS project version 1.5. Arch. Numer. Softw. **3**(100) (2015)
4. Alnæs, M.S., Logg, A., Ølgaard, K.B., Rognes, M.E., Wells, G.N.: Unified form language: a domain-specific language for weak formulations of partial differential equations. ACM Trans. Math. Softw. **40**(2) (2014)

5. Amann, H.: Ordinary Differential Equations. De Gruyter Studies in Mathematics, vol. 13. Walter de Gruyter & Co., Berlin (1990). (An introduction to nonlinear analysis, Translated from the German by Gerhard Metzen)
6. Amestoy, P.R., Duff, I.S., Koster, J., L'Excellent, J.Y.: A fully asynchronous multifrontal solver using distributed dynamic scheduling. SIAM J. Matrix Anal. A. **23**(1), 15–41 (2001)
7. Amestoy, P.R., Guermouche, A., L'Excellent, J.Y., Pralet, S.: Hybrid scheduling for the parallel solution of linear systems. Parallel Comput. **32**(2), 136–156 (2006)
8. Arrow, K.J., Hurwicz, L., Uzawa, H.: Studies in linear and non-linear programming. With contributions by Chenery, H.B., Johnson, S.M., Karlin, S., Marschak, T., Solow, R.M., Stanford Mathematical Studies in the Social Sciences, Vol. II. Stanford University Press, Stanford (1958)
9. Ascher, U., Osborne, M.R.: A note on solving nonlinear equations and the natural criterion function. J. Optim. Theory Appl. **55**(1), 147–152 (1987)
10. Aubin, J.P., Cellina, A.: Differential Inclusions, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 264. Springer, Berlin (1984). (Set-valued maps and viability theory)
11. Balay, S., Abhyankar, S., Adams, M.F., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Dener, A., Eijkhout, V., Gropp, W.D., Kaushik, D., Knepley, M.G., May, D.A., McInnes, L.C., Mills, R.T., Munson, T., Rupp, K., Sanan, P., Smith, B.F., Zampini, S., Zhang, H., Zhang, H.: PETSc Web page (2018). http://www.mcs.anl.gov/petsc
12. Balay, S., Abhyankar, S., Adams, M.F., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Dener, A., Eijkhout, V., Gropp, W.D., Kaushik, D., Knepley, M.G., May, D.A., McInnes, L.C., Mills, R.T., Munson, T., Rupp, K., Sanan, P., Smith, B.F., Zampini, S., Zhang, H., Zhang, H.: PETSc users manual. Tech. Rep. ANL-95/11 - Revision 3.10, Argonne National Laboratory (2018)
13. Benson, S.J., Munson, T.S.: Flexible complementarity solvers for large-scale applications. Optim. Methods Softw. **21**(1), 155–168 (2006)
14. Bertsekas, D.P.: Projected Newton methods for optimization problems with simple constraints. SIAM J. Control Optim. **20**(2), 221–246 (1982)
15. Bock, H.G., Kostina, E., Schlöder, J.P.: On the role of natural level functions to achieve global convergence for damped Newton methods. In: System Modelling and Optimization (Cambridge, 1999), pp. 51–74. Kluwer Acad. Publ., Boston (2000)
16. Calamai, P.H., Moré, J.J.: Projected gradient methods for linearly constrained problems. Math. Program. **39**(1), 93–116 (1987)
17. Casas, E., Tröltzsch, F.: First- and second-order optimality conditions for a class of optimal control problems with quasilinear elliptic equations. SIAM J. Control Optim. **48**(2), 688–718 (2009)
18. Cojocaru, M.G., Daniele, P., Nagurney, A.: Projected dynamical systems and evolutionary variational inequalities via Hilbert spaces with applications. J. Optim. Theory Appl. **127**(3), 549–563 (2005)
19. Cojocaru, M.G., Jonker, L.B.: Existence of solutions to projected differential equations in Hilbert spaces. Proc. Am. Math. Soc. **132**(1), 183–193 (2004)
20. Davidenko, D.F.: On a new method of numerical solution of systems of nonlinear equations. Doklady Akad. Nauk SSSR (N.S.) **88**, 601–602 (1953)
21. Deuflhard, P.: A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting. Numer. Math. **22**, 289–315 (1974)
22. Deuflhard, P.: Newton Methods for Nonlinear Problems. Springer Series in Computational Mathematics, vol. 35. Springer, Berlin (2004). (Affine invariance and adaptive algorithms)
23. Deuflhard, P.: The grand four: affine invariant globalizations of Newton's method. Vietnam J. Math. **46**(4), 761–777 (2018)
24. Deuflhard, P., Weiser, M.: Global inexact Newton multilevel FEM for nonlinear elliptic problems. In: Multigrid Methods V (Stuttgart, 1996), Lect. Notes Comput. Sci. Eng., Vol. 3, pp. 71–89. Springer, Berlin (1998)
25. Griewank, A., Walther, A.: Evaluating Derivatives, 2nd edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2008). (Principles and techniques of algorithmic differentiation)
26. Guignard, M.: Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space. SIAM J. Control **7**, 232–241 (1969)
27. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations. II. Springer Series in Computational Mathematics, vol. 14, Second edn. Springer, Berlin (1996). (Stiff and differential-algebraic problems)
28. Hante, F.M., Mommer, M.S., Potschka, A.: Newton–Picard preconditioners for time-periodic parabolic optimal control problems. SIAM J. Numer. Anal. **53**(5), 2206–2225 (2015)

