SpezialForschungsBereich F 32

Karl–Franzens Universität Graz
Technische Universität Graz
Medizinische Universität Graz

A Hybrid Semismooth Quasi-Newton Method Part 2: Applications

F. Mannel A. Rund

SFB-Report No. 2018-006 Sep 2018

A–8010 GRAZ, HEINRICHSTRASSE 36, AUSTRIA

Supported by the
Austrian Science Fund (FWF)

Der Wissenschaftsfonds.

SFB sponsors:

- **Austrian Science Fund** (FWF)
- **University of Graz**
- **Graz University of Technology**
- **Medical University of Graz**
- **Government of Styria**
- **City of Graz**

A hybrid semismooth quasi-Newton method Part 2: Applications

Florian Mannel* Armin Rund*

September 24, 2018

A hybrid semismooth quasi-Newton method for structured nonsmooth operator equations in Banach spaces is introduced in a bipartite work. In the first part we have developed the local convergence theory of the new method and, in particular, proven its local q-superlinear convergence under reasonable assumptions.

In this second part we focus on applications. First, it is demonstrated how the new method can be applied to generalized variational inequalities and nonsmooth optimization problems, particularly problems from PDE-constrained optimal control. For the latter we discuss two examples in detail. This includes working out different choices of function spaces that are covered by the hybrid approach, yielding convergence results with respect to different norms. Second, numerical realizations of the hybrid method are established for Broyden, SR1 and BFGS updates. Reduced matrix-free Newton-cg type methods are combined with different globalization techniques, including line search strategies as well as a trust-region globalization.

The numerical methods are thoroughly studied on nonsmooth optimal control problems. The studies include an investigation of mesh independence properties, a comparison of different globalization techniques, and an examination of the effect of the different update formulas. They also feature a nonsmooth and nonconvex application from magnetic resonance imaging to highlight the capability of the new method for solving large-scale real-world optimal control problems; here, a trust-region algorithm performs particularly well. In all the studies we observe that the new method is several times faster in runtime than semismooth Newton methods, ranging from a factor of five to a factor larger than seventy.

Keywords: Semismooth Newton methods, quasi-Newton methods, generalized variational inequalities, nonsmooth optimization, optimal control, Bloch equations

1. Introduction

We continue our work from [20] on the hybrid semismooth quasi-Newton method for the solution of nonsmooth operator equations of the form

$$H(q) := F(G(q)) + \hat{G}(q) = 0, \tag{P}$$

where $G : Q \to U$ and $\hat{G} : Q \to V$ are semismooth, $F : U \to V$ is smooth, Q and V are Banach spaces, and U is a Hilbert space. The precise setting along with the algorithm and its main convergence results from [20] are summarized in Section 2.

*University of Graz, Heinrichstr. 36, 8010 Graz, Austria (florian.mannel@uni-graz.at, armin.rund@uni-graz.at).

In this paper we are concerned with different aspects of applications of the hybrid method. First, we show for different problem classes, including generalized variational inequalities and problems from PDE-constrained optimal control, that their first order optimality conditions can be expressed in the form (P), making them amenable to the hybrid method. Here, proximal mappings will be the main tool. Subsequently, we discuss in detail for PDE-constrained optimal control problems under which conditions the assumptions of the theory developed in [20] are fulfilled. In particular, we elaborate possible choices of function spaces, resulting in convergence of the iterates with respect to different norms. In the second half of this paper we turn to numerics. We devise various numerical realizations of the hybrid method and compare them as part of an extensive numerical study. In fact, this study addresses many theoretical and practical aspects of the hybrid approach, such as the experimental verification of its superlinear convergence with respect to different norms, an investigation of its mesh independence properties, a comparison of different globalization strategies, and an examination of different quasi-Newton updates (Broyden, SR1, BFGS).
The numerical study is based on two optimal control problems that are quite different in character. The first problem consists in time-dependent tracking of the linear heat equation. As nonsmooth components we consider box constraints and an L^1 term in the objective. We regard this problem as a prototype from the important class of nonsmooth convex optimization problems. It will allows us to display very clearly the convergence properties of the hybrid method, e.g., its local superlinear convergence. The second problem deals with the design of radio-frequency pulses for magnetic resonance imaging, a topic from medical engineering. Here, a realistic modeling yields a nonsmooth and nonconvex optimal control problem that serves as a benchmark for the performance of the hybrid approach on real-world applications. The numerical results underline that the hybrid approach can be very powerful in practice and competitive on such problems. Indeed, a previous version of the presented matrix-free limited-memory truncated trust-region method formed the kernel of the code [31] that won the ISMRM challenge on radio-frequency pulse design in magnetic resonance imaging [14]. In this paper we provide an improved version of this method. Above all, we will observe throughout the entire study that the algorithms derived from the hybrid approach are considerably faster than their semismooth Newton counterparts.

For a comprehensive list of literature on the convergence analysis of quasi-Newton methods and methods that combine semismooth and quasi-Newton methods we refer to [20]. Since we focus on the application of the new approach to problems from PDE-constrained optimal control, we point the reader to the monographs [10, 17, 18, 36, 37]. They are all devoted to PDE-constrained optimal control and include a variety of applications and numerical examples. In addition, some of these monographs also contain material on variational inequalities. More specific references for the topics that we address are provided in the corresponding sections.

The main contributions of this paper consist in

- demonstrating that the hybrid method is applicable to various problem classes;
- deriving sophisticated numerical implementations of the hybrid method;
- providing extensive numerical studies which underline that the hybrid approach results in algorithms that are substantially faster than semismooth Newton methods, including on a nonsmooth and nonconvex real-world problem.

This paper is organized as follows. In Section 2 we state the problem setting, the hybrid method, and its main convergence results. Section 3 discusses in-depth the application of the new algorithm to different problem classes. In Section 4 we comment on implementation issues. Section 5 contains the aforementioned numerical study and in Section 6 we provide conclusions of this work. In Appendix A we outline the algorithm that we found most effective in the numerical studies—a matrix-free limited-memory truncated trust-region variant of the hybrid method.

2. Problem setting, algorithm and convergence results

In this section we repeat the parts of [20] that are relevant to this work. All linear spaces that appear in this work are taken over the field of real numbers. If X and Y are two such spaces, then we write $\mathcal{L}(X, Y)$ for the set of bounded linear operators between X and Y.

2.1. Problem setting and algorithm

For the remainder of Section 2 let be given

- Banach spaces Q, V and a Hilbert space U with scalar product $(\cdot,\cdot)_U$;
- mappings $G : Q \to U$, $F : U \to V$ and $\hat{G} : Q \to V$;
- $H : Q \to V$, $H(q) := F(G(q)) + \hat{G}(q)$.

In this setting the goal is to

$$\text{(P)} \qquad \text{find } \bar{q} \in Q \text{ such that } H(\bar{q}) = 0.$$

Supposing for the moment that G and $\hat{G}$ are semismooth and that F is smooth, we can apply a semismooth Newton method to the operator equation $H(q) = 0$. A semismooth Newton step $s^k \in Q$ at the current iterate $q^k \in Q$ is computed by solving

$$\left(F'(G(q^k))M_k + \hat{M}_k\right) s^k = -H(q^k),$$

where $M_k \in \partial G(q^k) \subset \mathcal{L}(Q, U)$ and $\hat{M}_k \in \partial\hat{G}(q^k) \subset \mathcal{L}(Q, V)$ are elements of the set-valued generalized derivatives ∂G of G, respectively, $\partial\hat{G}$ of $\hat{G}$, evaluated at q^k. The key idea of the novel method is to replace $F'(G(q^k)) \in \mathcal{L}(U, V)$ by a quasi-Newton approximation $B_k \in \mathcal{L}(U, V)$, while the generalized derivatives of G and $\hat{G}$ are left unchanged. The resulting algorithm therefore combines a quasi-Newton method with a semismooth Newton method and can be regarded as a hybrid approach. Specifically, the algorithm reads as follows.

Algorithm 1: Hybrid semismooth quasi-Newton method

Input: $q^0 \in Q$, $B_0 \in \mathcal{L}(U, V)$, $0 \le \sigma_{\min} \le \sigma_{\max} \le 2$

1 Let $u^0 := G(q^0)$
2 **for** $k = 0, 1, 2, \ldots$ **do**
3 **if** $H(q^k) = 0$ **then** let $\bar{q} := q^k$; STOP
4 Choose $M_k \in \partial G(q^k)$ and $\hat{M}_k \in \partial\hat{G}(q^k)$
5 Let $\tilde{M}_k := B_k M_k + \hat{M}_k$
6 Solve $\tilde{M}_k s^k = -H(q^k)$ for s^k
7 Let $q^{k+1} := q^k + s^k$ and $u^{k+1} := G(q^{k+1})$
8 Let $s_u^k := u^{k+1} - u^k$ and $y^k := F(u^{k+1}) - F(u^k)$
9 Choose $\sigma_k \in [\sigma_{\min}, \sigma_{\max}]$
10 **if** $s_u^k \neq 0$ **then** let $B_{k+1} := B_k + \sigma_k (y^k - B_k s_u^k) \frac{(s_u^k, \cdot)_U}{\|s_u^k\|_U^2}$;
11 **else** let $B_{k+1} := B_k$
12 **end**

Output: $\bar{q}$

Let us briefly comment on selected features of Algorithm 1. We start by pointing out that the stopping criterion $H(q^k) = 0$ in line 3 is replaced in the numerical experiments by a more

practical choice such as $\|H(q^k)\|_V \leq \mathrm{TOL}\|H(q^0)\|_V$ with an appropriate relative tolerance $\mathrm{TOL} > 0$. Furthermore, we observe in the numerical experiments of Section 5 that the simple choice $(\sigma_k) \equiv 1$—i.e., the classical Broyden update—results in an efficient algorithm. This is noteworthy because it shows that no "parameter tweaking" is necessary for the parameter σ. Nonetheless, the incorporation of iteration-dependent choices of σ_k may be advantageous, for instance because it can be used to ensure that B_{k+1} is invertible if B_k is invertible.

Regarding the update formula in line 10 we furthermore remark that while the convergence analysis of [20] is concerned with the generalized Broyden update, the numerical study conducted in Section 5 also includes the usage of other popular quasi-Newton updates, namely

$$B_{k+1} = B_k + \left(y^k - B_k s_u^k\right) \frac{\left(y^k - B_k s_u^k, \cdot\right)_U}{\left(y^k - B_k s_u^k, s_u^k\right)_U} \qquad \text{(SR1)}$$

and

$$B_{k+1} = B_k + y^k \frac{\left(y^k, \cdot\right)_U}{\left(y^k, s_u^k\right)_U} - B_k s_u^k \frac{\left(B_k s_u^k, \cdot\right)_U}{\left(B_k s_u^k, s_u^k\right)_U} \qquad \text{(BFGS).}$$

Given that both the SR1 and the BFGS update are symmetric, Broyden's method seems more appropriate for the general root-finding problem (P). In contrast, it is tempting to suspect that on optimization problems, where $F' = \nabla^2 f$ is symmetric, the SR1 and the BFGS update will outperform Broyden's method. Surprisingly, we will observe that this is false for BFGS.

Note that Algorithm 1 does not include globalization techniques. Accordingly, the convergence results of [20] are of local nature. However, since the issue of globalization is of vital importance for the effectiveness of practical optimization methods, we investigate different globalization strategies in the numerical study of Section 5. One of them consists in integrating Algorithm 1 into a trust-region framework. Since we observe that this trust-region variant of the hybrid method performs extremely well including on a nonsmooth and nonconvex real-world application, it is our belief that this method could be of high interest to practitioners. This view is further encouraged by the fact that an earlier version of this method is an integral part of the code [30,31] that won the 2015/2016 ISMRM challenge [14] on optimal control in magnetic resonance imaging. We provide the precise algorithm in Section A. In the numerical experiments we compare the trust-region globalization to a damping of the steps s^k in Algorithm 1 by a monotone and a non-monotone line search.

2.2. Assumptions and convergence results

When discussing the application of the hybrid approach we will investigate under which conditions the assumptions that underlie its convergence theory are satisfied. To this end, we repeat these assumptions here. For simplicity we supply a slightly stronger version than the original one from [20]. The definitions of the required concepts, e.g., semismoothness and η-strict differentiability, are contained in [20, Section 2].

Assumption 1. *Suppose that*

- *there is $\bar{q} \in Q$ with $H(\bar{q}) = 0$;*
- *$G : Q \to U$ and $\hat{G} : Q \to V$ are semismooth at $\bar{q}$;*
- *there is $C_M > 0$ such that $\|M\|_{\mathcal{L}(Q,U)} \leq C_M$ for all $M \in \partial G(q)$ and all q close to $\bar{q}$;*
- *$F : U \to V$ is η-strictly differentiable at $\bar{u} := G(\bar{q})$;*

- *there is $C_{\bar{M}^{-1}} > 0$ such that, for all q close to $\bar{q}$, all $\bar{M} \in \partial H(q)$ are invertible with $\|\bar{M}^{-1}\|_{\mathcal{L}(V,Q)} \leq C_{\bar{M}^{-1}}$, where the generalized derivative $\partial H : Q \rightrightarrows \mathcal{L}(Q,V)$ is given by*

$$\partial H(q) := \left\{ F'(\bar{u}) \circ M + \hat{M} : \quad M \in \partial G(q),\ \hat{M} \in \partial \hat{G}(q) \right\}. \tag{1}$$

Remark 1. *If F is Fréchet differentiable with Hölder continuous derivative in a ball around $\bar{u}$, then it is also η-strictly differentiable at $\bar{u}$. Moreover, it is argued in [20] that under Assumption 1 there are constants $L_G, L_F > 0$ such that*

$$\|G(q) - G(\bar{q})\|_U \leq L_G \|q - \bar{q}\|_Q \qquad \text{and} \qquad \|F(u) - F(\bar{u})\|_V \leq L_F \|u - \bar{u}\|_U$$

are satisfied for all q close to $\bar{q}$, respectively, for all u close to $\bar{u}$. The constants L_G and L_F will appear in the convergence results for Algorithm 1.

Next we state summarized versions of the convergence results from [20, Section 4]. The first result addresses the convergence of (q^k).

Theorem 2.1. *Let Assumption 1 hold and let $\beta \in (0,1)$. Then:*

1) *There exist $\delta, \varepsilon > 0$ such that for every pair of starting values $(q^0, B_0) \in Q \times \mathcal{L}(U,V)$ with $\|q^0 - \bar{q}\|_Q < \delta$ and $\|B_0 - F'(\bar{u})\|_{\mathcal{L}(U,V)} < \varepsilon$, Algorithm 1 is well-defined and either terminates after finitely many iterations or generates a sequence of iterates (q^k) that converges q-linearly with rate β to $\bar{q}$. If, in addition, $\sigma_{\min}, \sigma_{\max} \in (0,2)$ in Algorithm 1 and $B_0 - F'(\bar{u})$ is compact, then the convergence is q-superlinear.*
2) *If, in addition to Assumption 1, F is Gâteaux differentiable in a neighborhood of $\bar{u}$ and the Gâteaux derivative is continuous at $\bar{u}$, then $\|B_0 - F'(\bar{u})\|_{\mathcal{L}(U,V)} < \varepsilon$ in 1) can be replaced by $\|B_0 - F'(u^0)\|_{\mathcal{L}(U,V)} < \varepsilon$. In particular, this replacement is possible if F is Hölder continuously Fréchet differentiable in a neighborhood of $\bar{u}$.*
3) *If, in addition to Assumption 1, $F'(u^0) - F'(\bar{u})$ is compact, then the compactness of $B_0 - F'(\bar{u})$ in 1) can be replaced by the compactness of $B_0 - F'(u^0)$. In particular, this replacement is possible if F is affine linear.*

For the convergence of (u^k), $(F(u^k))$ and $(H(q^k))$ we draw the following conclusion.

Corollary 2.2. *Let Assumption 1 hold and let (q^k) be generated by Algorithm 1. If (q^k) converges q-linearly (q-superlinearly) to $\bar{q}$, then:*

1) *(u^k) converges r-linearly (r-superlinearly) to $\bar{u}$ and satisfies $\|u^k - \bar{u}\|_U \leq L_G \|q^k - \bar{q}\|_Q$ for all k sufficiently large.*
2) *$(F(u^k))$ converges r-linearly (r-superlinearly) to $F(\bar{u})$ and satisfies, for all k large enough, $\|F(u^k) - F(\bar{u})\|_V \leq L_F \|u^k - \bar{u}\|_U$ and $\|F(u^k) - F(\bar{u})\|_V \leq L_F L_G \|q^k - \bar{q}\|_Q$.*
3) *If there is $L_{\hat{G}} > 0$ such that $\|\hat{G}(q^k) - \hat{G}(\bar{q})\|_V \leq L_{\hat{G}} \|q^k - \bar{q}\|_Q$ for all k sufficiently large, then $(H(q^k))$ converges r-linearly (q-superlinearly) to zero.*

Remark 2. *We point out that the above convergence results do not require the Banach spaces Q and V to be reflexive. In Section 3.3 we will encounter optimization problems where this is beneficial.*

3. Applications

In this section we show how the results of Section 2 can be applied to generalized variational inequalities and nonsmooth optimization problems, particularly PDE-constrained optimal control problems. Here, an important tool are proximal mappings. Introduced and studied by Moreau, cf., e.g., [22], they continue to attract significant research interest until today. They are, for instance, used frequently in convex composite optimization, where many algorithms are built around them. We introduce these mappings and provide selected references below, but let us mention already here that [4, Section 6] and [3, Section 24] are good starting points for a finite-dimensional, respectively, infinite-dimensional treatment of proximal mappings. Also, this section requires some elementary facts from convex analysis, which are, e.g., contained in [11, Chapter 1–2].

Throughout Section 3 we will assume that

- U is a Hilbert space with scalar product $(\cdot,\cdot)_U$;
- $\varphi : U \to (-\infty, +\infty]$ is a proper, convex and lower semicontinuous function.

We write $\operatorname{dom}(\varphi) := \{u \in U : \varphi(u) < \infty\}$ for the essential domain of φ. Moreover, we introduce for $c > 0$ the notation $\varphi_c := \frac{\varphi}{c}$ and define the *proximal mapping of* φ_c by

$$\operatorname{Prox}_{\varphi_c} : U \to \operatorname{dom}(\varphi), \quad \operatorname{Prox}_{\varphi_c}(u) := \operatorname{argmin}_{\tilde{u}\in\operatorname{dom}(\varphi)} \left[\frac{1}{2} \|\tilde{u} - u\|_U^2 + \varphi_c(\tilde{u}) \right]. \tag{2}$$

It follows from standard arguments of convex analysis that $\operatorname{Prox}_{\varphi_c}$ is well-defined and that

$$u^* = \operatorname{Prox}_{\varphi_c}(u) \qquad \Longleftrightarrow \qquad c(u - u^*) \in \partial\varphi(u^*) \tag{3}$$

for all $u \in U$ and arbitrary $c > 0$.

3.1. Generalized variational inequalities

Generalized variational inequalities (GVIs) are a very broad framework. They comprise, for instance, complementarity problems, contact problems, certain Nash games and other equilibrium problems, saddle point problems and nonsmooth optimization problems, in particular from convex composite optimization. For a detailed treatment of finite-dimensional GVIs we refer to the survey paper [16] and the monographs [12, 13]. For a compact and current introduction to this subject we also point the reader to [21, Section 6], where the class of GVIs that we consider here is treated in finite dimensions in more detail and many references with further material are provided. Variational inequalities in infinite-dimensional spaces are, for instance, discussed in [19], [37, Section 5] and [24].

Our focus in this section is to work out how the hybrid method can be applied to GVIs. In particular, we will not discuss existence of solutions of GVIs. Instead, we will simply assume that a solution exists. The existence of solutions to GVIs is addressed in all of the previously mentioned references. Moreover, we will assume for simplicity that all mappings are defined on entire spaces, although it would be sufficient to have them defined only locally around solutions.

Concerning the Hilbert space U we mention that for the following considerations typical spaces would be $U = H_0^1(\Omega)$ or $U = \mathbb{R}^n$. To be consistent with the notation introduced in Section 2 we set $Q := V := U$. Let $T : U \to U$ be bijective and set $R := T^{-1}$. For a given mapping $\hat{F} : U \to V$ we consider the following generalized variational inequality that is also addressed in [33]:

Find $\bar{u} \in R(\mathrm{dom}(\varphi))$ such that

$$\text{(GVI)} \qquad \left(\hat{F}(\bar{u}), u - T(\bar{u})\right)_U + \varphi(u) - \varphi(T(\bar{u})) \geq 0 \quad \forall u \in U.$$

We note that for $T = \mathrm{id}$, (GVI) reduces to the important class of variational inequalities of the second kind. If, in addition, φ is the indicator function of a closed convex set, then (GVI) reproduces the classical variational inequality. We recall that Algorithm 1 is applied to problems of the form (P).

Lemma 3.1. *Problem* (GVI) *can be expressed in the form* (P).

Proof. By definition of the convex subdifferential, (GVI) is equivalent to $-\hat{F}(\bar{u}) \in \partial\varphi(T(\bar{u}))$. Denoting $\hat{F}_c := \frac{\hat{F}}{c}$ for $c > 0$ we deduce from (3) that $-\hat{F}(\bar{u}) \in \partial\varphi(T(\bar{u}))$ is satisfied if and only if $T(\bar{u}) = \mathrm{Prox}_{\varphi_c}(T(\bar{u}) - \hat{F}_c(\bar{u}))$. We rewrite this equation as

$$T(\bar{u}) = \mathrm{Prox}_{\varphi_c}(u) \quad \wedge \quad u = T(\bar{u}) - \hat{F}_c(\bar{u}).$$

Instead of looking for a solution $\bar{u} \in R(\mathrm{dom}(\varphi))$ of (GVI) we can thus search for $u \in U$ with

$$u = \mathrm{Prox}_{\varphi_c}(u) - \hat{F}_c(R(\mathrm{Prox}_{\varphi_c}(u)))$$

and, once u is found, obtain $\bar{u}$ through $\bar{u} := R(\mathrm{Prox}_{\varphi_c}(u))$. Since $Q = U$ we can summarize this as

$$\text{(4)} \qquad \bar{u} \in R(\mathrm{dom}(\varphi)) \subset U \text{ satisfies (GVI)} \;\Leftrightarrow\; \exists \bar{q} \in Q : \; \left[H(\bar{q}) = 0 \;\wedge\; \bar{u} = R(\mathrm{Prox}_{\varphi_c}(\bar{q}))\right],$$

where we used the mapping

$$H : Q \to V, \qquad H(q) := \hat{F}(R(\mathrm{Prox}_{\varphi_c}(q))) + c\left(q - \mathrm{Prox}_{\varphi_c}(q)\right)$$

for an arbitrary $c > 0$ (recalling that $Q = U = V$). Defining

$$F : U \to V, \quad F(u) := \hat{F}(R(u)), \qquad\qquad G : Q \to U, \quad G(q) := \mathrm{Prox}_{\varphi_c}(q),$$

and

$$\hat{G} : Q \to V, \quad \hat{G}(q) := c\left(q - \mathrm{Prox}_{\varphi_c}(q)\right)$$

we have therefore established the assertion. □

We mention that instead of including R in the definition of F, we can include R in the definition of G if R is nonsmooth. Moreover, let us point out the important special case that φ is the indicator function of a closed convex set C, i.e., $\varphi(u) = 0$ for $u \in C$ and $\varphi(u) = +\infty$ else. In this case, the definition of the proximal mapping yields that $\mathrm{Prox}_{\varphi_c}$ is the orthogonal projection onto C, denoted Π_C. Thus, H takes the form

$$H(q) = \hat{F}(R(\Pi_C(q))) + c\left(q - \Pi_C(q)\right).$$

For $R = \mathrm{id}$ and $c = 1$ such mappings were investigated by Robinson under the name *normal maps*, cf. [29].

To apply the main convergence result Theorem 2.1, we need to ensure that Assumption 1 is satisfied. In the present setting this assumption reads as follows:

1) There is $\bar{q} \in Q$ with $H(\bar{q}) = 0$, i.e., (GVI) has a solution;
2) $G = \mathrm{Prox}_{\varphi_c}$ is semismooth at $\bar{q}$;

3) there is $C_M > 0$ such that $\|M\|_{\mathcal{L}(Q,U)} \leq C_M$ for all $M \in \partial G(q)$ and all q close to $\bar{q}$;

4) F is η-strictly differentiable at $\bar{u} := G(\bar{q})$;

5) there is $C_{\bar{M}^{-1}} > 0$ such that, for all q close to $\bar{q}$, all $\bar{M} \in \partial H(q)$ are invertible with $\|\bar{M}^{-1}\|_{\mathcal{L}(V,Q)} \leq C_{\bar{M}^{-1}}$, where the generalized derivative $\partial H : Q \rightrightarrows \mathcal{L}(Q,V)$ is given in (1).

Here, we have used that the semismoothness requirement imposed on $\hat{G}$ in Assumption 1 follows from the semismoothness of G because of $\hat{G}(q) = c(q - G(q))$.

We emphasize that each of the requirements 1)–5) can be violated for (GVI). In particular, 1), 2) and 5) are delicate issues that have to be investigated individually for given problems. Let us, however, stress that these requirements are, indeed, satisfied for many problems of practical interest. Exemplarily, we will discuss 1)–5) for structured convex optimization problems, cf. Section 3.2. Furthermore, we point out that 5) is satisfied in finite dimensions if H is *BD-regular* or *CD-regular* at $\bar{q}$.

We now make some general remarks concerning 2) and 3). To begin with, let us mention that the notion of semismoothness that we use, cf. [20, Definition 2.1], is less demanding than many common notions of semismoothness, particularly those used in finite dimensions, e.g. the widely employed concept of semismoothness by Qi and Sun in [28, bottom of p. 355]. Therefore, 2) is, in particular, fulfilled if the operator $\mathrm{Prox}_{\varphi_c}$ is semismooth in any of the usual senses. For the finite-dimensional case, [38, Section 3] and [21, Section 3.3] discuss the semismoothness of several proximal maps and provide further references. The semismoothness of prox operators in infinite dimensions appears, e.g., in [27, Section 3.3], and Section 3.2 contains some examples, too. Moreover, since proximal maps are Lipschitz continuous with constant 1, cf. [3, Proposition 12.28], it can be expected that (due to $Q = U$) we have $\|\partial G\|_{\mathcal{L}(Q,U)} \leq 1$ globally and, consequently, 3) will often be satisfied, both in finite and infinite dimensions. For instance, if we use for ∂G Clarke's generalized Jacobian or—in the case of superposition operators—Ulbrich's generalized differential for superposition operators [37, Section 3], then this is true. Thus, the applicability of the hybrid method will in many cases reduce to the question if 1), 2) and 5) are satisfied.

As a final comment we point out that proximal mappings and/or their generalized derivatives can not always be evaluated efficiently. In these cases the application of the hybrid method is not possible, at least not in the manner just presented.

3.2. Optimization

In this section we discuss how the hybrid method can be applied to a class of structured optimization problems. We then consider the subclass of Hilbert space regularized structured problems. A particularly convenient feature of this subclass is that the question of uniform invertibility of ∂H can be treated in a general fashion.

3.2.1. Structured problems

We consider optimization problems having the form

$$\min_{u \in U} f(u) + \varphi(u), \tag{PO}$$

where $f : U \to \mathbb{R}$ is continuously Fréchet differentiable and U and φ satisfy the general assumptions listed at the beginning of Section 3. Note that (PO) is, in general, a *constrained*, *nonsmooth* and *nonconvex* optimization problem. By assuming additionally that f is convex we obtain the subclass of *structured convex optimization problems* that has received great attention in recent years. Structured convex optimization problems appear in statistics, finance, machine

learning and image reconstruction, for instance in the form of ℓ^1/L^1-regularized optimization problems such as the Lasso problem [35]. Besides being of high practical relevance they have the appealing theoretical property that their first order optimality conditions are not only necessary but also sufficient for optimality and, moreover, that every $\bar{u}$ which satisfies these conditions is a *global* minimizer. In the first half of Section 5 we investigate the numerical performance of the hybrid approach for L^1-regularized optimal control problems, which are large-scale structured convex optimization problems. We will find that Algorithm 1 performs very well on this class if augmented by a simple line search technique. In the following we will, however, not assume convexity of f. The relationship between (PO) and the mother problem (P) is as follows.

Lemma 3.2. *Consider* (PO) *and set* $Q := V := U$. *Then* $\bar{u} \in U$ *satisfies the first order optimality conditions of* (PO) *if and only if there exists* $\bar{q} \in Q$ *such that* $H(\bar{q}) = 0$ *for*

$$H : Q \to V, \qquad H(q) := \nabla f(\mathrm{Prox}_{\varphi_c}(q)) + c\left(q - \mathrm{Prox}_{\varphi_c}(q)\right), \tag{5}$$

where $c > 0$ *is arbitrary. If either* $\bar{u}$ *or* $\bar{q}$ *exists, then so does the other and there holds* $\bar{u} = \mathrm{Prox}_{\varphi_c}(\bar{q})$.

Moreover, if $f + \varphi$ *is additionally convex, then* $\bar{u} \in U$ *satisfies the first order optimality conditions of* (PO) *if and only if it solves* (PO).

Proof. The first order optimality conditions of (PO) read

$$(\nabla f(\bar{u}), u - \bar{u})_U + \varphi(u) - \varphi(\bar{u}) \geq 0 \quad \forall u \in U. \tag{6}$$

Setting $T := R := \mathrm{id}$ and $\hat{F} := \nabla f$ we note that this is (GVI). Hence, by (4) we can reformulate (6) equivalently as $H(\bar{q}) = 0$ with H as given in (5) and $c > 0$ arbitrary. Furthermore, (4) states that $\bar{u}$ exists if and only if $\bar{q}$ does, and that $\bar{u} = \mathrm{Prox}_{\varphi_c}(\bar{q})$, then. The final claim is a well-known fact from convex analysis. □

Remark 3. *Evidently, the mapping* H *given by* (5) *has the form required by the hybrid method:* $H = F \circ G + \hat{G}$ *for* $F(u) := \nabla f$, $G(q) := \mathrm{Prox}_{\varphi_c}(q)$ *and* $\hat{G}(q) := c(q - \mathrm{Prox}_{\varphi_c}(q))$.

Lemma 3.2 shows that the application of the hybrid method to (PO) involves the proximal mapping of φ, $\mathrm{Prox}_{\varphi_c}$. Very popular first order methods for (PO) are the so-called *proximal methods*, cf., e.g., [4, 25]. Well-known algorithms such as the proximal gradient method, the alternating direction method of multipliers (ADMM/Douglas-Rachford splitting) and FISTA all belong to this class, cf. [25, Section 4] for the two former and [5] for the latter. Generally speaking, these methods are all built around the proximal mapping $\mathrm{Prox}_{\varphi_c}$. They are most useful in situations where the proximal map can be evaluated efficiently, and many examples are known where this is actually the case. Similarly, solving $H(q) = 0$ with Algorithm 1 requires the evaluation of $\mathrm{Prox}_{\varphi_c}(q)$ and, in addition, a computable generalized derivative of $\mathrm{Prox}_{\varphi_c}$ such that $\mathrm{Prox}_{\varphi_c}$ is semismooth with respect to this derivative. It turns out that for many problems, for which the proximal map can be evaluated efficiently, it is also possible to compute an appropriate generalized derivative. Some examples are contained in the following sections; more examples are addressed, e.g., in [38, Section 3] and [21, Section 3.3] for finite dimensions as well as [27, Section 3.3] for infinite dimensions.

Using the definitions from Remark 3 we can write down Assumption 1 for (PO). This results in a specialization of 1)–5) from Section 3.1 to the case at hand and we therefore omit it. Let us, however, mention that the property 5) concerning ∂H can be difficult to establish, particularly in infinite-dimensional settings, where it is often investigated anew for each new problem. In contrast, we now consider a subclass of (PO) for which this issue can be dealt with in a general fashion, using an approach recently presented in [27, Section 3]. This approach also plays a role in some of the numerical experiments in Section 5.

3.2.2. Regularized structured problems

Following [27, Section 3] we consider (PO) in a Hilbert space U for an objective f that contains a regularization. That is, we are still interested in

$$\min_{u \in U} f(u) + \varphi(u) \tag{POR}$$

from Section 3.2.1, but from now on under the additional assumption that f is of the form

$$f(u) = \hat{f}(u) + \frac{\gamma}{2} \|u\|_U^2 , \tag{7}$$

where $\gamma > 0$ and $\hat{f} : U \to \mathbb{R}$ is continuously Fréchet differentiable (not necessarily convex). Recalling that in (5) we have worked with $Q = U = V$, we can choose $c := \gamma$ in (5) and conclude that $H : U \to U$ simplifies to

$$H(q) = \nabla \hat{f}(\mathrm{Prox}_{\varphi_\gamma}(q)) + \gamma q, \tag{8}$$

where $\varphi : U \to (-\infty, +\infty]$ is assumed to be proper, convex and lower semicontinuous, and $\mathrm{Prox}_{\varphi_\gamma}$ is defined according to (2). From Lemma 3.2 we obtain that $\bar{u} \in U$ satisfies the first order optimality conditions of (POR), respectively, solves (POR), if and only if there is $\bar{q} \in U$ with $H(\bar{q}) = 0$, and they adhere to the relation $\bar{u} = \mathrm{Prox}_{\varphi_\gamma}(\bar{q})$.

We now discuss a setting in which it is beneficial to choose Q and V different from U. To this end, let us assume in the following that $\nabla \hat{f}(u)$ satisfies

$$\nabla \hat{f}(U) \subset V \tag{9}$$

for a Banach space V that is continuously embedded in U. In fact, since we allow for the choice $V = U$ this assumption is possible without loss of generality, and we underline that if U is finite-dimensional, then it suffices to have $Q = V = U$ in the following considerations. To provide an infinite-dimensional example, we mention that a typical scenario in the context of elliptic partial differential equations is $U = L^2(\Omega)$ and $V = H_0^1(\Omega)$. In that case, (9) describes that $\nabla \hat{f}$ has a *smoothing property*.

Due to (9) the optimality condition $H(\bar{q}) = 0$, rewritten as $\bar{q} = -\frac{1}{\gamma} \nabla \hat{f}(\mathrm{Prox}_{\varphi_\gamma}(\bar{q}))$, implies $\bar{q} \in V$. Lemma 3.2 thus implies the following result.

Lemma 3.3. *Consider (POR) with f satisfying (7). Let $V \hookrightarrow U$ be a Banach space such that (9) holds and set $Q := V$. Then $\bar{u} \in U$ satisfies the first order optimality conditions of (POR) if and only if there is $\bar{q} \in Q$ that satisfies $H(\bar{q}) = 0$ for*

$$H : Q \to V, \qquad H(q) := \nabla \hat{f}(\mathrm{Prox}_{\varphi_\gamma}(q)) + \gamma q. \tag{10}$$

If either $\bar{u}$ or $\bar{q}$ exists, then so does the other and there holds $\bar{u} = \mathrm{Prox}_{\varphi_\gamma}(\bar{q})$.

Moreover, if $f + \varphi$ is additionally convex, then $\bar{u}$ satisfies the first order optimality conditions of (POR) if and only if it solves (POR).

Proof. There is nothing left to prove. □

Remark 4. *The mapping H given by (10) still has the form required by the hybrid method, although the components are different compared to (PO): $H = F \circ G + \hat{G}$ for $F(u) := \nabla \hat{f}(u)$, $G(q) := \mathrm{Prox}_{\varphi_\gamma}(q)$ and $\hat{G}(q) := \gamma q$.*

Lemma 3.3 shows that solving $H(q) = 0$ allows to tackle (POR). In view of the convergence result Theorem 2.1 the next step is to ensure that Assumption 1 is fulfilled for (POR). This will be the case if

1) there is $\bar{u} \in U$ that satisfies the first order optimality conditions of (POR);
2) $\mathrm{Prox}_{\varphi_\gamma} : Q \to U$ is semismooth at $\bar{q} := -\frac{1}{\gamma}\nabla \hat{f}(\bar{u})$;
3) there is $C_M > 0$ such that $\|M\|_{\mathcal{L}(Q,U)} \leq C_M$ for all $M \in \partial \,\mathrm{Prox}_{\varphi_\gamma}(q)$ and all q close to $\bar{q}$;
4) $\nabla \hat{f} : U \to V$ is η-strictly differentiable at $\bar{u}$;
5) there is $C_{\bar{M}^{-1}} > 0$ such that, for all q close to $\bar{q}$, all $\bar{M} \in \partial H(q)$ are invertible with $\|\bar{M}^{-1}\|_{\mathcal{L}(V,Q)} \leq C_{\bar{M}^{-1}}$, where the generalized derivative $\partial H : Q \rightrightarrows \mathcal{L}(Q,V)$ reads

$$\partial H(q) = \left\{ \nabla^2 \hat{f}(\mathrm{Prox}_{\varphi_\gamma}(\bar{q}))M + \gamma I : \; M \in \partial \,\mathrm{Prox}_{\varphi_\gamma}(q) \right\}. \tag{11}$$

For 2) we recall that in Section 3.1 we have provided references that are concerned with the question of semismoothness of proximal mappings. There, we have also motivated why 3) usually holds with $C_M = 1$ if $Q = U$. Here, we have $Q = V \hookrightarrow U$, hence we can still expect that 3) holds. Moreover, we point out that 4) is true if $\nabla \hat{f}$ is Fréchet differentiable with Hölder continuous derivative in a neighborhood of $\bar{u}$. Regarding 5) we are aware of two general approaches. The first one is based on operator perturbation theory, more precisely on the fact that the particular structure of ∂H often allows to show that $\cup_{q \text{ near } \bar{q}} \partial H(q)$ is a *collectively compact perturbation of the identity*. For the notion of collective compactness we refer to [1]. From this approach it follows that 5) holds, in particular, if $\nabla^2 \hat{f}(\mathrm{Prox}_{\varphi_\gamma}(\bar{q})) \in \mathcal{L}(U,V)$ is compact and the elements of $\partial H(q)$ are injective for all q near $\bar{q}$. The second approach also exploits the structure of the elements of ∂H, but does not require operator compactness. Instead, the main ingredient is a convexity-type assumption. The idea is that if $\hat{f}$ is convex, then the elements of $\partial H(q)$ are roughly of the form $\nabla^2 \hat{f}(\bar{u}) + \gamma I$ with $\nabla^2 \hat{f}$ positive semidefinite. This implies their invertibility and a norm bound of order $\frac{1}{\gamma}$ for the inverses. However, in this argument we have neglected the presence of the operator $M \in \partial \,\mathrm{Prox}_{\varphi_\gamma}(q)$ in the definition of $\partial H(q)$, and this complicates matters as it implies, for instance, that the elements of $\partial H(q)$ are not self-adjoint, in general. Fortunately, this obstacle can be overcome by the approach presented in [27, Section 3].

Lemma 3.4. *In the situation of Lemma 3.3 and with ∂H given by (11), condition 5) of Assumption 1 is satisfied if there exists $\nu > 0$ such that for all q near $\bar{q}$ and for all $M \in \partial \,\mathrm{Prox}_{\varphi_\gamma}(q)$ we have*

- $(Mh_1, h_1)_U \geq 0$ *for all* $h_1 \in Q$,
- $(Mh_1, h_2)_U = (h_1, Mh_2)_U$ *for all* $h_1, h_2 \in Q$,
- *and*

$$\gamma\,(h_1, Mh_1)_U + \left(\nabla^2 \hat{f}(\mathrm{Prox}_{\varphi_\gamma}(\bar{q}))Mh_1, Mh_1\right)_U \geq \nu\,(h_1, Mh_1)_U \qquad \text{for all } h_1 \in Q. \tag{12}$$

Proof. This is [27, Lemma 3.15] for the situation at hand. □

Remark 5. *The inequality (12) is, in particular, satisfied if $\nabla^2 \hat{f}(\bar{u})$ is positive semidefinite. Moreover, a bound for the norm of the inverse operators can be found in [27, (3.15)].*

To round off the discussion of Assumption 1 for (POR) and to explain why the use of different spaces Q and U is necessary in infinite-dimensional settings, let us provide an example. For simplicity we only treat the map $\mathrm{Prox}_{\varphi_\gamma}$ and do not fix the function $\hat{f}$.

Example 1. *Let $\Omega \subset \mathbb{R}^d$ be a nonempty and bounded domain, $Q := V := L^\infty(\Omega)$ and $U := L^2(\Omega)$ (both endowed with with their usual norms), $U_{\mathrm{ad}} := \{u \in L^2(\Omega) : a \le u \le b \text{ a.e. in } \Omega\}$ with real numbers $a < b$, and $\varphi(u) = \delta_{U_{\mathrm{ad}}}(u)$, i.e., $\varphi(u) = 0$ iff $u(x) \in [a,b]$ a.e. in Ω and $\varphi(u) = +\infty$ otherwise. The proximal map $\mathrm{Prox}_{\varphi_c} : U \to \mathrm{dom}(\varphi)$ is the L^2-projection onto U_{ad}, regardless of the value of c. We write $\Pi_{U_{\mathrm{ad}}} : U \to U$ for this projection. It is well-known that $\Pi_{U_{\mathrm{ad}}}$ is not semismooth at any $u \in U \setminus U_{\mathrm{ad}}$ when considered as an operator from $(U, \|\cdot\|_U)$ to $(U, \|\cdot\|_U)$, but semismooth at all $q \in Q$ when considered from $(Q, \|\cdot\|_Q)$ to $(U, \|\cdot\|_U)$. The underlying principle is that semismoothness of nonlinear superposition operators from $L^p(\Omega)$ to $L^q(\Omega)$ requires $p > q$ (except for $p = q = \infty$, to be precise); cf. [32, 37]. Apparently, this situation cannot be captured by working with $Q = U$.*

To define a generalized derivative $\partial G(q) \subset \mathcal{L}(Q,U)$ for the operator $G := \Pi_{U_{\mathrm{ad}}} : Q \to U$ we denote by $\partial^{Cl}\psi$ the Clarke subdifferential of

$$\psi : \mathbb{R} \to \mathbb{R}, \qquad \psi(t) := \max(a, \min(b,t)). \tag{13}$$

That is,

$$\partial^{Cl}\psi(t) = \begin{cases} \{1\} & \text{if } t \in (a,b), \\ \{0\} & \text{if } t \notin [a,b], \\ [0,1] & \text{if } t = a \text{ or } t = b. \end{cases} \tag{14}$$

Observing that $G(q)(x) = \psi(q(x))$ for a.e. $x \in \Omega$ we can take as generalized derivative the set

$$\partial G(q) = \bigcup_{\substack{r \in L^\infty(\Omega) \text{ with} \\ r(x) \in \partial^{Cl}\psi(q(x)) \text{ for a.e. } x \in \Omega}} \Big\{ M : \ Mh = rh \ \ \forall h \in Q \Big\}, \tag{15}$$

where rh is the pointwise multiplication of the functions r and h. The mapping $G : Q \to U$ is semismooth at every $q \in Q$ with respect to ∂G, cf., e.g., [17, Theorem 2.13]. That is, 2) is satisfied. Moreover, 3) is true because, as is readily checked, $\|M\|_{\mathcal{L}(Q,U)} \le |\Omega|^{1/2}$ for all $M \in \partial G(q)$ and all $q \in Q$. For 5) we easily confirm that $(Mh_1, h_1)_U \ge 0$, and $(Mh_1, h_2)_U = (h_1, Mh_2)_U$ are satisfied. Thus, 5) holds, for instance, if $\nabla^2 \hat{f}$ is positive semidefinite at $\bar{u} = \Pi_{U_{\mathrm{ad}}}(\bar{q})$.

To conclude this example let us point out that the demand for $\nabla \hat{f}$ to map $U = L^2(\Omega)$ to $V = L^\infty(\Omega)$ is a smoothing property. In the following section we discuss concrete examples for (POR) from PDE-constrained optimal control problem. This includes demonstrating that a smoothing property is fulfilled.

3.3. PDE-constrained optimal control

The aim of this section is to work out that PDE-constrained optimal control problems are well-suited for the application of the hybrid approach. In doing so we encounter several situations in which the use of non-reflexive Banach spaces $Q = V$ is beneficial, cf. Remark 2.

3.3.1. Distributed control of a semilinear elliptic equation

Let $Q := V := Y := P := H_0^1(\Omega)$, $U := L^2(\Omega)$, and $\gamma > 0$. We consider the nonconvex problem

$$\min_{(y,u) \in Y \times U_{\mathrm{ad}}} \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\gamma}{2}\|u\|^2_U \qquad \text{s.t.} \qquad \begin{cases} -\Delta y + y^3 = gu & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega, \end{cases} \tag{OCE}$$

where $\Omega \subset \mathbb{R}^d$, $1 \le d \le 3$, is a nonempty and bounded Lipschitz domain, $y_d \in L^2(\Omega)$, $g \in W^{1,\infty}(\Omega)$ ($g \in L^\infty(\Omega)$ is also possible by working with other spaces $Q = V$ than the ones that we use here), and

$$U_{\mathrm{ad}} := \{u \in U : \ a \le u \le b \ \text{ a.e. in } \Omega\}, \tag{16}$$

where $a \le b$ are fixed constants. The $L^2(\Omega)$-projection onto U_{ad} is denoted by $\Pi_{U_{\mathrm{ad}}}$.

We will now apply Lemma 3.3 to infer that the first order optimality conditions of (OCE) can be expressed in the form $H(q) = 0$. To begin with, we remark that the PDE has for every $u \in U$ a unique solution $y = y(u) \in Y$ and the solution operator $u \mapsto y(u)$ is three times continuously Fréchet differentiable from U to Y. This can be argued almost verbatim as in [37, Section 9]. Therefore, (OCE) can be reformulated equivalently as

$$\min_{u \in U} \frac{1}{2}\|y(u) - y_d\|^2_{L^2(\Omega)} + \frac{\gamma}{2}\|u\|^2_U + \delta_{U_{\mathrm{ad}}}(u), \tag{OCEr}$$

where $\delta_{U_{\mathrm{ad}}}(u) = 0$ for $u \in U_{\mathrm{ad}}$ and $\delta_{U_{\mathrm{ad}}}(u) = +\infty$ otherwise. Setting $\hat{f} : U \to \mathbb{R}$, $\hat{f}(u) := \frac{1}{2}\|y(u) - y_d\|^2_{L^2(\Omega)}$ and $\varphi : U \to (-\infty, +\infty]$, $\varphi(u) := \delta_{U_{\mathrm{ad}}}(u)$, problem (OCEr) has the form (POR). This allows us to conclude that aiming for the first order optimality conditions of (OCE) is the same as solving the operator equation $H(q) = 0$. Specifically, we obtain the following result.

Lemma 3.5. *There is at least one $\bar{q} \in Q$ with $H(\bar{q}) = 0$ for*

$$H : Q \to V, \qquad H(q) = g \cdot p(\Pi_{U_{\mathrm{ad}}}(q)) + \gamma q. \tag{17}$$

Here, $p = p(u) \in P$ is the adjoint state*, i.e., the unique solution of the* adjoint equation

$$\begin{cases} -\Delta p + 3y(u)^2 p = y(u) - y_d & \text{in } \Omega, \\ p = 0 & \text{on } \partial\Omega. \end{cases} \tag{18}$$

Setting $\bar{u} := \Pi_{U_{\mathrm{ad}}}(\bar{q})$ and $\bar{y} := y(\bar{u})$ for a given $\bar{q} \in Q$ with $H(\bar{q}) = 0$ yields a pair $(\bar{y}, \bar{u}) \in Y \times U_{\mathrm{ad}}$ that satisfies the first order optimality conditions of (OCE). In addition, every pair $(\bar{y}, \bar{u}) \in Y \times U_{\mathrm{ad}}$ that satisfies the first order optimality conditions of (OCE) can be obtained in this way.

Proof. It follows from the definition that $\mathrm{Prox}_{\varphi_\gamma} = \Pi_{U_{\mathrm{ad}}}$. Since $\nabla \hat{f}(u) = y'(u)^*(y(u) - y_d)$, Lemma 3.3 yields the claim if we can show that (OCE) has at least one solution and that $y'(u)^*(y(u) - y_d) = gp(u) \in V = H^1_0(\Omega)$. Since $gp(u) \in H^1_0(\Omega)$ follows if $p(u) \in H^1_0(\Omega)$ holds, all remaining claims are standard results, cf., e.g., [17, Chapter 1] or [37, Chapter 9]. □

Remark 6. *The optimality condition $H(\bar{q}) = 0$ of Lemma 3.5 implies higher regularity for $\bar{y}$, $p(\bar{u})$, $\bar{u}$ and $\bar{q}$. In fact, since the projection $\Pi_{U_{\mathrm{ad}}}$ can be represented by $\Pi_{U_{\mathrm{ad}}}(q)(x) = \psi(q(x))$ for a.e. $x \in \Omega$, with ψ given by (13), we obtain from [19, Chapter II, Corollary A.5] that $\Pi_{U_{\mathrm{ad}}}(W^{1,s}(\Omega)) \subset W^{1,s}(\Omega)$ for all $s \in [1, \infty]$. As $\bar{u} = \Pi_{U_{\mathrm{ad}}}(\bar{q})$, this yields $\bar{u} \in H^1(\Omega)$. In addition, it shows that if $\bar{q}$ admits $W^{1,s}$-regularity, then $\bar{u}$ will also exhibit this regularity. Since $\bar{q}$ satisfies $\bar{q} = -\frac{1}{\gamma} g \cdot p(\bar{u})$, $\bar{q}$ has the same $W^{1,s}$-regularity as $\bar{p} := p(\bar{u})$. The regularity of $\bar{p}$ is dictated by the adjoint equation (18) and is, under the current assumptions, at least $W^{1,s}$ for some $s > 2$. This follows by results of [15] and can be carried out in detail almost verbatim as in [9, Proposition 2.2]. In turn, this implies that $\bar{u} \in W^{1,s}(\Omega)$ for some $s \in (2, \infty]$. If y_d or Ω are assumed to be more regular, then $\bar{p}$ satisfies $\bar{p} \in W^{1,s}(\Omega)$ with an $s > 2$ that depends on the dimension d and the regularity of the data (y_d, Ω). This can give as much as $W^{1,\infty}$-regularity for $\bar{q}$ and $\bar{u}$.*

The convergence theory of the hybrid method is based on Assumption 1. This assumption is guaranteed to hold for (OCE) under the following sufficient condition.

Lemma 3.6. *Let $(\bar{y}, \bar{u}) \in Y \times U_{\mathrm{ad}}$ satisfy the first order optimality conditions of (OCE), let H be given by (17), and denote by $\bar{q} \in Q$ the root of H that satisfies $\bar{u} = \Pi_{U_{\mathrm{ad}}}(\bar{q})$. Then Assumption 1 is fulfilled provided there holds*

$$6\bar{y}\bar{p} \leq 1 \quad \text{a.e. in } \Omega, \tag{19}$$

where $\bar{p} := p(\bar{u})$ is the solution of the adjoint equation (18) for $u = \bar{u}$.

Proof. Since (OCEr) is of the form (POR), it suffices to verify 1)–5) as stated below Remark 4. Obviously, 1) is true. As we have already observed, it holds that $\nabla \hat{f}(u) = y'(u)^*(y(u) - y_d) = gp(u)$, where $p(u)$ solves (18). Furthermore, since $u \mapsto y(u) \in Y = V$ is thrice continuously differentiable, it follows that $\nabla \hat{f} : U \to V$ is twice continuously Fréchet differentiable. Thus, it is locally Lipschitz continuously differentiable for all $u \in U$ and, in particular, η-strictly differentiable at every $u \in U$. That is, 4) holds. The projection $\Pi_{U_{\mathrm{ad}}}$ can be represented by $\Pi_{U_{\mathrm{ad}}}(q)(x) = \psi(q(x))$ for a.e. $x \in \Omega$, with ψ from (13). The superposition operator $q \mapsto \Pi_{U_{\mathrm{ad}}}(q)$ is semismooth from $L^p(\Omega)$ to $L^2(\Omega) = U$ for any $p > 2$ (cf. [32, 37]), hence semismooth from Q to U at every $q \in Q$. Therefore, 2) holds. Moreover, the generalized derivative $\partial \Pi_{U_{\mathrm{ad}}}$, defined by (15), satisfies 3) and the first two conditions listed in Lemma 3.4, cf. the discussion below (15). Finally, since $(\nabla^2 \hat{f}(\bar{u})h, h)_U = (y'(\bar{u})h, (1 - 6\bar{y}\bar{p})y'(\bar{u})h)_U$ for all $h \in U$, (19) evidently yields (12) for $\nu := \gamma$. Thus, Lemma 3.4 implies that 5) holds. □

Remark 7. *The sufficient condition (19) is, for instance, satisfied if $\bar{p}$ is sufficiently small. In view of (18) this will be the case if $\bar{y}$ matches y_d well enough. Notice also that (19) is compatible with the homogeneous Dirichlet boundary conditions imposed on $\bar{p}$ and $\bar{y}$.*

Under Assumption 1 we obtain convergence results and various error bounds. We recall that $Q = Y = P = V = H_0^1(\Omega)$, $U = L^2(\Omega)$, and that U_{ad} is given by (16). Moreover, we denote $y^k := y(u^k)$ and $p^k := p(u^k)$ for all $k \geq 0$.

Theorem 3.7. *1) Let $(\bar{y}, \bar{u}) \in Y \times U_{\mathrm{ad}}$ satisfy the first order optimality conditions of (OCE), let H be given by (17), and denote by $\bar{q} \in Q$ the root of H that satisfies $\bar{u} = \Pi_{U_{\mathrm{ad}}}(\bar{q})$. Moreover, let Assumption 1 hold and let $\beta \in (0, 1)$. Then there exist $\delta, \varepsilon > 0$ such that for every pair of starting values $(q^0, B_0) \in Q \times \mathcal{L}(U, V)$ with $\|q^0 - \bar{q}\|_Q < \delta$ and $\|B_0 - \nabla^2 \hat{f}(u^0)\|_{\mathcal{L}(U,V)} < \varepsilon$, Algorithm 1 is well-defined and either terminates after finitely many iterations or generates a sequence of iterates (q^k) that converges q-linearly with rate β to $\bar{q}$ in Q. If, in addition, $\sigma_{\min}, \sigma_{\max} \in (0, 2)$ in Algorithm 1 and $(B_0 - \nabla^2 \hat{f}(\bar{u})) \in \mathcal{L}(U, V)$ is compact, then the convergence is q-superlinear.*

2) If (q^k) is generated by Algorithm 1, then $(u^k) \subset U_{\mathrm{ad}}$, i.e., every u^k is feasible for (OCE). Moreover, $(u^k), \{\bar{u}\} \subset H^1(\Omega)$ and there is $L_u \in (0, 1)$ such that

$$\|u^k - \bar{u}\|_{L^s(\Omega)} \leq \|q^k - \bar{q}\|_{L^s(\Omega)} \qquad \text{and} \qquad \|u^k - \bar{u}\|_U \leq L_u \|q^k - \bar{q}\|_Q$$

hold for all $k \in \mathbb{N}_0$ and all $s \in [1, \infty]$ that admit the continuous embedding $Q \hookrightarrow L^s(\Omega)$.

If, in addition, (q^k) converges to $\bar{q}$ in Q, then there are $L_y, L_p > 0$ such that

$$\|y^k - \bar{y}\|_Y \leq L_y \|q^k - \bar{q}\|_Q \qquad \text{and} \qquad \|p^k - \bar{p}\|_P \leq L_p \|q^k - \bar{q}\|_Q$$

hold for all k sufficiently large as well as $\|u^k - \bar{u}\|_{H^1(\Omega)} \to 0$ for $k \to \infty$.

3) If (q^k) is generated by Algorithm 1 and converges q-linearly (q-superlinearly) to $\bar{q}$ in Q, then (u^k), (y^k) and (p^k) converge r-linearly (r-superlinearly) in $L^s(\Omega)$, respectively, Y and P, where s is as in 2). Moreover, $(H(q^k))$ converges r-linearly (q-superlinearly) in V to zero, then.

Proof. **Proof of 1):** We recall that $G(q) = \Pi_{U_{\mathrm{ad}}}(q)$, $\hat{G}(q) = \gamma q$ and $F(u) = \nabla \hat{f}(u)$. Moreover, we established in the proof of Lemma 3.6 that $u \mapsto \nabla \hat{f}(u)$ is twice continuously Fréchet differentiable from U to V. Thus, all assertions follow from 1) and 2) of Theorem 2.1.

Proof of 2): The feasibility of the u^k follows since $G(Q) \subset U_{\mathrm{ad}}$ and since $u^k = G(q^k)$. Moreover, as noted in Remark 6 the projection $\Pi_{U_{\mathrm{ad}}}$ satisfies $\Pi_{U_{\mathrm{ad}}}(H^1(\Omega)) \subset H^1(\Omega)$. This implies $(u^k), \{\bar{u}\} \subset H^1(\Omega)$. From the pointwise representation of $\Pi_{U_{\mathrm{ad}}}$ it is straightforward to infer that $|u^k(x) - \bar{u}(x)| \leq |q^k(x) - \bar{q}(x)|$ for a.e. $x \in \Omega$. Since the embedding $Q \hookrightarrow L^s(\Omega)$ yields $(q^k), \{\bar{q}\} \subset L^s(\Omega)$, this implies the first error bound. The second error bound is Corollary 2.2, part 1), with a constant $L_u := L_G > 0$. However, the first error bound, specialized to $s = 2$, shows that L_u can be chosen in $(0, 1]$. The fact that L_u can be chosen strictly smaller than 1 is derived from the norm equivalence of $\|\cdot\|_{H^1(\Omega)}$ and the semi norm $|\cdot|_{H^1(\Omega)}$ in $H^1_0(\Omega)$. For the third and fourth error bound it is enough to recall that $u \mapsto y(u)$ and $u \mapsto p(u)$ are locally Lipschitz from $L^2(\Omega)$ to $Y = P = H^1_0(\Omega)$ and to apply the second error bound. The convergence of (u^k) to $\bar{u}$ in $H^1(\Omega)$ follows from $q^k \to \bar{q}$ in $H^1(\Omega)$ by continuity of $\Pi_{U_{\mathrm{ad}}} : H^1(\Omega) \to H^1(\Omega)$. The continuity of $\Pi_{U_{\mathrm{ad}}}$ can be established as in [2, Theorem 9.5].

Proof of 3): The claims follow from the error estimates in 2) and, for $(H(q^k))$, from part 3) of Corollary 2.2. □

Remark 8. *The compactness property in 1) of Theorem 3.7 holds, in particular, if B_0 is compact from $U = L^2(\Omega)$ to $V = H^1_0(\Omega)$ and Ω is convex or has C^2 boundary. In fact, we have for all $h \in U$ the representation $\nabla^2 \hat{f}(\bar{u})h = y'(\bar{u})^*\big((1 - 6y(\bar{u})p(\bar{u}))y'(\bar{u})h\big) = g \cdot S\big((1 - 6y(\bar{u})p(\bar{u}))y'(\bar{u})h\big)$, where $S(v)$ denotes for $v \in L^2(\Omega)$ the solution $p \in P$ of the adjoint equation (18) with the right-hand side $y(u) - y_d$ replaced by v and $y(u) = y(\bar{u})$ on the left-hand side. Using the regularity of Ω, elliptic regularity implies that S maps $L^2(\Omega)$ continuously to $H^2(\Omega) \cap H^1_0(\Omega)$, hence S is compact from $L^2(\Omega)$ to $H^1_0(\Omega)$. Since $(1 - 6y(\bar{u})p(\bar{u}))y'(\bar{u})h \in L^2(\Omega)$ due to the embedding of $H^1_0(\Omega)$ into $L^6(\Omega)$ in dimension $1 \leq d \leq 3$, we deduce that $h \mapsto g \cdot S\big((1 - 6y(\bar{u})p(\bar{u}))y'(\bar{u})h\big)$ is compact from $L^2(\Omega)$ to $H^1_0(\Omega)$. This demonstrates that $\nabla^2 \hat{f}(\bar{u}) \in \mathcal{L}(U, V)$ is compact from U to V, hence the compactness of $B_0 \in \mathcal{L}(U, V)$ is sufficient to obtain compactness of $B_0 - \nabla^2 \hat{f}(\bar{u})$. Moreover, the same reasoning as for $\nabla^2 \hat{f}(\bar{u})$ shows that $\nabla^2 \hat{f}(u) \in \mathcal{L}(U, V)$ is compact for every $u \in U$. This implies that any choice which ensures the compactness of $B_0 - \nabla^2 \hat{f}(u^0)$ from U to V is sufficient for compactness of $B_0 - \nabla^2 \hat{f}(\bar{u})$, e.g., $B_0 = 0$ or $B_0 = \nabla^2 \hat{f}(u^0)$.*

In Theorem 3.7 and Remark 8 we have worked with the space $H^1_0(\Omega)$ for state, adjoint state and the variable q, i.e., $Q = Y = P = V = H^1_0(\Omega)$. In fact, several other choices are possible as well. We point out that the use of a space $Q = Y = P = V$ with a stronger norm yields stronger convergence properties of the iterates, but, in view of Theorem 3.7 1), requires better initial approximations q^0 and B_0. This clarifies why we include a variety of choices in the following result, several of which have weaker norms than $H^1_0(\Omega)$.

Lemma 3.8. *Theorem 3.7 also holds for $Q = Y = P = V = H^1_0(\Omega) \cap L^\infty(\Omega)$ and for $Q = Y = P = V = L^s(\Omega)$ for any $s \in (2, \infty]$. Remark 8 remains true for each of these choices. For the choice $Q = Y = P = V = L^s(\Omega)$ with $s \in (2, 6)$, the requirement in Remark 8 that Ω is convex or has C^2 boundary can be dropped. If Ω is convex or has C^2 boundary, then the choice $Q = Y = P = V = H^1_0(\Omega) \cap L^\infty(\Omega) \cap W^{1,s}(\Omega)$ with $s \in (2, \infty]/(2, \infty)/(2, 6)$ can be used in Theorem 3.7 and Remark 8 for $d = 1/2/3$.*

Proof. We omit the proofs because they are completely analogue to the ones of Theorem 3.7 and Remark 8. However, we mention that $u \mapsto y(u)$ is still thrice continuously Fréchet differentiable from U to Y for $Y = H^1_0(\Omega) \cap L^\infty(\Omega)$ with Ω being a Lipschitz domain. □

Remark 9. *We notice in both Theorem 3.7 and Lemma 3.8 that the "Hilbert space variable" u^k converges with the analogue rate of q^k, e.g., if (q^k) converges q-superlinearly, then (u^k) converges r-superlinearly. Notably, this is true not only with respect to the norm of the Hilbert space $U = L^2(\Omega)$ that appears in the Broyden update, but also with respect to several stronger norms, e.g., the norm in $L^\infty(\Omega)$ in case of Lemma 3.8. Furthermore, it is important to observe in part 2) of Theorem 3.7 the feature that the L^s-error of (q^k) bounds the L^s-error of the controls (u^k) with constant 1, since the control error is often the one of interest.*

3.3.2. Time-dependent control of the heat equation

Let $N \in \mathbb{N}$, $Q := V := C([0,T])^N$, $U := L^2(I)^N$, and $Y := P := W(I; L^2(\Omega), H_0^1(\Omega))$. We recall from [36, §3.4] that Y is an appropriate space for weak solutions of the heat equation. As a non-stationary optimal control problem we consider the optimal tracking of the linear heat equation in $I \times \Omega$ with N time-dependent controls $u(t) = (u_1(t), \dots, u_N(t))^T$, where $\Omega \subset \mathbb{R}^d$, $1 \le d \le 3$, is a nonempty and bounded Lipschitz domain, and the time domain is $I := (0,T)$ for a fixed final time $T > 0$:

$$\min_{(y,u)\in Y\times U_{\text{ad}}} \frac{1}{2}\|y - y_{\text{d}}\|^2_{L^2(I\times\Omega_{\text{obs}})} + \sum_{i=1}^N \frac{\alpha_i}{2}\|u_i\|^2_{L^2(I)} + \sum_{i=1}^N \beta_i \|u_i\|_{L^1(I)}$$

$$\text{(OCP)} \qquad \text{s. t.} \quad \begin{cases} y_t - \Delta y = \sum_{i=1}^N g_i(x) u_i(t) & \text{in } I\times\Omega, \\ y = 0 & \text{on } \Sigma, \\ y(0,x) = y_0(x) & \text{in } \Omega, \end{cases}$$

where $\Sigma := I \times \partial\Omega$. Moreover, $y_{\text{d}} \in L^2(I \times \Omega_{\text{obs}})$ is the desired state, $\Omega_{\text{obs}} \subset \Omega$ is the observation domain, $\alpha_i > 0$ are the control cost parameters per control function, $\beta_i \ge 0$ influences the size of the support of u_i, $y_0 \in L^2(\Omega)$ is the initial state, and $g_i \in L^2(\omega_i)$ are fixed spatial functions living on the (not necessarily disjoint) control domains $\omega_i \subset \Omega$, $1 \le i \le N$. For instance, g_i could be the characteristic function χ_{ω_i} of ω_i for every i. The set of admissible controls is given by

$$U_{\text{ad}} := \left\{ u = (u_1, \dots, u_N)^T \in U : \ a_i \le u_i \le b_i \ \text{ a.e. in } I,\ 1 \le i \le N \right\} \tag{20}$$

with functions $a, b \in L^\infty(I)^N$ that satisfy $a \le b$ a.e. in I, where the inequality is meant componentwise. To stay within the framework of Section 3.2.2 we set $\gamma := 1$ and endow U with the norm $\|u\|_U = \|(u_1, u_2, \dots, u_N)\|_U := (\sum_{i=1}^N \alpha_i \|u_i\|^2_{L^2(I)})^{1/2}$. As this norm is induced by a scalar product and as it is equivalent to the standard norm of $U = L^2(I)^N$, it follows that U is a Hilbert space with respect to this norm. The $L^2(I)^N$-projection onto U_{ad} is denoted by $\Pi_{U_{\text{ad}}} : U \to U_{\text{ad}}$.

From [36, Theorem 3.13] we obtain that for every $u \in U$ there exists a unique $y = y(u) \in Y$ such that the heat equation appearing in (OCP) is satisfied. Since the solution operator $u \mapsto y(u)$ is linear and continuous from U to Y, it is infinitely many times continuously Fréchet differentiable. Moreover, the control reduced version of (OCP) is a convex problem with strongly convex objective, hence it possesses a unique solution $\bar{u} \in U_{\text{ad}}$ with associated state $y(\bar{u}) \in Y$. Proceeding in the same way as for (OCE) we can therefore derive the following result. We use the notation $(s)^+ := \max(0,s)$ and $(s)^- := \min(0,s)$ for $s \in \mathbb{R}$.

Lemma 3.9. *There is exactly one $\bar{q} \in Q$ with $H(\bar{q}) = 0$ for*

$$H : Q \to V, \qquad H_i(q)(t) = \int_{\omega_i} g_i(x) p\Big(\Pi_{U_{\text{ad}}}(\sigma(q))\Big)(t,x)\,dx + \alpha_i q_i(t), \quad 1 \le i \le N. \tag{21}$$

Here, $p = p(u) \in P$ is the adjoint state*, i.e., the unique solution of the* adjoint equation

$$\begin{cases} -p_t - \Delta p = \chi_{I\times\Omega_{\mathrm{obs}}} \cdot (y(u) - y_d) & \text{in } I \times \Omega, \\ p = 0 & \text{on } \Sigma, \\ p(T) = 0 & \text{in } \Omega, \end{cases} \tag{22}$$

and $\sigma : U \to U$ denotes the soft-shrinkage operator whose i-th component, $1 \le i \le N$, is

$$\sigma_i(u)(t) = \rho_i(u_i(t)) \qquad \text{for} \qquad \rho_i : \mathbb{R} \to \mathbb{R}, \quad \rho_i(s) := \left(s - \frac{\beta_i}{\alpha_i}\right)^+ + \left(s + \frac{\beta_i}{\alpha_i}\right)^- .$$

Defining $\bar{u} := \Pi_{U_{\mathrm{ad}}}(\sigma(\bar{q}))$ and $\bar{y} := y(\bar{u})$, the unique optimal solution of (OCP) *is $(\bar{y}, \bar{u}) \in Y \times U_{\mathrm{ad}}$.*

Proof. We only mention the main arguments since the proof is similar to the one of Lemma 3.5. From [36, Lemma 3.17] we infer that under the current assumptions on the problem data there holds $p(u) \in P$. Due to the continuous embedding $P \hookrightarrow C([0,T]; L^2(\Omega))$, cf. [8, Theorem 11.4], this implies $p(u) \in C([0,T]; L^2(\Omega))$. Therefore, defining $\lambda = \lambda(u)$ by $\lambda_i(t) := \int_{\omega_i} g_i(x) p(t,x)/\alpha_i \,\mathrm{d}x$, $t \in [0,T]$, where $p = p(u)$ and $1 \le i \le N$, yields $\lambda \in V$. Since $\lambda(u) = \nabla \hat{f}(u)$, it follows that $\nabla \hat{f}$ maps U to V. For $\varphi : U \to (-\infty, +\infty]$, $\varphi(u) := \delta_{U_{\mathrm{ad}}}(u) + \sum_{i=1}^N \beta_i \|u_i\|_{L^1(I)}$ we confirm by computation that $\mathrm{Prox}_{\varphi_1} : U \to U_{\mathrm{ad}}$ is given by $\mathrm{Prox}_{\varphi_1} = \Pi_{U_{\mathrm{ad}}} \circ \sigma$. The left-out details and further elaborations concerning this proximity operator can be found in [27, Section 3.3]. Since $\gamma = 1$, the assertions follow from Lemma 3.3. □

Remark 10. *The proof of Lemma 3.9 demonstrates that we have to choose $Q = V$ in such a way that $t \mapsto \int_{\omega_i} g_i(x) p(t,x) \,\mathrm{d}x$ belongs to V, where p solves* (22)*. Thus, the available regularity of the adjoint state p, respectively, of the multiplier λ, restricts the choice of $Q = V$. If additional regularity is available, then $Q = V$ may be chosen as a space of smoother functions than $C([0,T])^N$. For instance, from [17, Theorem 1.39] we deduce that if $\Omega_{\mathrm{obs}} = \Omega$ and $y_d \in Y$, then there holds $\frac{\partial p(u)}{\partial t} \in W(I; L^2(\Omega), H_0^1(\Omega))$, hence $p(u) \in H^1(I; H_0^1(\Omega))$. This implies $\lambda \in V = Q$ for the choice $Q := V := H^1(I)^N$. In fact, using $W(I; L^2(\Omega), H_0^1(\Omega)) \hookrightarrow C([0,T]; L^2(\Omega))$ we obtain $p \in C^1([0,T]; L^2(\Omega))$, which implies $\lambda \in V = Q$ for $Q := V := C^1([0,T])^N$. We stress that Lemma 3.9 is valid for all these choices of $Q = V$.*

Assumption 1 holds for (OCP) without further conditions.

Lemma 3.10. *Let $H : Q \to V$ be given by* (21)*. Then Assumption 1 is fulfilled.*

Proof. Since (OCP) is of the form (POR), it suffices to show that the conditions 1)–5) stated below Remark 4 are fulfilled. We have already argued that 1) is true. Concerning semismoothness of $G := \mathrm{Prox}_{\varphi_1} = \Pi_{U_{\mathrm{ad}}} \circ \sigma$ from $Q = C([0,T])^N$ to $U = L^2(I)^N$ we point out that G is a superposition operator. Based on [17, Theorem 2.13] it can thus be shown by standard arguments that G is semismooth from Q to U with respect to the generalized derivative $\partial G = \partial(\Pi_{U_{\mathrm{ad}}} \circ \sigma) \subset \mathcal{L}(Q,U)$ given by

$$\partial G(q) := \bigcup_{\substack{r \in L^\infty(I)^N \text{ with} \\ 0 \le r \le 1 \text{ a.e. in } I}} \Big\{ M(q,r) \Big\},$$

where $0 \le r \le 1$ is meant componentwise and $M = M(q,r) \in \mathcal{L}(Q,U)$ is for $(q,r) \in Q \times L^\infty(I)^N$ defined as

$$(Mh)_i(t) := \begin{cases} h_i(t) & \text{if } q_i(t) \in (a_i(t), b_i(t)), \\ 0 & \text{if } q_i(t) \notin [a_i(t), b_i(t)], \\ r_i(t) h_i(t) & \text{else} \end{cases} \tag{23}$$

if $i \in \{1, \dots, N\}$ is such that $\beta_i = 0$, and as

$$(Mh)_i(t) := \begin{cases} h_i(t) & \text{if } |q_i(t)| > \frac{\beta_i}{\alpha_i} \ \wedge \ \sigma_i(q)(t) \in (a_i(t), b_i(t)), \\ 0 & \text{if } |q_i(t)| < \frac{\beta_i}{\alpha_i} \ \vee \ \sigma_i(q)(t) \notin [a_i(t), b_i(t)], \\ r_i(t)h_i(t) & \text{else} \end{cases} \tag{24}$$

if $i \in \{1, \dots, N\}$ is such that $\beta_i > 0$. It is evident that $\|\partial G(q)\|_{\mathcal{L}(Q,U)}$ is uniformly bounded for all $q \in Q$. That is, 2) and 3) are satisfied. Since $\nabla \hat{f} : U \to V$ is linear and continuous, it is infinitely many times continuously Fréchet differentiable. This yields 4). To establish 5) we use Lemma 3.4. From the linearity of $u \mapsto y(u)$ we obtain that $\hat{f}(u) = \frac{1}{2}\|y(u) - y_d\|^2_{L^2(I\times\Omega_{\text{obs}})}$ is convex, hence (12) is fulfilled. Finally, it is readily checked that the first two properties listed in Lemma 3.4 are satisfied by the elements of $\partial G(q)$, $q \in Q$. Thus, 5) holds. □

Remark 11. *If $\Omega_{\text{obs}} = \Omega$ and $y_d \in Y$, then Lemma 3.10 is also true for the choices $Q := V := H^1(I)^N$ and $P := H^1(I; H^1_0(\Omega))$ as well as $Q := V := C^1([0,T])^N$ and $P := C^1([0,T]; L^2(\Omega))$, cf. Remark 10.*

We obtain the following convergence result for Algorithm 1, in which we denote $y^k := y(u^k)$ and $p^k := p(u^k)$ for $k \geq 0$. Since $\nabla^2 \hat{f}(u)$ is constant with respect to u, we write $\nabla^2 \hat{f}$. Moreover, we recall that $Q = V = C([0,T])^N$, $U = L^2(I)^N$ with norm $\|u\|_U = (\sum_{i=1}^N \alpha_i \|u_i\|^2_{L^2(I)})^{1/2}$, U_{ad} is given by (20), and $Y = P = W(I; L^2(\Omega), H^1_0(\Omega))$. Note in 2) and 3) that (u^k) converges to $\bar{u}$ not only with respect to $\|\cdot\|_U$, but also with respect to stronger norms.

Theorem 3.11. *1) Let $(\bar{y}, \bar{u}) \in Y \times U_{\text{ad}}$ be the solution of (OCP), let H be given by (21), and denote by $\bar{q} \in Q$ the unique root of H. Moreover, let $\beta \in (0,1)$. Then there exist $\delta, \varepsilon > 0$ such that for every pair of starting values $(q^0, B_0) \in Q \times \mathcal{L}(U,V)$ with $\|q^0 - \bar{q}\|_Q < \delta$ and $\|B_0 - \nabla^2 \hat{f}\|_{\mathcal{L}(U,V)} < \varepsilon$, Algorithm 1 is well-defined and either terminates after finitely many iterations or generates a sequence of iterates (q^k) that converges q-linearly with rate β to $\bar{q}$ in Q. If, in addition, $\sigma_{\min}, \sigma_{\max} \in (0,2)$ in Algorithm 1 and $(B_0 - \nabla^2 \hat{f}) \in \mathcal{L}(U,V)$ is compact, then the convergence is q-superlinear.*

2) If (q^k) is generated by Algorithm 1, then $(u^k) \subset U_{\text{ad}}$, i.e., every u^k is feasible for (OCP). Moreover, $(u^k), \{\bar{u}\} \subset L^\infty(I)^N$ and there are $L_y, L_p > 0$ such that

$$\begin{aligned} &\|u^k - \bar{u}\|_{L^s(I)^N} \leq \|q^k - \bar{q}\|_{L^s(I)^N}, && \|u^k - \bar{u}\|_{L^2(I)^N} \leq T^{\frac{1}{2}} \|q^k - \bar{q}\|_{C([0,T])^N}, \\ &\|y^k - \bar{y}\|_Y \leq L_y \|q^k - \bar{q}\|_Q, && \|p^k - \bar{p}\|_P \leq L_p \|q^k - \bar{q}\|_Q \end{aligned}$$

hold for all $k \in \mathbb{N}_0$ and all $s \in [1, \infty]$.

If, in addition, $a, b \in Q$ holds, then we have $(u^k), \{\bar{u}\} \subset Q$ and for all $k \in \mathbb{N}_0$

$$\|u^k - \bar{u}\|_Q \leq \|q^k - \bar{q}\|_Q. \tag{25}$$

In particular, $\|q^k - \bar{q}\|_Q \to 0$ for $k \to \infty$ implies $\|u^k - \bar{u}\|_Q \to 0$.

3) If (q^k) is generated by Algorithm 1 and converges q-linearly (q-superlinearly) to $\bar{q}$ in Q, then (u^k), (y^k) and (p^k) converge r-linearly (r-superlinearly) in $L^s(I)^N$, respectively, Y and P, where s is as in 2). Moreover, $(H(q^k))$ converges r-linearly (q-superlinearly) in V to zero, then.

Proof. **Proof of 1):** The claim of 1) follows from Lemma 3.9 and Theorem 2.1, part 1). The latter holds since Assumption 1 is satisfied, cf. Lemma 3.10.

Proof of 2): The feasibility of the u^k is valid since $G(Q) \subset U_{\text{ad}}$ and since $u^k = G(q^k)$. Moreover, the property $(u^k), \{\bar{u}\} \subset L^\infty(I)^N$ follows from $U_{\text{ad}} \subset L^\infty(I)^N$. To show the first inequality, we remark that $q^k, \bar{q} \in Q$ implies $q^k, \bar{q} \in L^s(I)^N$ for all $s \in [1, \infty]$ and all $k \geq 0$. It is straightforward to

infer for $q, \bar{q} \in L^s(I)^N$ that $|\rho_i(q_i(t)) - \rho_i(\bar{q}_i(t))| \leq |q_i(t) - \bar{q}_i(t)|$ for a.e. $t \in I, 1 \leq i \leq N$. The same can be established for ρ_i replaced by $(\Pi_{U_{\mathrm{ad}}})_i$. Together, this implies $|u_i(t) - \bar{u}_i(t)| \leq |q_i(t) - \bar{q}_i(t)|$ for a.e. $t \in I$, $1 \leq i \leq N$, proving the first error bound in 2). Moreover, this also implies (25) provided that $(\Pi_{U_{\mathrm{ad}}} \circ \sigma)(Q) \subset Q$ if $a, b \in Q$. This property of $\Pi_{U_{\mathrm{ad}}} \circ \sigma$ is elementary to see, for instance by showing it separately for σ and $\Pi_{U_{\mathrm{ad}}}$. The second error bound follows from the first for $s = 2$ by use of $\|q^k - \bar{q}\|_{L^2(I)^N} \leq |I|^{\frac{1}{2}} \|q^k - \bar{q}\|_{C([0,T])^N}$, where $|I|$ is the Lebesgue measure of I. Since $u \mapsto y(u)$ and $u \mapsto p(u)$ are linear and continuous from U to $Y = P$, they are globally Lipschitz, too. This in combination with the estimate $\|u^k - \bar{u}\|_{L^2(I)^N} \leq \|q^k - \bar{q}\|_{L^2(I)^N}$ and the continuous embedding $Q \hookrightarrow L^2(I)^N$ yields the third and fourth error bound.

Proof of 3): The claims follow from the error estimates in 2) and, for $(H(q^k))$, from part 3) of Corollary 2.2. □

Remark 12. *It is not difficult to argue that Theorem 3.11 also holds for $Q := V := L^{\hat{s}}(I)^N$ for any $\hat{s} \in (2, \infty]$. Of course, the L^s estimates of that theorem are then only true for $s \in [1, \hat{s}]$. Note that the use of a weaker norm in Q and V relaxes the assumption on (q^0, B_0) and the compactness requirement in that theorem.*

In Theorem 3.11 we have worked with $Q = V = C([0,T])^N$ and $P = W(I; L^2(\Omega), H_0^1(\Omega))$. If $\Omega_{\mathrm{obs}} = \Omega$ and y_d is more regular, then we can employ stronger spaces, resulting in stronger convergence properties (but also in the need for better initial data (q^0, B_0)). For simplicity we consider constant bounds.

Lemma 3.12. *Let $\Omega_{\mathrm{obs}} = \Omega$, $y_d \in Y$, and let a_i, b_i be constant for each $1 \leq i \leq N$. Then all claims of Theorem 3.11 except (25) are true for $Q := V := H^1(I)^N$, $U := L^2(I)^N$, $Y := W(I; L^2(\Omega), H_0^1(\Omega))$, and $P := H^1(I; H_0^1(\Omega))$.*

Proof. The proof is completely analogue to the one of Theorem 3.11, except for the claim above and the one below (25). We thus establish only these two assertions, i.e., $(u^k), \{\bar{u}\} \subset Q = H^1(I)^N$ and $u^k \to \bar{u}$ in Q provided that $q^k \to \bar{q}$ in Q. In fact, this follows since $G = \Pi_{U_{\mathrm{ad}}} \circ \sigma$ satisfies $G(Q) \subset Q$ and since $G : Q \to Q$ is continuous. Both properties can be proven separately for $\Pi_{U_{\mathrm{ad}}}$ and σ. For $\Pi_{U_{\mathrm{ad}}}$ these properties have already been observed in Remark 6; for σ the corresponding proof is elementary. □

We infer from Lemma 3.12 a result that includes the compactness of $\nabla^2 \hat{f}$. Here we can work with nonconstant bounds again.

Corollary 3.13. *Let $\Omega_{\mathrm{obs}} = \Omega$, $y_d \in Y$, and let $a, b \in L^\infty(I)^N$. Then all claims of Theorem 3.11 except for the second error estimate in part 2) are true for $Q := V := L^{\hat{s}}(I)^N$ with arbitrary $\hat{s} \in (2, \infty]$, $U := L^2(I)^N$, $Y := W(I; L^2(\Omega), H_0^1(\Omega))$, and $P := H^1(I; H_0^1(\Omega))$, provided that the interval $[1, \infty]$ in part 2) of Theorem 3.11 is replaced by $[1, \hat{s}]$. Furthermore, the operator $\nabla^2 \hat{f} \in \mathcal{L}(U, V)$ is compact.*

Proof. This follows from Remark 12 and the fact that $H^1(I)^N \hookrightarrow L^{\hat{s}}(I)^N$ is a compact embedding for any $\hat{s} \in [1, \infty]$. □

Still higher regularity is available in the situation of Lemma 3.12.

Lemma 3.14. *Let $\Omega_{\mathrm{obs}} = \Omega$, $y_d \in Y$, and let a_i, b_i be constant for each $1 \leq i \leq N$. Then all claims of Theorem 3.11 except (25) and the convergence assertion below it are true for $Q := V := C^{0,\hat{s}}([0,T])^N$ with arbitrary $\hat{s} \in (0, 1]$, $U := L^2(I)^N$, $Y := W(I; L^2(\Omega), H_0^1(\Omega))$, and $P := C^1([0,T]; L^2(\Omega))$. In this setting $q^k \to \bar{q}$ in Q implies $u^k \to \bar{u}$ in $C^{0,\hat{s}-\tau}([0,T])^N$ for any $\tau \in (0, \hat{s})$. Moreover, if $\hat{s} \neq 1$, then $\nabla^2 \hat{f} \in \mathcal{L}(U, V)$ is compact.*

Proof. The main task is to establish that $G = \Pi_{U_{\mathrm{ad}}} \circ \sigma$ satisfies $G(Q) \subset Q$ and that $G : Q \to C^{0,\hat{s}-\tau}([0,T])^N$ is continuous at $\bar{q}$ for any $\tau \in (0, \hat{s})$. The proof can be undertaken separately for $\Pi_{U_{\mathrm{ad}}}$ and σ. The property $G(Q) \subset Q$ follows from [2, Theorem 7.1], and the continuity at $\bar{q}$ is a consequence of [2, first part of Theorem 7.6]. The compactness of $\nabla^2 \hat{f} \in \mathcal{L}(U, V)$ is implied by the fact that $C^{0,1}([0,T])^N \hookrightarrow C^{0,\hat{s}}([0,T])^N$ is compact for $\hat{s} \in (0,1)$. □

4. Implementation

The hybrid framework is tested with six different quasi-Newton implementations: both full and limited-memory implementations of Broyden, SR1, and BFGS. The limited-memory implementations store the last up to L (called the limit) updates as vectors and the application of B_k is computed in a matrix-free manner. Consequently, the Newton systems are solved with iterative methods, specifically with GMRES or cg. The limited-memory BFGS method is implemented in the compact variant according to [7], see also [23, (7.24)]. Additionally, the quasi-Newton methods are compared to Newton's method itself, i.e., setting $B_k = \nabla^2 \hat{f}(u^k)$ for all k and dropping lines 8–11 in Algorithm 1. Here, the matrix-free evaluation in a direction is implemented via forward-backward solve. We stress that when Newton's method is used, Algorithm 1 is a semismooth Newton method (as opposed to a semismooth Newton-type method in the quasi-Newton case).

The methods are applied within three different globalization frameworks. First, a standard backtracking line search on the residual norm $\|H\|_U$ is used together with GMRES. This scheme is denoted ls-GMRES. The line search selects the smallest integer $k \geq 0$ such that both $\|H(q^k + \rho^k s^k)\|_U < \|H(q^k)\|_U$ and $\rho^k \geq 10^{-6}$ are satisfied, where $\rho = 0.5$ in all experiments. If there is no such k, then we take $k = 19$, which yields the smallest possible step size ρ^k that is larger than 10^{-6}. GMRES from MATLAB is used with a tolerance of 10^{-10} and a maximum number of 50 iterations.

For the nonconvex examples two additional frameworks are applied. On the one hand, we use a non-monotone line search (denoted nls-GMRES) with $N_{\mathrm{ls}} \in \mathbb{N}$ steps, where the step size $\rho^k = 0.5^k$, $k \geq 0$, is accepted as soon as $R(q^k + \rho^k s^k) < \max\{R(q^k), R(q^{k-1}), \ldots, R(q^{k-N_{\mathrm{ls}}+1})\}$ holds. If this does not happen for $0 \leq k \leq 19$, then we take $k = 19$. Therein, $R(v)$ stands for the residual norm $\|H(v)\|_U$ or, in the notation of Section 3.2, the objective $f(v) + \varphi(v)$.

On the other hand, a trust-region framework based on [34] with Steihaug-cg (denoted tr-cg) is applied. The precise algorithm is included in Appendix A. It is started with a radius $\varrho_0 = 0.1$ and stopped with a relative tolerance of 10^{-5}. The parameters are $\sigma_1 = 0.05, \sigma_2 = 0.25, \sigma_3 = 0.7$, radius factors $f_1 = 0.4, f_2 = 2, f_3 = 0.6$, and a maximum radius $\varrho_{\max} = 2$. Up to 300 iterations are allowed. The update of the quasi-Newton matrix is also carried out at rejected steps. To be able to use Steihaug-cg, the linear system is first reduced to a symmetric one by restricting to the Hilbert space induced by the inner product $(\cdot, M_k \cdot)_U$ with $M_k \in \partial G(q^k)$, confer Lemma 3.4 and [27, Def. 3.4]. Then a correction step gives the full update, confer [27, (3.25)]. The cg method is limited to 100 iterations and stopped with a relative tolerance of 10^{-10}. We emphasize that we use this small tolerance to suppress the influence of inexact linear system solves in the numerical results.

5. Numerical experiments

The numerical experiments below show the application of the hybrid method to optimal control problems. The first example considers the heat equation, the second the bilinear control of the

Bloch equation in magnetic resonance imaging. As nonsmooth problem parts we deal with pointwise box constraints on the controls and sparsity promoting objectives. All computations are carried out using MATLAB 2017a and a workstation with two Intel Xeon X5675, which in total possesses twelve cores with 3.06GHz each and 24GB RAM. All time measurements are performed on a single CPU without multi-threading.

5.1. Time-dependent tracking of the heat equation

The first example problem is sparse control of the linear heat equation with $\Omega_{\text{obs}} = \Omega = (-1,1)^2$ and time domain $I = (0,1)$. Specifically, we consider

$$\min_{(y,u)\in Y\times U_{\text{ad}}} \frac{1}{2}\|y - y_{\text{d}}\|^2_{L^2(I\times\Omega_{\text{obs}})} + \frac{\alpha}{2}\|u\|^2_{L^2(I)} + \beta\|u\|_{L^1(I)}$$
$$\text{s. t.} \quad \begin{cases} y_t - \Delta y = \chi_\omega(x)u(t) & \text{in } I\times\Omega, \\ \partial_\nu y = 0 & \text{on } \Sigma, \\ y(0,x) = y_0(x) & \text{in } \Omega. \end{cases}$$

Therein, $U_{\text{ad}} = \{u \in U : a \le u(t) \le b \text{ for a.e. } t \in I\}$ with $a = -1$ and $b = 1$, $y_{\text{d}}(t,x) = 2\sin(2\pi t)$, $\alpha > 0$, $\beta \ge 0$, $\chi_\omega(x) \in L^\infty(\Omega)$ is the characteristic function of the right half $\omega = (0,1)\times(-1,1)$, and $y_0 \equiv 0$. The example is a special case of (OCP) with $N = 1$, however using Neumann instead of Dirichlet boundary conditions. We emphasize that the results of Section 3.3.2 can also be developed for these boundary conditions. In view of Corollary 3.13 and Lemma 3.14 and since $y_{\text{d}} \in W(I; L^2(\Omega), H^1(\Omega))$, $\Omega_{\text{obs}} = \Omega$, and the bounds a, b are constant, we expect convergence in rather strong norms. Specifically, we are interested in q-superlinear convergence of (q^k) in $H^1(I)$ and in $C^{0,s}([0,T])$ for all $s \in (0,1)$, convergence of (u^k) in $H^1(I)$ and in $C^{0,s-\tau}([0,T])$ for all $\tau \in (0,s)$, and r-superlinear convergence of (u^k) in $L^2(I)$ and $L^\infty(I)$. Furthermore, we should be able to observe the error estimates $\|u^k - \bar u\|_{L^s(I)} \le \|q^k - \bar q\|_{L^s(I)}$ for $s = 2, \infty$, cf. Theorem 3.11 2). We will investigate these properties numerically.

We use an unstructured triangular mesh with 725 P1 elements. It is generated using the MATLAB function `initmesh` with `Hmax`=0.1, and it is displayed in Figure 1. The time domain $[0,1]$ is discretized equidistantly with 101 points, and we use the CG(1) method as time-stepping scheme (corresponding to the Crank-Nicolson method) together with a piecewise constant discretization of q and u. We initialize the algorithm with $q^0 = 0$. The choice of B_0 is specified below. Since the problem is strongly convex the solution is possible based on the simple globalization ls-GMRES.

The problem falls into the category considered in Section 3.2. From Lemma 3.9 we obtain that we have to find the unique solution $\bar q$ of $H(q) = F(G(q)) + \hat G(q) = 0$, where $F(u) = \nabla\hat f(u)$ with $\hat f(u) = \frac{1}{2}\|y(u) - y_{\text{d}}\|^2_{L^2(I\times\Omega)}$, $G(q) = \Pi_{U_{\text{ad}}}(\sigma(q))$ with L^2-projection $\Pi_{U_{\text{ad}}}(q) = \max(a, \min(b, q))$ and soft-shrinkage operator $\sigma(q) = (q - \frac{\beta}{\alpha})^+ + (q + \frac{\beta}{\alpha})^-$, and $\hat G(q) = \alpha q$. The methods are analyzed for varying parameters α, β, limits, and initial matrices B_0. If not explicitly stated otherwise, then we use $\alpha = 0.01$ and a limit $L = 30$ for the limited-memory methods.

At first we investigate the case without sparsity, i.e., $\beta = 0$. Here, the proximal mapping is the projection onto the admissible set, $G(q) = \Pi_{U_{\text{ad}}}(q) = \max(a, \min(b, q))$. In the algorithm we select the differential $M_k \in \partial G(q^k)$ that satisfies, for all $h \in Q$, $(M_k h)(t) = h(t)$ for $a < q^k(t) < b$ and $(M_k h)(t) = 0$ for all other $t \in [0,T]$; cf. (23). The optimal controls for different α are depicted in Figure 1. The control constraints are active in the optimum for $\alpha \le 1.04$ but not for $\alpha \ge 1.045$. The optimal control for $\alpha = 10^{-3}$ is nearly bang-bang (99 of 100 control points are on the bounds), therefore we do not investigate smaller α.

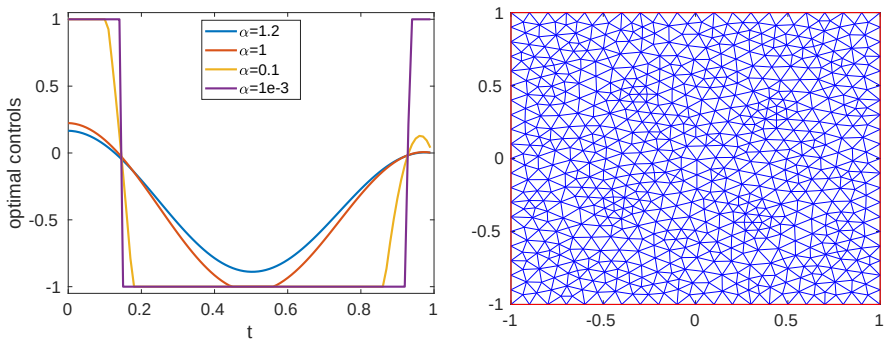

Figure 1: Optimal control $\bar{u}$ for different α (left) and domain Ω with triangulation (right).

		Broyden				SR1				BFGS			
α	$\lvert\mathcal{A}\rvert$	αI	$\frac{y^Ts}{s^Ts}I$	$\frac{y^Ty}{y^Ts}I$	0	αI	$\frac{y^Ts}{s^Ts}I$	$\frac{y^Ty}{y^Ts}I$	0	αI	$\frac{y^Ts}{s^Ts}I$	$\frac{y^Ty}{y^Ts}I$	I
10^0	12	12	10	10	6	8	9	10	5	12	7	9	12
10^{-1}	80	12	15	–	7	10	17	–	6	14	11	14	18
10^{-2}	95	11	19	–	7	9	–	–	7	10	13	20	22
10^{-3}	99	10	17	–	9	9	–	–	12	9	12	12	14

	SN	Broyden				SR1			BFGS	
α		$\nabla^2\hat{f}(u^0)$	I	αI	0	I	αI	0	I	αI
10^0	2	2	12	12	6	8	8	5	12	12
10^{-1}	4	4	16	12	7	14	10	6	18	14
10^{-2}	6	6	17	11	7	13	9	7	22	10
10^{-3}	11	11	14	10	9	16	10	13	14	10

Table 1: Iterations of the quasi-Newton implementations for different α (rows) and for different B_0 (columns per method). The upper part shows the limited-memory implementations, the lower part the full implementations. $|\mathcal{A}|$ depicts the number of control components out of 100 that are on the upper or lower bound.

The first study is devoted to the initialization of B_0. We compare the different methods by means of the iteration numbers that are needed to reach the relative tolerance 10^{-8}. If a method does not reach this tolerance within 50 iterations, we indicate this in the tables by the symbol – . Table 1 shows the iteration numbers of the different methods. The upper part of that table is concerned with the limited-memory methods. The first two columns show α and the number of active points $|\mathcal{A}|$, i.e., points where the optimal control assumes the value of either upper or lower bound. Different initializations are compared, including the scaled identity αI, the zero matrix 0, and the two formulas [7, Eq.(3.23)] $B_0 = y^Ts/(s^Ts)I$ and [23, Eq.(7.20)] $B_0 = y^Ty/(y^Ts)I$. We note that both formulas are implemented in the first step with $B_0 = 0$ for Broyden/SR1 and αI for BFGS. The table shows that the performance strongly depends on the initialization, especially for small α. A good performance for any α and for all three methods can be obtained by choosing $B_0 = \alpha I$. However, Broyden and SR1 show faster convergence with the zero initialization. The lower part of the table provides the iteration counts of the full implementations including for the semismooth Newton method, denoted SN. The iteration numbers for $B_0 = 0, \alpha I$ coincide with those of the limited-memory implementations since the iteration numbers are always below the limit $L = 30$. Additionally, the iteration counts with an exact initial Hessian $B_0 = \nabla^2\hat{f}(u^0)$, computed with $N_t - 1$ Hessian directions, is included in the comparison (depicted only for

Broyden). With this exact initialization all methods converge exactly as fast as the semismooth Newton method in this example and the tolerance is met by far. This is not surprising since in this setting all three variants of the hybrid method coincide with the semismooth Newton method (pointing out once again that the limit L is irrelevant due to the low iteration number). However, the relative tolerance of 10^{-8} is also fulfilled quickly for the other initialization of B_0. In direct comparison, the scaled identity αI outperforms I. Based on these results the initialization is from now on in general set to $B_0 = 0$ for Broyden and SR1, and to $B_0 = y^T s/(s^T s)I$ for BFGS. We point out that, from an infinite-dimensional point of view, the choices $B_0 = \nabla^2 \hat{f}(u^0) = \nabla^2 \hat{f}(\bar{u})$ and $B_0 = 0$ yield to $B_0 - \nabla^2 \hat{f}(\bar{u})$ being compact, cf. Corollary 3.13 and Lemma 3.14, while this is not the case for the choices $B_0 = I$ and $B_0 = \alpha I$. We recall that operator compactness is required in the result on superlinear convergence, Theorem 3.11, part 1).

The iteration counts of the limited-memory methods for different discretizations are depicted in Table 2 for $\alpha = 0.01$. The time domain $[0, 1]$ is discretized equidistantly with $N_c + 1$ points, the spatial meshes are again generated with `initmesh`. The results show mesh independence for all three quasi-Newton methods with respect to both the spatial and the temporal discretization.

	Broyden			SR1			BFGS		
$N_c \setminus N_x$	725	1938	7701	725	1938	7701	725	1938	7701
100	7	7	7	7	7	7	13	13	13
400	7	8	7	7	7	7	14	14	14
1600	7	7	7	7	7	7	14	14	14
6400	7	8	7	7	7	7	14	14	14

Table 2: Iterations of the limited-memory quasi-Newton implementations for different number of control points N_c (rows) and for different N_x (three columns each): The three columns denote triangulations with $N_x = 725, 1938, 7701$ nodes.

The runtimes of the limited-memory methods and the semismooth Newton method are provided in Table 3. Three different spatial discretizations are applied (columns) as well as several different time discretizations (rows). All values are averages of five runs. Broyden and SR1 show nearly identical runtimes in this example and are twice as fast as BFGS. All three methods outperform the semismooth Newton method in runtime. For all time and space discretizations a speed-up factor of five to six is observed for Broyden and SR1, and three to four for BFGS.

	SN			Broyden			SR1			BFGS		
$N_c \setminus N_x$	725	1938	7701	725	1938	7701	725	1938	7701	725	1938	7701
100	11	40	210	2	5	33	2	6	31	3	9	56
400	40	128	681	6	22	112	6	20	111	12	37	211
1600	158	486	2755	27	83	446	27	82	449	51	154	841
6400	643	1971	10855	137	456	2095	148	419	2102	290	763	3945

Table 3: Runtimes of the semismooth Newton and the limited-memory quasi-Newton methods for different discretizations. The rows show different temporal discretizations, the columns different spacial discretizations. All runtimes are mean values of five runs.

5.2. Sparse control with box constraints

In the following we change to $\beta > 0$ and investigate sparse control with the box constraints given by U_{ad} as before. We have discussed the corresponding proximal mapping in Lemma 3.9. It coincides with the soft-shrinkage operator and subsequent projection onto U_{ad}, i.e., $G(q) = \Pi_{U_{\text{ad}}}(\sigma(q))$. In the algorithm we select the differential $M_k \in \partial G(q^k)$ that satisfies, for all $h \in Q$, $(M_k h)(t) = h(t)$ if $|q^k(t)| \geq \beta/\alpha$ and $a < \sigma(q^k)(t) < b$, and $(M_k h)(t) = 0$ otherwise; cf. (24). The optimal solutions for different α, β are depicted in Figure 2. The structure of the optimal control is analyzed in the left part of Table 4 using two different numbers of time points $|O|$ and $|\mathcal{A}|$, which collect the number of time points with zero control and with control on the bound ($u(t) = a$ or $u(t) = b$). It can be seen that the inequality constraints are in general inactive in the optimum if $\alpha \geq 1$. On the other hand, smaller values $\alpha \leq 0.01$ result in only one or two inactive points (the number of inactive points is $100 - |\mathcal{A}| - |O|$). Increasing the parameter β between 0.1 and 1 increases the sparsity from around 20% to 80%. For $\beta \geq 2$ the optimal solution is zero.

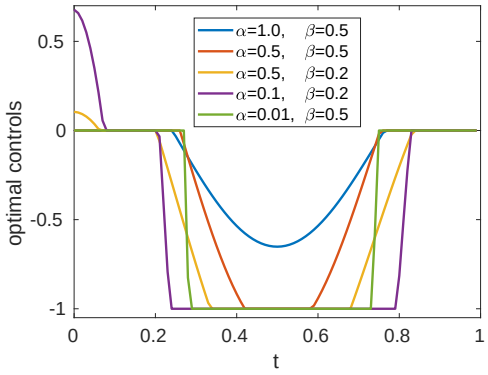

Figure 2: Sparse optimal control for different α, β.

The sparse control problem is solved with the different hybrid semismooth quasi-Newton methods using their limited-memory implementations. The resulting iteration counts of these methods are depicted for different α, β in Table 4. All methods show fast convergence for all values of α and β. The methods tend to require less iterations for larger β (which corresponds to more degrees of freedom fixed to zero). Broyden (Br) and SR1 always converge in under 8 iterations, BFGS needs up to 17 iterations.

The next study analyses the superlinear convergence properties numerically. For comparison the optimal solution $\bar{q}$ is first computed in high precision with the semismooth Newton method and GMRES using fine relative tolerances of 10^{-14} for both. Then the indicators of superlinear convergence $r_u^k := \|u^{k+1} - \bar{u}\|_Z / \|u^k - \bar{u}\|_Z$ and $r_q^k := \|q^{k+1} - \bar{q}\|_Z / \|q^k - \bar{q}\|_Z$ are computed for different norms, namely $Z = L^2(I), L^\infty(I), H^1(I), C^{0,1}([0,T])$. For q-superlinear convergence these indicators should converge to zero in the last steps of an optimization run. Table 5 depicts these indicators for the different methods showing the last four iterations in each case. The results are obtained for $\alpha = 0.1, \beta = 0.2$, and a relative tolerance of 10^{-8}. We observe that the semismooth Newton method converges in one step as soon as the active set has converged. Broyden and SR1 show fast superlinear convergence with final indicators between $7 \cdot 10^{-3}$ and $6 \cdot 10^{-4}$ both for the control u and the optimization variable q. For BFGS, the indicators are slightly larger, but also decrease towards the end. If α is further reduced to 0.01, we observe also a one-step convergence for all three limited-memory methods, which can be explained by the fact that all but one time point are either active or zero then.

α	β	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{O}\rvert$	SN	Br	SR1	BFGS
1.00	0.1	0	18	4	6	5	8
0.10	0.1	68	20	4	8	6	12
0.01	0.1	79	20	6	6	6	13
1.00	0.2	0	34	3	6	5	7
0.10	0.2	56	30	4	8	7	12
0.01	0.2	68	31	5	7	6	17
1.00	0.5	0	48	3	5	5	7
0.10	0.5	41	53	3	6	5	8
0.01	0.5	45	53	5	7	6	9
1.00	1.0	0	71	3	4	4	6
0.10	1.0	14	77	4	6	5	10
0.01	1.0	21	77	5	7	6	9

Table 4: Iteration counts of the semismooth Newton method and of the limited-memory quasi-Newton methods for different α and β (last four columns). $\lvert\mathcal{A}\rvert$ and $\lvert\mathcal{O}\rvert$ stand for the number out of 101 control points that are on the bounds, respectively, that are zero.

	SN		Broyden		SR1		BFGS	
	r_u^k	r_q^k	r_u^k	r_q^k	r_u^k	r_q^k	r_u^k	r_q^k
L^2-norm	$3 \cdot 10^{-2}$	$2 \cdot 10^{-1}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$6 \cdot 10^{-2}$	$3 \cdot 10^{-2}$
	$3 \cdot 10^{-1}$	$7 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$2 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$4 \cdot 10^{-1}$	$3 \cdot 10^{-1}$
	$4 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$6 \cdot 10^{-2}$	$1 \cdot 10^{-1}$
	$2 \cdot 10^{-10}$	$9 \cdot 10^{-11}$	$6 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	$6 \cdot 10^{-4}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$
L^∞-norm	$1 \cdot 10^{0}$	$3 \cdot 10^{0}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$7 \cdot 10^{-2}$
	$3 \cdot 10^{-1}$	$9 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-1}$	$3 \cdot 10^{-1}$
	$4 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
	$3 \cdot 10^{-10}$	$3 \cdot 10^{-10}$	$9 \cdot 10^{-3}$	$9 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$
H^1-norm	$1 \cdot 10^{0}$	$7 \cdot 10^{-1}$	$5 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$6 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	$1 \cdot 10^{-1}$
	$3 \cdot 10^{-1}$	$8 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$4 \cdot 10^{-2}$	$4 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$4 \cdot 10^{-1}$	$3 \cdot 10^{-1}$
	$6 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
	$3 \cdot 10^{-10}$	$3 \cdot 10^{-9}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-2}$
$C^{0,1}$-norm	$4 \cdot 10^{1}$	$1 \cdot 10^{1}$	$4 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$2 \cdot 10^{-1}$
	$3 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$2 \cdot 10^{-2}$	$4 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$3 \cdot 10^{-1}$	$3 \cdot 10^{-1}$
	$5 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
	$4 \cdot 10^{-10}$	$8 \cdot 10^{-9}$	$9 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$2 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-2}$

Table 5: Indicators of superlinear convergence with $\alpha = 0.1$, $\beta = 0.2$: The two columns per method show r_u^k and r_q^k for the last four steps of an optimization run. Each group of four rows shows these indicators for different norms displayed in the first column.

For the limited-memory methods we take a closer look at the convergence of u and q in different norms. Table 6 displays the following expressions for the last four steps of each optimization method: $e^k_{u,L^2} := \|u^k - \bar{u}\|_{L^2}$, $e^k_{u,H^1} := \|u^k - \bar{u}\|_{H^1}$, $e^k_{u,L^\infty} := \|u^k - \bar{u}\|_{L^\infty}$, and analogue definitions e^k_{q,L^2}, e^k_{q,H^1}, e^k_{q,L^∞} for q. The semismooth Newton method exhibits one-step convergence as

soon as the active sets are converged. The other methods show a quick decrease of all values towards the relative tolerance during the last four steps of the optimization run. In particular, SR1 yields the fastest reduction, while BFGS shows a significantly slower convergence here. A closer look confirms that $e^k_{u,L^2} \le e^k_{q,L^2}$ and $e^k_{u,L^\infty} \le e^k_{q,L^\infty}$ are true in any case. Theorem 3.11, part 2) demonstrates that these inequalities are always satisfied (which has nothing to do with Broyden's method but with the fact that $u^k = G(q^k)$ and that $G : L^s \to L^s$ is nonexpansive for any $s \in [1, \infty]$; therefore this estimate appears in all four methods under consideration). We observe that, in contrast, $e^k_{u,H^1} \le e^k_{q,H^1}$ does not hold in general. However, the results indicate that e^k_{u,H^1} still goes to zero, which agrees with part 2) of Theorem 3.11 for the choice $Q = V = H^1(I)$ described in Lemma 3.12.

SN						Broyden					
e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}	e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}
$3 \cdot 10^{-2}$	$2 \cdot 10^{-1}$	$1 \cdot 10^{0}$	$7 \cdot 10^{-1}$	$1 \cdot 10^{0}$	$3 \cdot 10^{0}$	$3 \cdot 10^{-6}$	$5 \cdot 10^{-6}$	$2 \cdot 10^{-4}$	$6 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	$9 \cdot 10^{-5}$
$9 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$4 \cdot 10^{-1}$	$5 \cdot 10^{-2}$	$3 \cdot 10^{-1}$	$3 \cdot 10^{-1}$	$3 \cdot 10^{-7}$	$6 \cdot 10^{-7}$	$2 \cdot 10^{-5}$	$2 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$
$4 \cdot 10^{-4}$	$8 \cdot 10^{-4}$	$2 \cdot 10^{-2}$	$2 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-8}$	$3 \cdot 10^{-8}$	$7 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$4 \cdot 10^{-7}$	$4 \cdot 10^{-7}$
$6 \cdot 10^{-14}$	$7 \cdot 10^{-14}$	$7 \cdot 10^{-12}$	$7 \cdot 10^{-12}$	$4 \cdot 10^{-12}$	$4 \cdot 10^{-12}$	$9 \cdot 10^{-11}$	$2 \cdot 10^{-10}$	$7 \cdot 10^{-9}$	$2 \cdot 10^{-9}$	$4 \cdot 10^{-9}$	$4 \cdot 10^{-9}$

SR1						BFGS					
e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}	e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}
$1 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$4 \cdot 10^{-3}$	$6 \cdot 10^{-4}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$1 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$8 \cdot 10^{-6}$	$8 \cdot 10^{-6}$
$2 \cdot 10^{-7}$	$5 \cdot 10^{-7}$	$2 \cdot 10^{-5}$	$7 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$3 \cdot 10^{-8}$	$3 \cdot 10^{-8}$	$3 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$
$5 \cdot 10^{-9}$	$1 \cdot 10^{-8}$	$4 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$1 \cdot 10^{-8}$	$1 \cdot 10^{-8}$	$1 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$8 \cdot 10^{-7}$	$8 \cdot 10^{-7}$
$4 \cdot 10^{-12}$	$7 \cdot 10^{-12}$	$4 \cdot 10^{-10}$	$3 \cdot 10^{-10}$	$2 \cdot 10^{-10}$	$2 \cdot 10^{-10}$	$1 \cdot 10^{-9}$	$3 \cdot 10^{-9}$	$1 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$6 \cdot 10^{-8}$	$6 \cdot 10^{-8}$

Table 6: Convergence of the semismooth Newton method and the limited-memory methods in different norms for $\alpha = 0.1, \beta = 0.2$: Depicted are $e^k_{u,L^2} = \|u^k - \bar{u}\|_{L^2}$, $e^k_{q,L^2} = \|q^k - \bar{q}\|_{L^2}$, $e^k_{u,H^1} = \|u^k - \bar{u}\|_{H^1}$, $e^k_{q,H^1} = \|q^k - \bar{q}\|_{H^1}$, $e^k_{u,L^\infty} = \|u^k - \bar{u}\|_{L^\infty}$, and $e^k_{q,L^\infty} = \|q^k - \bar{q}\|_{L^\infty}$ for the last four steps of each optimization run.

5.3. Sparse control of the Bloch equation

As example for a nonconvex optimization problem we investigate the bilinear control of the Bloch equations in magnetic resonance imaging (without relaxation, in the rotating frame, and on-resonance). A realistic optimal control modeling for radio-frequency (RF) pulse design in slice-selective imaging is considered based on [30]. However, we add sparsity to the control model, which is desired in practice in particular since the duty cycle of the RF amplifier is often limited. For background on magnetic resonance imaging confer, e.g., [6]. As model problem we consider the slice-selective imaging with a single slice. Here, imaging data of a whole slice is to be acquired. The spatial field of view is described by its extent $\Omega \subset \mathbb{R}$ perpendicular to the slice direction. The slice itself is described by $\Omega_{\text{in}} \subset \Omega$ while the remaining part of Ω is denoted by $\Omega_{\text{out}} = \Omega \setminus \Omega_{\text{in}}$. The latter should not contribute to the data acquisition. The control problem is modeled as tracking of the nuclear magnetization vector $\mathbf{M} = \mathbf{M}(u) = (M_1, M_2, M_3)$ at the terminal time T. Specifically, we consider

$$\min_{u \in U_{\text{ad}}} \quad \hat{f}(u) + \frac{\alpha}{2}\|u\|^2_{L^2(I)} + \beta\|u\|_{L^1(I)} \tag{26}$$

$$\text{s.t.} \quad \dot{\mathbf{M}}(t,x) = \gamma \mathbf{M}(t,x) \times \mathbf{B}(t,x) \quad \text{a.e. in } I \times \Omega, \qquad \mathbf{M}(0,x) = \mathbf{M}_0(x) \quad \text{a.e. in } \Omega \tag{27}$$

with $\alpha > 0$, proton gyromagnetic ratio $\gamma = 267.5380$ [rad/s/μT], given initial condition $\mathbf{M}_0(x)$, spatial domain $x \in \Omega = (-c, c)$ with $c = 0.06$ [m], and time $t \in I = (0, T)$ with $T = 2.69$ [ms].

The term $\hat{f}(u)$ is a tracking-type functional at the terminal time T describing the intended use of the RF pulse, see below. The external magnetic field $\mathbf{B}(t,x) = (u(t), v(t), w(t)x)$ depends on the RF pulse $(u,v) \in L^2(I)^2$ and the slice-selective gradient amplitude $w = w(t) \in L^2(I)$. While these three time-dependent functions can often be controlled, we consider for simplicity the situation in which $w \equiv 2$ is given and the RF pulses are real-valued, i.e., $v \equiv 0$. Hence, u is the control variable. Technical constraints of the scanner hardware in form of limitations of the RF amplifier are modeled as control constraints $U_{\text{ad}} = \{u \in L^2(I) : |u| \le u_{\max}\}$ with $u_{\max} = 1.2$ $[10^2 \mu\text{T}]$. The values reflect a typical 3T magnetic resonance scanner hardware.

The specific example here is the optimization of a refocusing pulse, which is, among others, a central building block of the clinically important turbo spin echo based sequences. The initial condition results from assuming that an ideal 90°-excitation pulse for the same slice has been applied before, keeping the net magnetization vectors out of the slice in the steady state $(0,0,1)^T$ while exciting the slice itself. In particular, we set $\mathbf{M}_0 = \chi_{\Omega_{\text{out}}}(x)(0,0,1)^T + \chi_{\Omega_{\text{in}}}(x)(0,1,0)^T$. A slice of 1.65 [cm] thickness is assumed: $\Omega_{\text{in}} = [-0.00825, 0.00825]$, $\Omega_{\text{out}} = \Omega \setminus \Omega_{\text{in}}$.
The aim of the refocusing is to flip the magnetization vectors in the $x-y-$plane in the interior Ω_{in} of the slice, which is modeled as rotation around the x axis with angle π. This desired magnetization pattern at the end time $t = T$ of the refocusing pulse is given by

$$\hat{f}(u) = \frac{1}{2}\int_{\Omega_{\text{in}}} (M_1(T,x))^2 + (M_2(T,x)+1)^2 dx + \frac{1}{2}\int_{\Omega_{\text{out}}} (1 - M_3(T,x))^2 dx,$$

recalling that $\mathbf{M} = \mathbf{M}(u)$. However, this tracking term for basic refocusing pulses is typically not used in numerical practice. Instead, we apply a more involved formulation of the desired state at the terminal time for advanced refocusing pulses, that we describe now. Because of practical reasons including robustness issues, refocusing pulses are generally applied within crusher gradients, confer [6], which are additional sequence elements surrounding the RF pulse. These crusher gradients cannot be modeled by the depicted $\hat{f}(u)$. It seems that the only practical way to model tracking terms for refocusing pulses with ideal crusher gradients is to define them in the spin domain, confer [6, 30]. Therefore, we choose an equidistant time grid $t_k = (k-1)\tau$, $k = 1, \dots, N_t$ with $N_t = 270$ points and step size $\tau = T/(N_t - 1) = 0.01$ [ms], together with piecewise constant w and control u with values w_m, u_m, $m = 1, \dots, N_t - 1$. This implies that the magnetic field $\mathbf{B}$ is piecewise constant. It is well-known that for piecewise constant magnetic field the Bloch equations (27) in a spatial point x_0 can be solved analytically as a sequence of rotations. This is expressed by using the Cayley-Klein parameters $(a_m), (b_m) \in \mathbb{C}$, $m = 1, \dots, N_t - 1$ with evolution

$$a_m = \alpha_m a_{m-1} - \beta_m^* b_{m-1}, \qquad b_m = \beta_m a_{m-1} + \alpha_m^* b_{m-1},$$

and with initial conditions $a_0 = 1$, $b_0 = 0$, confer [26]. For the formula relating a_m, b_m and $\mathbf{M}(t_{m-1}, x_0)$ confer [6, eq.(2.15)]. The coefficients a_m, b_m are given by

$$\alpha_m = \cos(\phi_m/2) + i\gamma\tau x_0 w_m \sin(\phi_m/2)/\phi_m, \qquad \beta_m = i\gamma\tau u_m \sin(\phi_m/2)/\phi_m,$$

with $\phi_m = -\gamma\tau\sqrt{u_m^2 + (x_0 w_m)^2}$. Since it is well-known that perfect refocusing with ideal crusher gradients is obtained through $|b(T,x)|^2 = \chi_{\Omega_{\text{in}}}(x)$ for a.e. $x \in \Omega$, the tracking term is given by

$$\hat{f}(u) = \frac{1}{2}\| \, |b(T,x)|^2 - \chi_{\Omega_{\text{in}}}(x)\|^2_{L^2(\Omega)}. \tag{28}$$

Note that $b = b(u)$. In the numerical experiments we use $\hat{f}$ as defined in (28). The adjoint equation and the reduced gradient for this formulation are derived in the appendix of [30]. The spatial domain is discretized equidistantly in $N_x = 481$ points.

In accordance with the presentation in Section 3.3 we use a control-reduced problem formulation for (26–27). For the problem at hand this bears the advantage to iterate only on the small control

vector with $N_t - 1 = 269$ entries, but not on the large state vector with $3N_xN_t = 389610$ entries. Consequently, B_k is a rather small matrix of format 269×269, and the linear algebra operations for its update and evaluation are cheap. In effect, the numerical effort is dominated clearly by the state and adjoint solves; concrete values are reported below. We stress that the use of control-reduced formulations is common practice in optimal control.

We consider the same four optimization methods as in Section 5.1, i.e., a semismooth Newton method and the hybrid method with the Broyden, SR1 and BFGS update, respectively. If not mentioned otherwise, these methods are globalized with tr-cg. The stopping criterion is a relative tolerance of 10^{-5} for the residual norm $\|H\|_{L^2(I)}$. Unless declared otherwise, the following settings are applied: a limit of $L = 75$ for all methods, $B_0 = 0$ for Broyden and SR1 methods, and $B_0 = \alpha I$ for BFGS. These initializations are selected because they turned out to be the most effective for the respective methods on this problem. Note that this agrees with the numerical results for the optimal control of the heat equation in Section 5.1, cf. Table 1.

5.3.1. Comparison of the optimization methods

The optimization is initialized with a sinc-shaped RF pulse $f^0(t) = 1.8 \cdot \mathrm{sinc}(-2.2 + 4t/T)$. To maintain a good initial slice profile we use $q^0 = f^0 + \mathrm{sign}(f^0)\beta/\alpha$, which implies $f^0 = \sigma(q^0)$ and $u^0 = \Pi_{U_{\mathrm{ad}}}(\sigma(q^0)) = \Pi_{U_{\mathrm{ad}}}(f^0)$ using again the soft-shrinkage operator σ introduced in Lemma 3.9. The initialization and the corresponding slice profile $M_3(T, x)$ with the desired slice profile are depicted in Figure 3, the optimal controls for different α, β are depicted in Figure 4. The sparsity and bound properties of the solutions for different α, β are depicted in Table 7. If not mentioned otherwise, then we use $\alpha = 5 \cdot 10^{-4}$ and $\beta = 10^{-4}$ below. In all experiments we monitor that only runs leading to the same local minimizer are compared, which is important since the problem possesses several different minimizers. The topics of different minimizers and convergence from random initials q^0 are discussed in the last study.

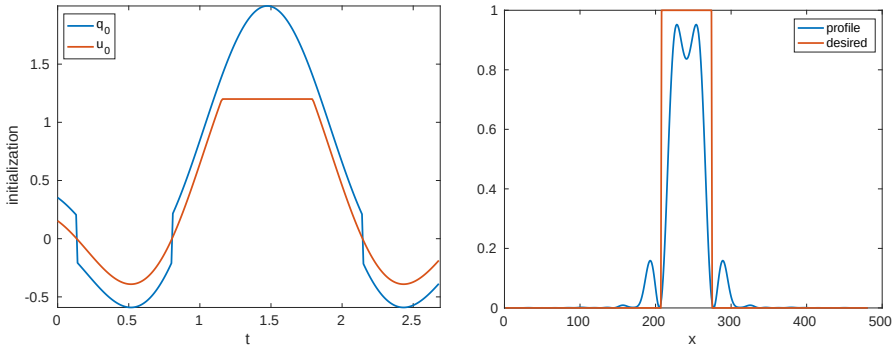

Figure 3: Initial q^0 and $u^0 = \Pi_{U_{\mathrm{ad}}}(\sigma(q^0))$ (left) and the corresponding slice profile $M_3(T, x)$ compared to the desired slice profile (right).

	$\lvert O\rvert$					$\lvert\mathcal{A}\rvert$				
$\alpha \setminus \beta$	$5 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	10^{-4}	$5 \cdot 10^{-5}$	10^{-5}	$5 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	10^{-4}	$5 \cdot 10^{-5}$	10^{-5}
10^{-3}	73	58	37	31	11	47	47	44	43	42
$5 \cdot 10^{-4}$	93	76	20	5	1	57	57	54	51	50
10^{-4}	104	75	12	8	3	68	66	65	65	67
$5 \cdot 10^{-5}$	106	75	14	10	4	70	69	71	75	101

Table 7: Number of time instances with sparsity ($|O|$) or active box constraint ($|\mathcal{A}|$) out of 269 for different α (rows) and β (five columns each).

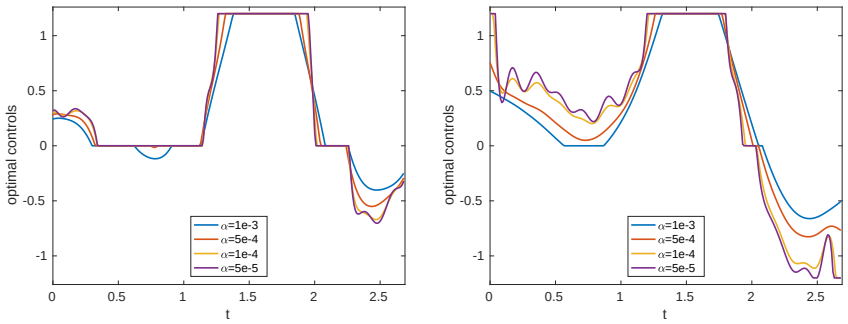

Figure 4: Optimal controls for different α, β computed with the semismooth Newton method. On the left $\beta = 5 \cdot 10^{-4}$, on the right $\beta = 5 \cdot 10^{-5}$.

The first study compares the performance of four different semismooth Newton-type methods embedded in a trust-region cg framework for varying α, β and for different initializations of the quasi-Newton matrix B_0. First, the limited-memory methods are analyzed in Table 8 and compared to the semismooth Newton method. The up to four columns per method show the iteration counts for different B_0. The last row shows the mean value per column taken over the converged runs only. The runs that do not converge are marked with $-$. The first three columns display the parameters α, β and the iteration number of SN. The next two column groups show that the hybrid method with Broyden updates behaves quite similar to the variant with SR1 updates, the latter often requiring slightly fewer iterations. The choices $B_0 = 0$ and $B_0 = \alpha I$ yield fast convergence throughout all (α, β)-pairings, while the other two formulas turn out to be less efficient in this setting. Looking at the hybrid method with BFGS in the last column group we observe that its performance degenerates for large values of α. Apart from this surprising phenomenon the scaled identity yields good results also for BFGS. Using the best choice B_0 for each method, Broyden requires 41 iterations on average, SR1 35 and BFGS 83, compared to 16 for the semismooth Newton method. Since the latter has much more costly iterations due to the forward-backward solve of the second-order equations, it is important to compare the corresponding runtimes. They are included below.

α	β	SN	Broyden				SR1				BFGS		
			αI	$\frac{y^T s}{s^T s} I$	$\frac{y^T y}{y^T s} I$	0	αI	$\frac{y^T s}{s^T s} I$	$\frac{y^T y}{y^T s} I$	0	αI	$\frac{y^T s}{s^T s} I$	$\frac{y^T y}{y^T s} I$
10^{-3}	$3 \cdot 10^{-4}$	9	44	102	–	28	33	84	–	28	296	222	–
$5 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	11	29	97	–	38	31	143	–	27	137	143	–
10^{-4}	$3 \cdot 10^{-4}$	28	52	206	–	51	48	–	–	49	70	90	128
$5 \cdot 10^{-5}$	$3 \cdot 10^{-4}$	32	64	243	–	61	52	–	–	61	67	76	149
10^{-3}	10^{-4}	9	43	70	–	23	27	71	–	23	–	122	–
$5 \cdot 10^{-4}$	10^{-4}	9	32	65	–	35	30	58	–	30	58	84	211
10^{-4}	10^{-4}	21	40	101	174	48	43	–	–	39	44	58	76
$5 \cdot 10^{-5}$	10^{-4}	30	51	145	–	56	49	–	–	54	54	63	89
10^{-3}	$5 \cdot 10^{-5}$	9	27	59	–	22	28	53	–	23	104	98	–
$5 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	15	34	53	106	38	27	65	170	23	39	55	298
10^{-4}	$5 \cdot 10^{-5}$	16	44	104	135	50	40	131	–	41	41	46	71
$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	21	49	115	284	54	46	265	–	49	49	46	90
10^{-3}	10^{-5}	10	27	52	–	22	20	66	174	19	218	83	–
$5 \cdot 10^{-4}$	10^{-5}	7	31	69	86	33	33	67	162	23	44	38	72
10^{-4}	10^{-5}	16	37	78	–	50	33	191	–	34	37	39	60
$5 \cdot 10^{-5}$	10^{-5}	15	46	124	168	46	45	223	–	42	49	52	78
	$\varnothing$	16	41	105	159	41	38	109	127	35	87	120	83

Table 8: Iteration numbers of the limited-memory trust-region quasi-Newton implementations with $L = 75$ for different α, β (rows) and for $B_0 = \alpha I$, $B_0 = \frac{y^T s}{s^T s} I$, $B_0 = \frac{y^T y}{y^T s} I$, $B_0 = 0$. The symbol – indicates that the relative tolerance is not met within 300 iterations. The last line depicts the mean iteration counts per column for the converged runs.

We turn to the hybrid methods with full quasi-Newton implementations in Table 9. First note that all three full quasi-Newton methods yield basically the same results for $B_0 = \alpha I$ and $B_0 = 0$ as the corresponding limited-memory implementations (which use $L = 75$). Additionally, an exact initialization $B_0 = \nabla^2 \hat{f}(u^0)$ via $N_t - 1$ evaluations of the Hessian direction is formed. This initialization yields small iteration counts, too. However, in view of the high costs of the exact initialization, the use of $B_0 = 0$ or $B_0 = \alpha I$ is preferable. Based on these results only the limited-memory variants of Broyden and SR1 with $B_0 = 0$ as well as of BFGS with $B_0 = \alpha I$ are investigated further.

For the limited-memory implementations the question arises how to set the limit parameter L adequately. To address this issue we compare the performance of the hybrid methods for different limits in Table 10 based on the iteration counts. Depicted are four columns per method which differ in the choice of the limit ranging from $L = 25$ to $L = 100$. The rows show results for different α while keeping $\beta = 10^{-4}$ fixed. We stress that in combination with tr-cg it is appropriate to choose a limit L that is larger than typical values from the literature for globalization by line search methods. This is due to the fact that Steihaug-cg employs earlier breaks in the cg method leading to smaller and more steps. We observe that a limit of 25 is only adequate for Broyden and SR1 in the case of large $\alpha \geq 5 \cdot 10^{-4}$. For smaller α the limit should be increased to 50. In contrast, the performance of BFGS is less sensitive to the values of L that are investigated.

In the particular setting of optimal control problems with many state variables and few control variables, it pays off in runtime to use larger limits since this helps to save some iterations of the trust-region method while it increases the required time per trust-region iteration only marginally. This is due to the fact that the costs per trust-region iteration are largely dominated

α	β	SN	Broyden				SR1				BFGS		
			$\nabla^2 \hat{f}(u^0)$	I	αI	0	$\nabla^2 \hat{f}(u^0)$	I	αI	0	$\nabla^2 \hat{f}(u^0)$	I	αI
10^{-3}	$3 \cdot 10^{-4}$	9	41	–	47	28	26	233	33	28	–	–	298
$5 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	11	33	–	28	37	26	260	31	27	113	–	140
10^{-4}	$3 \cdot 10^{-4}$	28	53	–	49	57	43	–	48	50	76	–	70
$5 \cdot 10^{-5}$	$3 \cdot 10^{-4}$	32	69	–	60	68	45	–	50	61	81	–	67
10^{-3}	10^{-4}	9	28	–	31	23	21	185	27	23	269	–	–
$5 \cdot 10^{-4}$	10^{-4}	9	31	–	32	35	25	221	30	30	60	–	58
10^{-4}	10^{-4}	21	48	–	40	45	41	–	43	39	55	–	44
$5 \cdot 10^{-5}$	10^{-4}	30	51	–	52	57	53	280	49	54	87	–	54
10^{-3}	$5 \cdot 10^{-5}$	9	29	–	27	22	18	199	28	23	270	–	104
$5 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	15	34	–	34	29	28	148	27	23	53	–	39
10^{-4}	$5 \cdot 10^{-5}$	16	50	–	44	49	41	–	40	41	57	–	41
$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	21	52	–	48	54	43	289	46	49	104	–	49
10^{-3}	10^{-5}	10	26	–	28	22	26	155	20	19	–	–	218
$5 \cdot 10^{-4}$	10^{-5}	7	29	–	31	33	34	192	33	23	64	–	44
10^{-4}	10^{-5}	16	50	–	37	51	45	257	33	34	52	–	37
$5 \cdot 10^{-5}$	10^{-5}	15	45	–	48	46	42	–	44	42	83	–	49
	$\varnothing$	16	42	–	40	41	35	220	36	35	102	–	87

Table 9: Iterations of the full trust-region quasi-Newton implementations for different α, β (rows) and for $B_0 = \nabla^2 \hat{f}(u^0)$, $B_0 = I$, $B_0 = \alpha I$, $B_0 = 0$ (columns).

by the evaluation of objective and gradient. For example, the runtime of Broyden with $\alpha = 5 \cdot 10^{-5}$ for a limit of 75 (56 iterations) is 5.1 seconds, which is significantly lower than the 6.4 seconds that are needed with a limit of 50 (73 iterations). In both cases, around 90% of the runtime is spent on the evaluation of objective and gradient. Therefore, a limit of $L = 75$ is chosen for all subsequent studies. We emphasize that all runs converge to the same local minimizer independently of the limit parameter.

	Broyden				SR1				BFGS			
$\alpha \setminus L$	25	50	75	100	25	50	75	100	25	50	75	100
10^{-3}	23	23	23	23	23	23	23	23	–	–	–	–
$5 \cdot 10^{-4}$	61	35	35	35	69	30	30	30	54	58	58	58
10^{-4}	82	48	48	48	–	39	39	39	53	44	44	44
$5 \cdot 10^{-5}$	127	73	56	56	150	67	54	54	64	54	54	54

Table 10: Iterations of the limited-memory quasi-Newton methods for different α and different limits L. The symbol – stands for not converging within 300 iterations

To investigate mesh independence properties of the hybrid methods we perform runs with different temporal and spatial mesh sizes. The results are shown in Table 11 using iteration counts in the left table, respectively runtime (mean runtime in 5 runs, in seconds) in the right table. The rows depict results for different temporal refinements with $N_c = N_t - 1$ control points. The two columns per method show different spatial grids with $N_x = 481$, respectively, $N_x = 4811$ points. The finest example with $N_c = 17216$ and $N_x = 4811$ features 248 million degrees of freedom for the state variable. In all cases the same initial guess is used. Furthermore, the same local minimizer is attained in all runs, which allows for a direct comparison. The left table shows

that the iteration counts of all four methods do not increase with N_c or N_x. In particular, SN, Broyden and SR1 exhibit nearly the same iteration count for any of the discretizations. The hybrid method with BFGS displays a constant iteration number per column with a reduced iteration count for the right column (larger N_x). Interestingly, its performance is rather similar to that of Broyden and SR1 for $N_x = 4811$, while for $N_x = 481$ it requires roughly twice as many iterations and twice as much runtime as Broyden and SR1.

The right table shows the mean values of five runtimes in seconds, measured for a single CPU without parallelization. Despite their higher iteration counts, the three limited-memory methods have much smaller runtimes than the semismooth Newton method. The bottom line depicts the mean value per column of the runtime in microseconds divided by N_x and N_c, a number that varies only slightly per column (and per row, although not displayed). We regard this quantity as an efficiency index and denote it by $\mathcal{E}$. In contrast, the runtime of the semismooth Newton method increases faster in N_c than linearly. We attribute this to the fact that the cg method requires more iterations for larger systems and that these iterations involve expensive operations for SN. Therefore, the speedup factor of the Broyden variant of the hybrid method over the semismooth Newton method rises with the number of time instances, starting at 7 and reaching 70 for $N_c = 17216$ and $N_x = 481$. Using SR1 updates leads to similar runtimes with a speedup of up to 68. As already seen in the left table, the use of BFGS updates produces higher iteration counts resulting in an increased runtime. Still, a speedup factor of up to 37 over the semismooth Newton method is reached.

	Iteration count								Runtime							
N_c	SN		Broyden		SR1		BFGS		SN		Broyden		SR1		BFGS	
269	10	10	32	34	35	27	89	41	23	181	3	33	3	26	8	39
538	9	9	28	34	29	27	73	38	46	507	7	69	5	52	13	73
1076	9	9	31	33	30	27	55	32	117	1193	13	139	11	114	20	137
2152	9	9	31	39	25	26	54	40	322	3199*	23	317	18	213	40	319
4304	9	9	29	28	29	26	53	36	1106*		43	438	45	407	77	559
8608	9	9	28	34	30	26	53	36	4485*		90	1092	95	841	168	1154
17216	9	9	28	28	27	30	53	36	12755*		183	2120*	188	1796*	341	2390*
∅	9.1	9.1	29.6	32.9	29.3	27	61.4	37		$\mathcal{E}$	23.2	26.1	21.2	20.6	44.0	28.5

Table 11: Iterations (left table) and runtime in seconds (right table) of the limited-memory quasi-Newton implementations for different numbers of control points N_c (rows) and $N_x = 481, 4811$ (two columns per method). The runtimes are mean values from five runs, respectively, one run if marked with an asterisk. For the efficiency index $\mathcal{E}$ a smaller value indicates greater efficiency

We now take a closer look at the convergence properties of the different quasi-Newton updates in the trust-region method. To this end, the solution is first computed in high precision with the semismooth Newton method and a relative tolerance of 10^{-13}. The error norms are defined as in the first example. Then the different optimization runs are performed with a relative tolerance of 10^{-7}, measured in $\|H(\cdot)\|_{L^2(I)}$. The results are displayed in Table 12. The table shows that all methods are capable of reducing all six errors to approximately the size of the relative tolerance. The semismooth Newton method converges in three steps as soon as a local neighborhood of the minimizer is reached. The other methods need significantly more steps for the local convergence at the end of the optimization run. It seems likely that this is due to the strong nonconvexity of the bilinear control problem at hand. Moreover, the table confirms that $e^k_{u,L^2} \leq e^k_{q,L^2}$ and $e^k_{u,L^\infty} \leq e^k_{q,L^\infty}$ are valid, cf. Theorem 3.7 part 2) and Theorem 3.11 part 2). It also shows that this relationship is not satisfied with respect to the $H^1(I)$-norm.

SN						Broyden					
e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}	e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}
$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$4 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$6 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$2 \cdot 10^{-5}$	$6 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$
$6 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	$2 \cdot 10^{-2}$	$6 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$4 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$
$1 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$4 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$8 \cdot 10^{-8}$	$1 \cdot 10^{-7}$	$2 \cdot 10^{-6}$	$8 \cdot 10^{-7}$	$1 \cdot 10^{-6}$	$1 \cdot 10^{-6}$
$2 \cdot 10^{-9}$	$3 \cdot 10^{-9}$	$4 \cdot 10^{-8}$	$2 \cdot 10^{-8}$	$3 \cdot 10^{-8}$	$5 \cdot 10^{-8}$	$7 \cdot 10^{-9}$	$9 \cdot 10^{-9}$	$2 \cdot 10^{-7}$	$8 \cdot 10^{-8}$	$1 \cdot 10^{-7}$	$1 \cdot 10^{-7}$

SR1						BFGS					
e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}	e^k_{u,L^2}	e^k_{q,L^2}	e^k_{u,H^1}	e^k_{q,H^1}	e^k_{u,L^∞}	e^k_{q,L^∞}
$8 \cdot 10^{-7}$	$2 \cdot 10^{-6}$	$3 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$4 \cdot 10^{-5}$	$5 \cdot 10^{-8}$	$8 \cdot 10^{-8}$	$2 \cdot 10^{-6}$	$6 \cdot 10^{-7}$	$1 \cdot 10^{-6}$	$1 \cdot 10^{-6}$
$3 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$1 \cdot 10^{-5}$	$8 \cdot 10^{-6}$	$6 \cdot 10^{-6}$	$2 \cdot 10^{-5}$	$4 \cdot 10^{-8}$	$7 \cdot 10^{-8}$	$1 \cdot 10^{-6}$	$5 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$1 \cdot 10^{-6}$
$8 \cdot 10^{-8}$	$1 \cdot 10^{-7}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$1 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$4 \cdot 10^{-8}$	$6 \cdot 10^{-8}$	$1 \cdot 10^{-6}$	$4 \cdot 10^{-7}$	$7 \cdot 10^{-7}$	$9 \cdot 10^{-7}$
$1 \cdot 10^{-8}$	$3 \cdot 10^{-8}$	$2 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$6 \cdot 10^{-7}$	$3 \cdot 10^{-8}$	$5 \cdot 10^{-8}$	$1 \cdot 10^{-6}$	$3 \cdot 10^{-7}$	$6 \cdot 10^{-7}$	$8 \cdot 10^{-7}$

Table 12: Convergence of the hybrid limited-memory methods in different norms: The six columns per method show $e^k_{u,L^2} = \|u_k - \bar{u}\|_{L^2}$, $e^k_{q,L^2} = \|q^k - \bar{q}\|_{L^2}$, $e^k_{u,H^1} = \|u^k - \bar{u}\|_{H^1}$, $e^k_{q,H^1} = \|q^k - \bar{q}\|_{H^1}$, $e^k_{u,L^\infty} = \|u^k - \bar{u}\|_{L^\infty}$, and $e^k_{q,L^\infty} = \|q^k - \bar{q}\|_{L^\infty}$ for the last four steps of the optimization run.

5.3.2. Comparison of the globalization techniques

This section is devoted to comparing different globalization techniques for the hybrid methods. This topic is particularly relevant for the bilinear control problem at hand because it is expected to posses several local minimizers. We analyze the performance of the three globalizations tr-cg, ls-GMRES and nls-GMRES. They are paired with the semismooth Newton method and the three limited-memory quasi-Newton methods. For each of these twelve combinations, 2000 optimization runs from a random initial (MATLAB `rand`) q^0 are performed on the small-scale example with $N_c = 269$ and $N_x = 481$. The monotone line search operates on the residual $\|H(\cdot)\|_{L^2(I)}$, while the non-monotone line search is tested with $N_{\text{ls}} = 2, 3, 4, 5$ based on the objective ($R(u) = f(u) + \varphi(u)$) and based on the residual ($R(q) = \|H(q)\|_{L^2(I)}$). Due to space limitations we show only the best results, which are obtained with $N_{\text{ls}} = 2$ and $R(u) = f(u) + \varphi(u)$. The maximum iteration number is set to 300 for all optimization methods and the relative tolerance is set to 10^{-4}. For cg and GMRES a tolerance of 10^{-10} and a maximum iteration number of 100 is used.

Throughout the 8000 optimization runs with tr-cg twelve different stationary points are observed, whose controls are depicted in Figure 5, divided into three sets (top row). Every set of four controls yields identical optimal values $\bar{m} := f(\bar{u}) + \varphi(\bar{u})$, control norms $\|\bar{u}\|_U$, and final magnetization (bottom row). The four controls are related by axial symmetry to the t-axis and the axis $t = T/2$. From an application point of view these four optimal controls are therefore equivalent. Thus, we introduce the multiplicity of a minimizer and count these four solutions henceforth as one local minimizer. In this way we end up with three local minimizers in total. The relative occurrences of these three minimizers are depicted in the upper three rows of Table 13 in columns five to eight. The lower part of the table additionally displays the relative occurrences of runs that do not satisfy the termination criterion, i.e., that do not reach the prescribed relative tolerance within 300 iterations; they are labeled "not converged". The average objective value returned by the optimizer $\varnothing\ \bar{m}$ and the average runtime [s] are also shown. In contrast to previous experiments they are taken over all runs here, i.e., they include also the "not converged" runs. In particular, we observe that SN, Br and SR always meet the relative tolerance. In contrast,

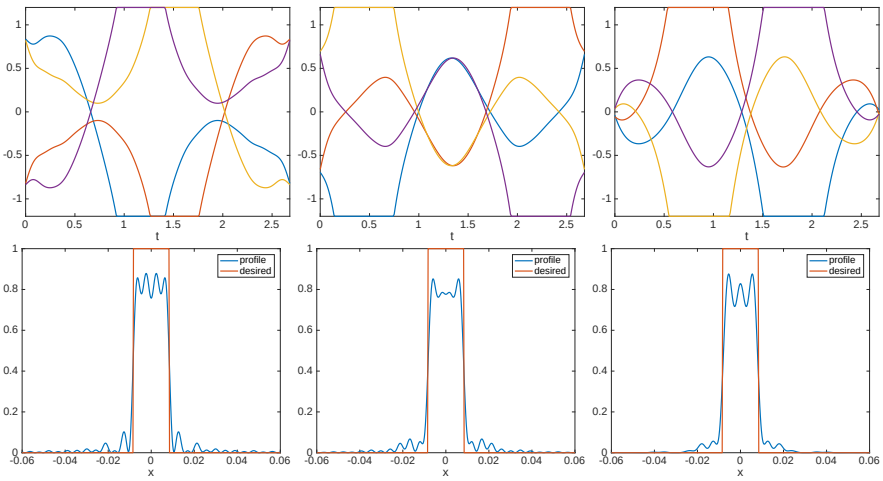

Figure 5: Four optimized controls with the same objective value (top) and their slice profiles (bottom, desired state in red) for the three best observed candidates $\bar{m} \cdot 10^3 = 0.61906$ (left), $\bar{m} \cdot 10^3 = 0.64959$ (mid), and $\bar{m} \cdot 10^3 = 0.67280$ (right).

one seventh of the BFGS runs does not converge. In fact, these runs yield the same twelve minimizers but fail to reach the prescribed tolerance, which is underlined by the agreeing values of $\varnothing\, \bar{m}$. This mean objective value is $\varnothing\, \bar{m} = 0.64 \cdot 10^{-3}$ for all four methods with tr-cg, which is significantly smaller than those achieved with the other globalization techniques. The mean runtime shows a clear speedup of the hybrid methods compared to SN, despite the fact that this is a small scale example with $N_t = 270$ and $N_x = 481$.

In contrast to the trust-region method, the line search globalizations find many more stationary points, 78 in total. However, 49 of these have a prohibitively high objective value. They are summarized in lines number 10 and 11 of Table 13. The results with monotone line search (ls-GMRES) are depicted in the fourth column group of Table 13. We observe that the semismooth Newton method with a basic monotone line search on the residual $\|H(\cdot)\|_{L^2(I)}$ quickly converges to a noncompetitive minimizer in nearly all cases. The three quasi-Newton methods yield smaller objective values in the mean, but most of the runs fail to match the prescribed relative tolerance. The mean optimal values of all four methods are much larger than those obtained with tr-cg.

The non-monotone line search nls-GMRES is more effective than ls-GMRES for all three quasi-Newton methods, see the last column group of Table 13. We note that the number of runs that do not converge is smaller than for the monotone line search. In particular, Broyden and BFGS converge in most of the cases. Moreover, the best control and the top three controls are found more often leading to much better average optimal values compared to ls-GMRES. However, excellent values similar to those of tr-cg are attained only for BFGS. Notably, the semismooth Newton method does not benefit from the non-monotone line search; it behaves similarly as with ls-GMRES. It is also worth mentioning that the average runtimes of the tr-cg hybrid quasi-Newton methods are only slightly above those of the nls-GMRES variant.

Summarizing, in this application problem the tr-cg globalization robustly delivers the top-three candidates for all four optimization methods. In contrast, the line search globalizations often have difficulties with convergence for the quasi-Newton methods, and tend to noncompetitive solutions for the semismooth Newton method. Thus, for optimal control of the Bloch equations tr-cg should clearly be preferred over a line search globalization in the case of the semismooth Newton method and the hybrid Broyden or SR1 method. For BFGS, both tr-cg and nls-GMRES work equally well.

nr	mult	$\bar{m} \cdot 10^3$	$\|\bar{u}\|_{L^2}$	% (tr-cg)				% (ls-GMRES)				% (nls-GMRES)			
				SN	Br	SR	BF	SN	Br	SR	BF	SN	Br	SR	BF
1	4	0.61906	1.1961	41.6	36.7	38.4	37.3		11.1	0.2	7.7	0.2	39.6	18.8	44.8
2	4	0.64959	1.1367	45.2	43.8	43.7	40.9		3.8	0.2	1.9	0.1	20.9	7.8	31.7
3	4	0.67280	1.1215	13.2	19.6	17.9	7.7		1.1	0.1	1.0		6.7	4.3	6.6
4	2	0.67608	1.1020									0.1		0.1	
5	3	0.69636	1.1488					0.1	0.2			0.1			
6	3	1.11919	1.1522						0.1				0.1	0.5	
7	1	1.57083	1.0094									0.1		0.1	
8	4	1.61472	1.0902						0.1			0.1		0.2	
9	4	1.82260	1.6886										1.4	0.3	0.2
10	36	$\in [2, 3]$						0.1	4.2	0.3	1.9		14.6	2.8	4.8
11	13	> 3						99.2				88.4			
		not converged					14.2	0.7	79.5	99.2	87.6	11.1	16.9	65.3	11.9
		$\varnothing\ \bar{m} \cdot 10^3$		0.64	0.64	0.64	0.64	4.18	2.78	2.78	2.78	4.14	0.94	1.17	0.72
		$\varnothing$ runtime		23	4	3	6	19	3	2	3	19	3	4	5

Table 13: Quality of the solutions for different optimization and globalization methods for $\alpha = 5 \cdot 10^{-4}$, $\beta = 10^{-4}$. 2000 optimization runs from randomly chosen initials q^0 are performed for each combination. Depicted are the multiplicity, the optimal value $\bar{m}$, and the norm of the optimal control. The next three column groups show the relative occurrence (%) of the respective solution, separately for tr-cg, ls-GMRES and nls-GMRES. Each column group is divided into the four methods semismooth Newton (SN), Broyden (Br), SR1 (SR), and BFGS (BF). The three bottom lines depict the percentage of runs that did not converge, the average objective value $\varnothing\ \bar{m}$ that was returned by the optimizer, and the average runtime, both as average over all runs.

6. Conclusions

In this paper we have demonstrated that the hybrid approach developed in [20] is widely applicable and leads to competitive algorithms that can solve large-scale real-world nonsmooth and nonconvex PDE-constrained optimal control problems in a fraction of the time that semismooth Newton methods require for this task. In particular, we observed this to be true for a matrix-free limited-memory truncated trust-region variant of the hybrid approach, a method that we believe to hold tremendous potential for nonsmooth large-scale optimization problems.

All in all, we conclude from this and the preceding paper [20] that the novel approach is widely applicable, has favorable theoretical properties, and leads to highly efficient numerical schemes.

A. Trust-region globalization of the hybrid method

We state the precise algorithm of tr-cg that is employed in the numerical experiments. It is designed for solving (POR) from Section 3.2.2 and uses the notation of that section. The objective of (POR) is denoted by $J : U \to \mathbb{R}$, i.e., $J(u) := f(u) + \varphi(u) = \hat{f}(u) + \frac{\gamma}{2}\|u\|_U^2 + \varphi(u)$.

Algorithm 2: Hybrid semismooth quasi-Newton-cg method with trust-region globalization

Input: $0 < \mathrm{tol}_{\mathrm{tr}}, \mathrm{tol}_{\mathrm{cg}} \ll 1$; $\mathrm{maxit}_{\mathrm{tr}}, \mathrm{maxit}_{\mathrm{cg}} \in \mathbb{N}$; initial guess (q^0, B_0); trust-region parameters $0 < \varrho_0 \le \varrho_{\max}$, $0 < \sigma_1 < \sigma_2 < \sigma_3 < 1$, $0 < f_1, f_3 < 1 < f_2$

Set $k = 0, \varrho = \varrho_0$; compute $H(q^k)$; choose $M \in \partial G(q^k)$, $\hat{M} \in \partial \hat{G}(q^k)$
Define $\langle x, y\rangle = (x, My)_U$
while $\left[\|H(q^k)\| > \mathrm{tol}_{\mathrm{tr}}\|H(q^0)\| \quad \textit{and} \quad k \le \mathrm{maxit}_{\mathrm{tr}}\right]$ **do** // trust-region loop
  Set $p^0 = r^0 = -H(q^k)$, $\delta q = 0$, $i = 0$
  while $\left[\|r^i\| > \mathrm{tol}_{\mathrm{cg}}\|r^0\| \quad \textit{and} \quad i \le \mathrm{maxit}_{\mathrm{cg}}\right]$ **do** // Steihaug-cg loop
    Set $\tilde{M} = B_k M + \hat{M}$; compute $\tilde{M}p^i$
    if $\langle p^i, \tilde{M}p^i\rangle \le 0$ **then** // negative curvature
      Compute $\max\{\tau : \|\delta q + \tau p^i\| \le \varrho\}$ // go to boundary of trust-region
      Set $\delta q = \delta q + \tau p^i$
      break
    end
    Compute $\alpha = \|r^i\|/\langle p^i, \tilde{M}p^i\rangle$
    if $\|\delta q + \alpha p^i\| \ge \varrho$ **then** // step too large
      Compute $\max\{\tau : \|\delta q + \tau p^i\| \le \varrho\}$ // go to boundary of trust-region
      Set $\delta q = \delta q + \tau p^i$
      break
    end
    Set $r^{i+1} = r^i - \alpha\tilde{M}p^i$
    Set $p^{i+1} = r^{i+1} + \|r^{i+1}\|^2/\|r^i\|^2 p^i$
    Set $\delta q = \delta q + \alpha p^i$, $\quad i = i + 1$
  end
  Compute $\delta J_a = J(G(q^k)) - J(G(q^k + \delta q))$ // actual decrease
  Compute $\delta J_m = -\frac{1}{2}\langle \delta q, \tilde{M}\delta q\rangle - \langle \delta q, H(q^k)\rangle$ // predicted decrease
  if $\left[\delta J_a > \varepsilon \quad \textit{and} \quad \delta J_a > \sigma_1 \delta J_m\right]$ **then** // accept step
    Set $q^{k+1} = q^k + \delta q$
    Compute $H(q^{k+1})$; choose $M \in \partial G(q^{k+1})$, $\hat{M} \in \partial \hat{G}(q^{k+1})$
    Define $\langle x, y\rangle = (x, My)_U$
    if $|\delta J_a/\delta J_m - 1| \le 1 - \sigma_3$ **then** // increase radius
      Set $\varrho = \min\{f_2\varrho, \varrho_{\max}\}$
    else if $|\delta J_a/\delta J_m - 1| > 1 - \sigma_2$ **then** // decrease radius
      Set $\varrho = f_3\varrho$
    end
  else
    Set $\varrho = f_1\varrho$, $q^{k+1} = q^k$ // decrease radius
  end
  Set $s_u^k = G(q^{k+1}) - G(q^k)$, $y^k = \nabla\hat{f}(G(q^{k+1})) - \nabla\hat{f}(G(q^k))$
  Compute B_{k+1} by quasi-Newton update
  Set $k = k + 1$
end
Output: q^k

Acknowledgments

This work was supported in part by the Austrian Science Fund (FWF) in the context of the "SFB F32-N18" (Mathematical Optimization and Applications in Biomedical Sciences) in the projects F3201 and F3202.

References

[1] P. M. Anselone. *Collectively compact operator approximation theory and applications to integral equations. With an appendix by Joel Davis.* Prentice-Hall Series in Automatic Computation. Prentice-Hall Inc., 1971.

[2] J. Appell and P. P. Zabrejko. *Nonlinear Superposition Operators.* Cambridge Tracts in Mathematics. Cambridge University Press, 1990. doi:10.1017/CBO9780511897450.

[3] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* Springer, 2nd edition, 2017. doi:10.1007/978-3-319-48311-5.

[4] A. Beck. *First-order methods in optimization.* MOS-SIAM Series on Optimization. SIAM, 2017. doi:10.1137/1.9781611974997.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009. doi:10.1137/080716542.

[6] M. A. Bernstein, K. F. King, and X. J. Zhou. *Handbook of MRI Pulse Sequences.* Elsevier Academic Press, 2004. doi:10.1016/B978-012092861-3/50003-0.

[7] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Math. Program.*, 63(1 (B)):129–156, 1994. doi:10.1007/BF01582063.

[8] M. Chipot. *Elements of nonlinear analysis.* Birkhäuser, 2000. doi:10.1007/978-3-0348-8428-0.

[9] C. Clason, F. Kruse, and K. Kunisch. Total variation regularization of multi-material topology optimization. *ESAIM: M2AN*, 52(1):275–303, 2018. doi:10.1051/m2an/2017061.

[10] J. C. De los Reyes. *Numerical PDE-constrained optimization.* Springer, 2015. doi:10.1007/978-3-319-13395-9.

[11] I. Ekeland and R. Témam. *Convex analysis and variational problems.* SIAM, 1999. doi:10.1137/1.9781611971088.

[12] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems. Vol. I.* Springer, 2003. doi:10.1007/b97543.

[13] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems. Vol. II.* Springer, 2003. doi:10.1007/b97544.

[14] W. A. Grissom, K. Setsompop, S. A. Hurley, J. Tsao, J. V. Velikina, and A. A. Samsonov. Advancing RF pulse design using an open-competition format: Report from the 2015 ISMRM challenge. *Magnetic Resonance in Medicine*, 78(4):1352–1361. doi:10.1002/mrm.26512.

[15] K. Gröger. A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations. *Math. Ann.*, 283(4):679–687, 1989. doi:10.1007/BF01442860.

[16] P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. *Math. Program.*, 48(2 (B)):161–220, 1990. doi:10.1007/BF01582255.

[17] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009. doi:10.1007/978-1-4020-8839-1.

[18] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. SIAM, 2008. doi:10.1137/1.9780898718614.

[19] D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, reprint of the 1980 original edition, 2000. doi:10.1137/1.9780898719451.

[20] F. Mannel and A. Rund. A hybrid semismooth quasi-Newton method. Part 1: Theory. *Submitted (2018); preprint available at https://imsc.uni-graz.at/mannel/sqn1.pdf*.

[21] A. M. Milzarek. *Numerical Methods and Second Order Theory for Nonsmooth Problems*. PhD Thesis, Technische Universität München, Munich, 2016. URL: https://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20160712-1289514-1-6.

[22] J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. Fr.*, 93:273–299, 1965. doi:10.24033/bsmf.1625.

[23] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2006. doi:10.1007/978-0-387-40065-5.

[24] M. Noor and K. Noor. On general quasi-variational inequalities. *Journal of King Saud University*, 24, 2010. doi:10.1016/j.jksus.2010.07.002.

[25] N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014. doi:10.1561/2400000003.

[26] J. Pauly, P. Le Roux, D. Nishimura, and A. Macovski. Parameter relations for the Shinnar–Le Roux selective excitation pulse design algorithm. *IEEE Transactions on Medical Imaging*, 10(1):53–65, 1991. doi:10.1109/42.75611.

[27] K. Pieper. *Finite element discretization and efficient numerical solution of elliptic and parabolic sparse control problems*. PhD Thesis, Technische Universität München, Munich, 2015. URL: https://nbn-resolving.de/urn/resolver.pl?nbn:de:bvb:91-diss-20150420-1241413-1-4.

[28] L. Qi and J. Sun. A nonsmooth version of Newton's method. *Math. Program.*, 58(3 (A)):353–367, 1993. doi:10.1007/BF01581275.

[29] S. M. Robinson. Normal maps induced by linear transformations. *Math. Oper. Res.*, 17(3):691–714, 1992. doi:10.1287/moor.17.3.691.

[30] A. Rund, C. Aigner, K. Kunisch, and R. Stollberger. Magnetic resonance RF pulse design by optimal control with physical constraints. *IEEE Transactions on Medical Imaging*, 37(2):461–472, 2018. doi:10.1109/TMI.2017.2758391.

[31] A. Rund, C. S. Aigner, K. Kunisch, and R. Stollberger. Simultaneous multislice refocusing via time optimal control. *Magnetic Resonance in Medicine*, 80(4):1416–1428, 2018. doi:10.1002/mrm.27124.

[32] A. Schiela. A simplified approach to semismooth Newton methods in function space. *SIAM J. Optim.*, 19(3):1417–1432, 2008. doi:10.1137/060674375.

[33] M. V. Solodov. Merit functions and error bounds for generalized variational inequalities. *J. Math. Anal. Appl.*, 287(2):405–414, 2003. doi:10.1016/S0022-247X(02)00554-1.

[34] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20:626–637, 1983. doi:10.1137/0720042.

[35] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B*, 58(1):267–288, 1996. URL: http://www.jstor.org/stable/2346178.

[36] F. Tröltzsch. *Optimal control of partial differential equations. Theory, methods and applications*, volume 112. AMS, 2010. doi:10.1090/gsm/112.

[37] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM Series on Optimization. SIAM, 2011. doi:10.1137/1.9781611970692.

[38] X. Xiao, Y. Li, Z. Wen, and L. Zhang. A regularized semi-smooth newton method with projection steps for composite convex programs. *Journal of Scientific Computing*, 76(1):364–389, 2018. doi:10.1007/s10915-017-0624-3.