29. Hauswirth, A., Bolognani, S., Hug, G., Dörfler, F.: Projected gradient descent on Riemannian manifolds with applications to online power system optimization. In: 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 225–232. IEEE (2016)
30. Hauswirth, A., Subotić, I., Bolognani, S., Hug, G., Dörfler, F.: Time-varying projected dynamical systems with applications to feedback optimization of power systems. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 3258–3263. IEEE (2018)
31. Hintermüller, M.: Semismooth Newton methods and applications. Tech. rep., Department of Mathematics, Humboldt-University of Berlin (2010)
32. Hintermüller, M., Ito, K., Kunisch, K.: The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**(3), 865–888 (2002)
33. Hintermüller, M., Ulbrich, M.: A mesh-independence result for semismooth Newton methods. Math. Program. **101**(1, Ser. B), 151–184 (2004)
34. Hohmann, A.: Inexact Gauss Newton Methods for Parameter Dependent Nonlinear Problems. Ph.D. thesis, Freie Universität Berlin (1994)
35. Ito, K., Kunisch, K.: The primal-dual active set method for nonlinear optimal control problems with bilateral constraints. SIAM J. Control Optim. **43**(1), 357–376 (2004)
36. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications. Advances in Design and Control, vol. 15. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2008)
37. Jongen, H.T., Stein, O.: Nonconvex optimization: gradient flows and deformation. J. Dyn. Control Syst. **7**(3), 425–446 (2001)
38. Kelley, C.T., Sachs, E.W.: Multilevel algorithms for constrained compact fixed point problems. SIAM J. Sci. Comput. **15**(3), 645–667 (1994). (Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992))
39. Logg, A., Mardal, K.A., Wells, G.N. (eds.): Automated Solution of Differential Equations by the Finite Element Method. Springer (2012)
40. Logg, A., Wells, G.N.: DOLFIN: automated finite element computing. ACM Trans. Math. Softw. **37**(2) (2010)
41. Logg, A., Wells, G.N., Hake, J.: DOLFIN: a C++/Python finite element library. In: Logg, A., Mardal, K.A., Wells, G.N. (eds.) Automated Solution of Differential Equations by the Finite Element Method. Lecture Notes in Computational Science and Engineering, Chap. 10, vol. 84. Springer, Berlin, Heidelberg (2012)
42. Lubkoll, L., Schiela, A., Weiser, M.: An affine covariant composite step method for optimization with PDEs as equality constraints. Optim. Methods Softw. **32**(5), 1132–1161 (2017)
43. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM J. Control Optim. **15**(6), 959–972 (1977)
44. Moreau, J.J.: Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. C. R. Acad. Sci. Paris **255**, 238–240 (1962)
45. Munson, T.S., Facchinei, F., Ferris, M.C., Fischer, A., Kanzow, C.: The semismooth algorithm for large scale complementarity problems. INFORMS J. Comput. **13**(4), 294–311 (2001)
46. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006)
47. Pang, J.S., Stewart, D.E.: Differential variational inequalities. Math. Program. **113**(2, Ser. A), 345–424 (2008)
48. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends® Optim. **1**(3), 127–239 (2014)
49. Potschka, A.: A Direct Method for Parabolic PDE Constrained Optimization Problems. Advances in Numerical Mathematics. Springer, Wiesbaden (2013)
50. Potschka, A.: Direct multiple shooting for parabolic PDE constrained optimization. In: Carraro, T., Geiger, M., Körkel, S., Rannacher, R. (eds.) Multiple Shooting and Time Domain Decomposition Methods, pp. 159–181. Springer International Publishing, Cham (2015)
51. Potschka, A.: Backward step control for global Newton-type methods. SIAM J. Numer. Anal. **54**(1), 361–387 (2016)
52. Potschka, A.: Backward step control for Hilbert space problems. Numer. Algorithms 1–30 (2018)
53. Potschka, A., Mommer, M.S., Schlöder, J.P., Bock, H.G.: Newton–Picard-based preconditioning for linear-quadratic optimization problems with time-periodic parabolic PDE constraints. SIAM J. Sci. Comput. **34**(2), A1214–A1239 (2012)

54. Qi, L.Q., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**(3, Ser. A), 353–367 (1993)
55. Shikhman, V., Stein, O.: Constrained optimization: projected gradient flows. J. Optim. Theory Appl. **140**(1), 117–130 (2009)
56. Ulbrich, M.: Semismooth Newton methods for operator equations in function spaces. SIAM J. Optim. **13**(3), 805–842 (2002)
57. Ulbrich, M.: Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces, MOS-SIAM Series on Optimization, Vol. 11. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, (2011)
58. Yosida, K.: Functional Analysis. Classics in Mathematics. Springer, Berlin (1995). (Reprint of the sixth (1980) edition)
59. Zarantonello, E.H.: Projections on convex sets in Hilbert space and spectral theory. I. Projections on convex sets. In: Contributions to nonlinear functional analysis (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1971), pp. 237–341. Academic Press, New York (1971)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Andreas Potschka[1] · **Hans Georg Bock**[1]

✉ Andreas Potschka
potschka@iwr.uni-heidelberg.de

Hans Georg Bock
bock@iwr.uni-heidelberg.de

[1] Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany