Computational Optimization and Applications (2021) 80:639–677
https://doi.org/10.1007/s10589-021-00308-0

A proximal gradient method for control problems with non-smooth and non-convex control cost

Carolin Natemeyer[1] · Daniel Wachsmuth[1]

Received: 22 July 2020 / Accepted: 4 August 2021 / Published online: 3 September 2021
© The Author(s) 2021

Abstract

We investigate the convergence of the proximal gradient method applied to control problems with non-smooth and non-convex control cost. Here, we focus on control cost functionals that promote sparsity, which includes functionals of L^p-type for $p \in [0, 1)$. We prove stationarity properties of weak limit points of the method. These properties are weaker than those provided by Pontryagin's maximum principle and weaker than L-stationarity.

Keywords Proximal gradient method · Non-smooth and non-convex optimization · Sparse control problems

1 Introduction

In this article, we consider a possibly non-smooth optimal control problem of type

$$\min_{u \in L^2(\Omega)} f(u) + \int_\Omega g(u(x))\,\mathrm{d}x, \tag{P}$$

where $\Omega \subset \mathbb{R}^n$ is Lebesgue measurable. The functional $f : L^2(\Omega) \to \mathbb{R}$ is assumed to be smooth. Here, we have in mind to choose $f(u) := f(y(u))$ as the smooth part of an optimal control problem incorporating the state equation and a smooth cost functional. The function $g : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is allowed to be non-convex and non-smooth. Examples include

This research was partially supported by the German Research Foundation DFG under project Grant Wa 3626/3-2.

✉ Daniel Wachsmuth
daniel.wachsmuth@mathematik.uni-wuerzburg.de

Carolin Natemeyer
carolin.natemeyer@mathematik.uni-wuerzburg.de

[1] Institut für Mathematik, Universität Würzburg, 97074 Würzburg, Germany

$$g(u) = |u|^p, \quad p \in (0,1),$$

and

$$g(u) = |u|_0 := \begin{cases} 1 & \text{if } u \neq 0 \\ 0 & \text{if } u = 0. \end{cases}$$

In particular, g is chosen to promote sparsity, that is, local solutions of (P) are zero on a significant part of Ω. We will make the assumptions on the ingredients of the control problem precise below in Sect. 2.

Due to lack of convexity of g, the resulting integral functional $j(u) := \int_\Omega g(u(x))\,\mathrm{d}x$ is not weakly lower semicontinuous in $L^2(\Omega)$, so it is impossible to prove existence of solutions of (P) by the direct method of the calculus of variations. Still it is possible to prove that the Pontryagin maximum principle is a necessary optimality condition. This principle does not require differentiability of g. In this paper, we will address the question whether weak limit points of the proposed optimization method satisfy the maximum principle or weaker conditions.

In order to guarantee existence of solutions one has to modify the problem, e.g., by introducing some compactness. This is done in [18], where a regularization term of the type $\frac{\alpha}{2}\|u\|^2_{H^1}$ is added to the functional in (P). These regularized problems are solvable. However, the maximum principle cannot be applied anymore. In addition, due to the non-local nature of H^1-optimization problems, it is much more difficult to compute solutions numerically. Convergence for $\alpha \searrow 0$ of *global* solutions of the regularized problem to solutions of the original problem has been proven in [18], but it is not clear how this can be exploited algorithmically.

In this paper, we propose to use the proximal gradient method (also called forward-backward algorithm [3]) to compute candidates for solutions. The main idea of this method is as follows: Suppose the objective is to minimize a sum $f + j$ of two functions f and j on the Hilbert space H, where f is smooth. Here, we have in mind to choose $H = L^2(\Omega)$ and $j(u) = \int_\Omega g(u(x))\,\mathrm{d}x$.

Given an iterate u_k, the next iterate u_{k+1} is computed as

$$u_{k+1} \in \operatorname*{arg\,min}_{u \in H} \left(f(u_k) + \nabla f(u_k)\cdot(u - u_k) + \frac{L}{2}\|u - u_k\|_H^2 + j(u) \right), \tag{1.1}$$

where $L > 0$ is a proximal parameter, and L^{-1} can be interpreted as a step-size. In our setting, the functional to be minimized in each step is an integral function, whose minima can be computed by minimizing the integrand pointwise. Let us introduce the so-called prox-map, which is defined by

$$\operatorname{prox}_{\gamma j}(z) := \operatorname*{arg\,min}_{x \in H} \left(\frac{1}{2}\|x - z\|_H^2 + \gamma j(x) \right), \tag{1.2}$$

where $\gamma > 0$. If j is weakly lower semicontinuous and bounded from below, then $\operatorname{prox}_{\gamma j}(z)$ is non-empty for all $z \in H$. Let us emphasize that due to the non-convexity of j, the solution set argmin is multi-valued in general, so that $\operatorname{prox}_{\gamma j} : H \rightrightarrows H$ is a set-valued mapping. Then, (1.1) can be written as

$$u_{k+1} \in \operatorname{prox}_{L^{-1}j}\Big(u_k - \frac{1}{L}\nabla f(u_k)\Big).$$

If $j \equiv 0$, the method reduces to the steepest descent method. If j is the indicator function of a convex set, then the method is a gradient projection method. The convergence analysis of this method is based on the following observation: under suitable assumptions on L, the iterates satisfy $\|u_{k+1} - u_k\|_H \to 0$. If f and j are convex, then the convergence properties of the method are well-known: under mild assumptions, the iterates (u_k) converge weakly to a global minimum of $f + j$, see, e.g., [3, Corollary 27.9]. If f is non-convex and H is finite-dimensional, then sequential limit points u^* of (u_k) are stationary, that is, they satisfy

$$-\nabla f(u^*) \in \partial j(u^*), \tag{1.3}$$

where ∂j is the convex subdifferential of j, [5, Theorems 6.39, 10.15]. For infinite-dimensional spaces a similar result can be proven, if one assumes strong convergence (or in the case of $H = L^2(\Omega)$ pointwise convergence almost everywhere) of $\nabla f(u_k)$, see below Remark 4.22. Literature on the convergence analysis of the simple method (1.1) in infinite-dimensional spaces if either f or j is non-convex is relatively scarce. There are results for projected gradient methods, see, e.g., [14, 17]. Recently, a stochastic version of the algorithm was analyzed in [16]. However, in these papers no convergence results for weakly converging subsequences of iterates are given.

If in addition j is non-convex, then much less has been proven. For finite-dimensional problems it has been shown that limit points u^* are fixed points of the iteration, that is

$$u^* \in \operatorname{prox}_{L^{-1}j}\Big(u^* - \frac{1}{L}\nabla f(u^*)\Big). \tag{1.4}$$

Similar results for problems in the space ℓ^2 can be found in [8], where it was shown that weak limit points are fixed points in the sense of (1.4). There, the setting of the problem in ℓ^2 was important, as it could be proven that the active sets $\{n \in \mathbb{N} : u_k(n) \neq 0\}$ only change finitely often, as is the case in finite-dimensional problem. This result is not available for problems on $L^2(\Omega)$, where the underlying measure space is atom-free. In [6] and [4, Chapter 10], points satisfying (1.4) are called L-stationary. For convex and lower semicontinuous j, conditions (1.3) and (1.4) are equivalent. For non-convex j it is natural to consider inclusions of type (1.3), where the convex subdifferential is replaced by some generalized derivative (e.g., Fréchet or limiting subdifferential). Here it turns out that conditions of type (1.3) involving generalized derivatives are weaker than L-stationary. Consider the case $H = \mathbb{R}^1$, and $g(u) = |u|_0$ or $g(u) = |u|^p$ ($p \in (0, 1)$). Then the Fréchet and the limiting subdifferential of g at $u^* = 0$ is equal to $\mathbb{R}$, so the inclusion (1.3) is trivially satisfied. In contrast to this, the L-stationarity condition still gives some information of the following type: if $u^* = 0$ is L-stationary, then $|\nabla f(0)|$ is small, since it can be shown that $0 \in \operatorname{prox}_{L^{-1}j}(q)$ if and only if $|q| \leq q_0$ for some finite q_0, compare Lemmas 3.5 and 3.6 below.

Hence, we are interested in proving that weak limit points in $L^2(\Omega)$ of the proximal gradient method are L-stationary. Unfortunately, weak convergence leads to convexification in the following sense: Let $\mathcal{R} \subset H \times H$ be such that $(u^*, -\nabla f(u^*)) \in \mathcal{R}$ if and only if $(u^*, -\nabla f(u^*))$ satisfies (1.4). The iterates of the method satisfy

$$\big(u_{k+1},\ L(u_{k+1} - u_k) - \nabla f(u_k)\big) \in \mathcal{R}.$$

Let us assume for simplicity that $u_k \rightharpoonup u^*$, $\nabla f(u_k) \to \nabla f(u^*)$, and $u_{k+1} - u_k \to 0$ in H. Passing to the limit will lead to an inclusion $(u^*, \nabla f(u^*)) \in \overline{conv}\,\mathcal{R}$, where $\overline{conv}$ denotes the closed convex hull.

In order to partially prevent this convexification, we will employ an idea of [27]. There the method was analyzed when applied to control problems with L^0-control cost.

An essential ingredient of the analysis in [27] was that the function $g(u) := |u|_0$ is sparsity promoting: solutions of the proximal step (1.1) are either zero or have a positive distance to zero in the following sense: there is $\sigma > 0$ such that $u_{k+1}(x) = 0$ or $|u_{k+1}(x)| \geq \sigma$ for almost all x. In Sect. 3, we investigate conditions on g under which this property can be obtained.

Still this is not enough to conclude L-stationarity of weak limit points. We will show that weak limit points satisfy a weaker condition in general, see Theorem 4.20. Under stronger assumptions on $(\nabla f(u_k))$, L-stationarity can be obtained (Theorems 4.21, 4.23). Pointwise almost everywhere and strong convergence of (u_k) is proven under additional assumptions in Theorem 4.26. We apply these results to $g(u) = |u|^p$, $p \in (0, 1)$ in Sect. 5.1.

Interestingly, the proximal gradient method sketched above is related to algorithms based on proximal minimization of the Hamiltonian in control problems. These algorithms are motivated by Pontryagin's maximum principle. First results for smooth problems can be found in [25]. There, stationarity of pointwise limits of (u_k) was proven. Under weaker conditions it was proved in [7] that the residual in the optimality conditions tends to zero. These results were transferred to control problems with parabolic partial differential equations in [9].

Notation We will frequently use $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. Let $A \subseteq \Omega$ be a set. We define the indicator function of A by

$$\delta_A(x) = \begin{cases} 0 & \text{if } x \in A, \\ +\infty & \text{otherwise,} \end{cases}$$

and the characteristic function of A by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

The convex hull and closed convex hull of the set A is denoted by $\operatorname{conv} A$ and $\overline{\operatorname{conv}} A$, respectively.

For measurable A, we denote the Lebesgue measure of A by $|A|$. We will abbreviate the quantifiers "almost everywhere" and "for almost all" by "a.e." and "f.a.a.", respectively. Let X be a non-empty set. For a given function $F : X \to \bar{\mathbb{R}}$ we define its domain by $\operatorname{dom} F := \{x : F(x) < +\infty\}$. The open ball centered at $x \in \mathbb{R}^n$ with radius $r > 0$ is denoted by $B_r(x)$.

2 Preliminary considerations

2.1 Necessary optimality conditions

In the following we are going to derive a necessary optimality condition for (P), known as Pontryagin maximum principle, where no derivatives of the functional are involved. We formulate the Pontryagin maximum principle (PMP) as in [27]. A control $\bar{u} \in L^2(\Omega)$ satisfies (PMP) if and only if for almost all $x \in \Omega$

$$\nabla f(\bar{u})(x)\bar{u}(x) + g(\bar{u}(x)) \leq \nabla f(\bar{u})(x) \cdot v + g(v) \tag{2.1}$$

holds true for all $v \in \mathbb{R}$. This relation can be rewritten equivalently as

$$\bar{u}(x) \in \operatorname*{arg\,min}_{u\in\mathbb{R}} \, (f(\bar{u}(x)) + \nabla f(\bar{u})(x) \cdot (u - \bar{u}(x)) + g(u)) \quad \text{f.a.a.}\, x \in \Omega.$$

Hence, the iteration (1.1) is nothing else than a fixed point iteration for (2.1) with an additional proximal term. The following result is shown in [27, Thm. 2.5] for the special choice $g(u) := |u|_0$.

Theorem 2.1 (Pontryagin maximum principle) *Let $\bar{u} \in L^\infty(\Omega)$ be a local solution to (P) in $L^2(\Omega)$. Furthermore, assume f satisfies*

$$f(u) - f(\bar{u}) = \nabla f(\bar{u}) \cdot (u - \bar{u}) + o(\|u - \bar{u}\|_{L^1(\Omega)}).$$

Then $\bar{u}$ satisfies the Pontryagin maximum principle (2.1).

Proof We will use needle perturbations of the optimal control. Let $E := \{(v_i, t_i),\ i \in \mathbb{N}\}$ be a countable dense subset of

$$\operatorname{epi}(g) := \{(v, t) \in \mathbb{R} \times \mathbb{R} : g(v) \leq t\}.$$

For arbitrary $x \in \Omega$, $r > 0$, and $i \in \mathbb{N}$ we define $u_{r,i,x} \in L^2(\Omega)$ by

$$u_{r,i,x}(t) := \begin{cases} v_i & t \in B_r(x), \\ \bar{u}(t) & \text{otherwise.} \end{cases}$$

Let $\chi_r := \chi_{B_r(x)}$, then we have $u_{r,i,x} = (1 - \chi_r)\bar{u} + \chi_r v_i$ and

$$\begin{aligned} \|u_{r,i,x} - \bar{u}\|_{L^1(\Omega)} &= \|\chi_r(v_i - \bar{u})\|_{L^1(\Omega)} \leq (|v_i| + \|\bar{u}\|_{L^\infty(\Omega)})\|\chi_r\|_{L^1(\Omega)} \\ &= (|v_i| + \|\bar{u}\|_{L^\infty(\Omega)})|B_r(x)|. \end{aligned}$$

With $j(u) := \int_\Omega g(u(t))\,\mathrm{d}t$ we get

$$
\begin{aligned}
0 &\le f(u_{r,i,x}) + j(u_{r,i,x}) - f(\bar u) - j(\bar u)\\
&= \int_\Omega \nabla f(\bar u)(u_{r,i,x} - \bar u)\,\mathrm{d}t + o(\|u_{r,i,x} - \bar u\|_{L^1(\Omega)}) + \int_\Omega (g(u_{r,i,x}) - g(\bar u))\,\mathrm{d}t\\
&\le \int_{B_r(x)} \nabla f(\bar u)(v_i - \bar u) + (t_i - g(\bar u))\,\mathrm{d}t + o(\|u_{r,i,x} - \bar u\|_{L^1(\Omega)})
\end{aligned}
$$

After dividing above inequality by $|B_r(x)|$ and passing to the limit $r \searrow 0$, we obtain by Lebesgue's differentiation theorem

$$
0 \le \nabla f(\bar u)(x) \cdot (v_i - \bar u(x)) + (t_i - g(\bar u(x))) \tag{2.2}
$$

for every Lebesgue point $x \in \Omega$ of the integrands, i.e., for all $x \in \Omega \setminus N_i$, where N_i is a set of zero Lebesgue measure, on which the above inequality is not satisfied. Since the countable union $\bigcup_{i\in\mathbb{N}} N_i$ is also of measure zero, (2.2) holds true for all $x \in \Omega \setminus \bigcup_i N_i$ for all i. Due to the density of E in $\operatorname{epi}(g)$, we find for $(v, g(v)) \in \operatorname{epi}(g)$ a sequence $(\tilde v_k, \tilde t_k) \to (v, g(v))$ with $(\tilde v_k, \tilde t_k) \in E$, and hence for almost all $x \in \Omega$ it holds

$$
0 \le \nabla f(\bar u)(x) \cdot (v - \bar u(x)) + (g(v) - g(\bar u(x)))
$$

for all $v \in \mathbb{R}$ which is the claim. □

2.2 Standing assumptions

We define the functional $j : L^2(\Omega) \to \bar{\mathbb{R}}$ by

$$
j(u) := \int_\Omega g(u(x))\,\mathrm{d}x,
$$

where we set $j(u) = +\infty$ if $g(u)$ is not integrable. Let us define $\operatorname{dom} j := \{u : j(u) < +\infty\}$.

Throughout the paper, we will assume the following standing assumption on f and g. Another set of structural assumptions on g will be developed in Sect. 3.

Assumption A

(A1) The function $g : \mathbb{R} \to \bar{\mathbb{R}}$ is lower semicontinuous.

(A2) The functional $f : L^2(\Omega) \to \mathbb{R}$ is bounded from below. Moreover, f is Fréchet differentiable and $\nabla f : L^2(\Omega) \to L^2(\Omega)$ is Lipschitz continuous with constant L_f on $\operatorname{dom} j$, i.e.,

$$
\|\nabla f(u_1) - \nabla f(u_2)\|_{L^2(\Omega)} \le L_f \|u_1 - u_2\|_{L^2(\Omega)}
$$

holds for all $u_1, u_2 \in \operatorname{dom} j \subset L^2(\Omega)$.

Here, (A1) implies that g is a normal integrand, and $g(u)$ is measurable for each measurable u, see [15, Section VIII.1.1]. The Lipschitz continuity of the ∇f as in (A2) will be important to prove the basic convergence result Theorem 4.5 below. For $u \in L^2(\Omega)$, we have $\nabla f(u) \in L^2(\Omega)$. With a slight abuse of notation, we will use the notation $\nabla f(u)v := \int_\Omega (\nabla f(u)(x))v(x)\,\mathrm{d}x$ for $v \in L^2(\Omega)$.

The following optimal control example is covered by Assumption A. Let $\Omega_{\mathrm{pde}} \supset \Omega$ be a bounded domain in $\mathbb{R}^n$, $n \le 3$. It will be the domain of the state $y \in H_0^1(\Omega_{\mathrm{pde}})$ associated to the control $u \in L^2(\Omega)$. Let us define

$$f(u) := \int_{\Omega_{\mathrm{pde}}} L(x, y_u(x))\,\mathrm{d}x,$$

where $y_u \in H_0^1(\Omega_{\mathrm{pde}})$ is defined to be the unique weak solution of the elliptic partial differential equation

$$(-\Delta y)(x) + d(x, y(x)) = \chi_\Omega(x)u(x) \quad \text{a.e. in } \Omega_{\mathrm{pde}}.$$

Let us assume that L and d are Carathéodory functions, continuously differentiable with respect to y and such that the derivatives of L, d with respect to y are bounded on bounded sets. In addition, d is assumed to be monotonically increasing with respect to y. Then the mapping $u \mapsto y_u$ is Lipschitz continuous from $L^2(\Omega)$ to $H_0^1(\Omega_{\mathrm{pde}}) \cap L^\infty(\Omega_{\mathrm{pde}})$, see [26, Section 4.5]. The gradient of f is given by $\nabla f(u) = \chi_\Omega p_u$, where $p_u \in H_0^1(\Omega_{\mathrm{pde}})$ is the unique weak solution of the adjoint equation

$$(-\Delta p)(x) + d_y(x, y_u(x))p(x) = L_y(x, y_u(x)) \quad \text{a.e. in } \Omega_{\mathrm{pde}},$$

where d_y, L_y denote the partial derivatives of d, L with respect to the argument y.

Suppose that the optimal control problem contains control constraints of the type $|u(x)| \le b$ f.a.a. $x \in \Omega$. This can be modeled by setting $g(u) = +\infty$ for all u with $|u| > b$. Then the domain of j is a bounded subset of $L^2(\Omega)$. The Lipschitz continuity of $u \mapsto \nabla f(u) = \chi_\Omega p_u$ can be proven by standard techniques, see, e.g., [23, Lemma 4.1]. The maximum principle holds for such problems as well, see [11].

3 Sparsity promoting proximal operators

The focus of this section is to investigate under which assumptions prox_{sg} is sparsity promoting. Here, we want to prove that there is $\sigma > 0$ such that for all q

$$u \in \operatorname{prox}_{sg}(q) \;\Rightarrow\; u = 0 \text{ or } |u| \ge \sigma. \tag{3.1}$$

In [21, 22], this was also investigated for some special cases of non-convex functions. We will show that the following assumption is enough to guarantee the sparsity promoting property. It contains the requirements from e.g. [21, Theorem 3.3] and [8, Lemma 3.1] as a special case.

Assumption B

(B1) $g : \mathbb{R} \to \bar{\mathbb{R}}$ is lower semicontinuous, $g(x) = g(-x)$ for all $x \in \mathbb{R}$, and $g(0) = 0$.
(B2) There is $u \neq 0$ such that $g(u) \in \mathbb{R}$.
(B3) g satisfies one of the following properties:

(B3.a) g is twice differentiable on an interval $(0, \epsilon)$ for some $\epsilon > 0$ and $\limsup\limits_{u \searrow 0} g''(u) \in (-\infty, 0)$,
(B3.b) g is twice differentiable on an interval $(0, \epsilon)$ for some $\epsilon > 0$ and $\lim\limits_{u \searrow 0} g''(u) = -\infty$,
(B3.c) $0 < \liminf_{u \searrow 0} g(u)$.

(B4) $g(u) \geq 0$ for all $u \in \mathbb{R}$.

By Assumption B, the function g is non-convex in a neighborhood of 0 and non-smooth at 0. Some examples are given below.

Example 3.1 Functions satisfying Assumption B:

(1) $g(u) := |u|_0 := \begin{cases} 1 & u \neq 0, \\ 0 & u = 0, \end{cases}$
(2) $g(u) := |u|^p, \quad p \in (0, 1)$,
(3) $g(u) := \ln(1 + \alpha|u|)$, with a given positive constant α,
(4) the indicator function of the integers $g(u) := \delta_{\mathbb{Z}}(u)$.

In order to prove the desired property (3.1), we have to analyze the structure of the solution set of

$$\min_{u \in \mathbb{R}} h_{q,s}(u) \tag{3.2}$$

for $s > 0$ with

$$h_{q,s}(u) := -qu + \frac{1}{2}u^2 + sg(u).$$

Let us begin with stating basic properties of prox_{sg}.

Lemma 3.2 *Let $g : \mathbb{R} \to \bar{\mathbb{R}}$ satisfy (B1) and (B4). Then $\operatorname{prox}_{sg}(q)$ is non-empty for all $q \in \mathbb{R}$. In addition, the graph of prox_{sg} is a closed set. Moreover, $q \rightrightarrows \operatorname{prox}_{sg}(q)$ is monotone, i.e., the inequality $0 \leq (q_1 - q_2)(u_1 - u_2)$ is satisfied for all $q_1, q_2 \in \mathbb{R}$ and $u_1 \in \operatorname{prox}_{sg}(q_1)$, $u_2 \in \operatorname{prox}_{sg}(q_2)$.*

Proof The function $h_{q,s}$ is lower semicontinuous, thus closed. Further, it is coercive, i.e., $h_{s,q}(u) \to +\infty$ as $|u| \to +\infty$. This implies the non-emptiness of prox_{sg}, see [5, Theorem 6.4]. The closedness of the graph of prox_{sg} is a consequence of the lower semicontinuity of g. The monotonicity can be verified by using the optimality for (3.2). That is for $u_1 \in \operatorname{prox}_{sg}(q_1)$ and $u_2 \in \operatorname{prox}_{sg}(q_2)$ it holds

$$h_{q_1,s}(u_1) \leq h_{q_1,s}(u_2) \text{ and } h_{q_2,s}(u_2) \leq h_{q_2,s}(u_1),$$

respectively. Elementary computations yield the claimed inequality. □

Lemma 3.3 *Let $g : \mathbb{R} \to \bar{\mathbb{R}}$ satisfy (B1). Let $u \in \operatorname{prox}_{sg}(q)$. Then $u \geq 0$ if and only if $q \geq 0$.*

Proof Due to (B1), we have $u \in \operatorname{prox}_{sg}(q)$ if and only if $-u \in \operatorname{prox}_{sg}(-q)$. The claim now follows from the monotonicity of the prox-map. □

Lemma 3.4 *Let $g : \mathbb{R} \to \bar{\mathbb{R}}$ satisfy (B1) and (B4). Then the growth condition*

$$|u| \leq 2|q| \quad \forall u \in \operatorname{prox}_{sg}(q)$$

is satisfied.

Proof Let $u \in \operatorname{prox}_{sg}(q)$. By optimality, the following inequality

$$\frac{1}{2}u^2 - qu + sg(u) \leq g(0) = 0$$

is true. Since $g(u) \geq 0$, the claim follows. □

Next, we have to make sure that the image of prox_{sg} is not equal to $\{0\}$.

Lemma 3.5 *Let H be a Hilbert space. Let $f : H \to \bar{\mathbb{R}}$ be a function with $f(0) \in \mathbb{R}$. Then $0 \in \operatorname{prox}_f(q)$ for all $q \in H$ if and only if f is of the form $f(x) = f(0) + \delta_{\{0\}}(x)$.*

Proof If f is of the claimed form, then clearly $\operatorname{prox}_f(q) = \{0\}$ for all q. Now, let $0 \in \operatorname{prox}_f(q)$ for all $q \in H$. Then it holds

$$\frac{1}{2}\|u - q\|_H^2 + f(u) \geq \frac{1}{2}\|q\|_H^2 + f(0) \quad \forall u, q \in H.$$

This is equivalent to

$$f(u) + \frac{1}{2}\|u\|_H^2 \geq f(0) + (u, q)_H \quad \forall u, q \in H.$$

Setting $q := tu$ and letting $t \to +\infty$ shows $f(u) = +\infty$ for all $u \neq 0$. □

Lemma 3.6 *Let $g : \mathbb{R} \to \bar{\mathbb{R}}$ satisfy (B1). Let $s > 0$. Assume there is $q_0 \geq 0$ such that*

$$q_0|u| \leq \frac{1}{2}u^2 + sg(u) \quad \forall u \in \mathbb{R}. \tag{3.3}$$

Then the following statements hold:

(1) $u = 0$ *is a global solution to* (3.2) *if* $|q| \leq q_0$. *If* $|q| < q_0$, *then* $u = 0$ *is the unique global solution to* (3.2).

(2) *Moreover, if*

$$q_0 := \sup\{q \geq 0 : q|u| \leq \frac{1}{2}u^2 + sg(u) \quad \forall u \in \mathbb{R}\}, \tag{3.4}$$

then $|q| \leq q_0$ *is also necessary for* $u = 0$ *to be a global solution to* (3.2).

Proof Let $|q| \leq q_0$. Take $u \neq 0$, then we have

$$h_{q,s}(u) = \frac{1}{2}u^2 + sg(u) - uq \geq \frac{1}{2}u^2 + sg(u) - |u| \cdot |q| \geq \frac{1}{2}u^2 + sg(u) - q_0|u| \geq 0 = h_{q,s}(0).$$

Note that the second inequality is strict if $|q| < q_0$. To prove (2), assume $u = 0$ is a global solution to (3.2). Assume $q > 0$. Then it holds

$$qu \leq \frac{1}{2}u^2 + sg(u) \quad \forall u \geq 0.$$

Since $g(u) = g(-u)$, this implies

$$q|u| \leq \frac{1}{2}u^2 + sg(u) \quad \forall u \in \mathbb{R}.$$

By the definition of q_0, the inequality $q \leq q_0$ follows. Similarly, one can prove $|q| \leq q_0$ for negative q. □

Together with Assumption B, these results allow us to show the desired sparsity promoting property (3.1). A similar statement to the following can be found in [22, Theorem 1.1].

Theorem 3.7 *Let* $g : \mathbb{R} \to \bar{\mathbb{R}}$ *satisfy Assumption* B. *Let us set*

$$s_0 := \begin{cases} -\frac{1}{\limsup_{u \searrow 0} g''(u)} & \text{if (B.3a) is satisfied,} \\ 0 & \text{if (B.3b) or (B3.c) is satisfied.} \end{cases} \tag{3.5}$$

Then the following statements hold:

(1) *For every* $s > s_0$ *there is* $u_0(s) > 0$ *such that for all* $q \in \mathbb{R}$ *every global minimizer* u of (3.2) *satisfies*

$$u = 0 \text{ or } |u| \geq u_0(s).$$

(2) *Moreover, for all $s > 0$ there is $q_0 := q_0(s) > 0$ such that $u = 0$ is a global solution to* (3.2) *if and only if $|q| \leq q_0$. If $|q| < q_0$ then $u = 0$ is the unique global solution to* (3.2).

Proof We prove the first claim (1) by contradiction. Therefore, assume g satisfies Assumption B but the first claim does not hold, i.e., there exists $s > s_0$ such that for all $u_0 > 0$ there is q and u with $u \in \operatorname{prox}_{sg}(q)$ and $0 < |u| < u_0$. Then there are sequences (u_n) and (q_n) with $u_n \in \operatorname{prox}_{sg}(q_n)$, $u_n \neq 0$, and $u_n \to 0$. W.l.o.g., (u_n) is a monotonically decreasing sequence of positive numbers, and hence (q_n) is monotonically decreasing and non-negative by Lemma 3.3. Let u and q denote the limits of both sequences. Since $u_n \neq 0$ is a global minimum of $h_{q_n,s}$, it follows $h_{q_n,s}(u_n) \leq h_{q_n,s}(0) = 0$. Passing to the limit in this inequality, we obtain

$$\liminf_{n\to+\infty} h_{q_n,s}(u_n) = \liminf_{n\to+\infty} g(u_n) \leq 0.$$

Hence, (B3.c) is violated, so at least one of (B.3a) or (B.3b) is satisfied. For n sufficiently large, we have $0 < u_n < \epsilon$, and the necessary second-order optimality condition $h''_{q_n,s}(u_n) \geq 0$ holds, and we obtain

$$\limsup_{n\to+\infty} h''_{q_n,s}(u_n) \geq 0,$$

which implies

$$1 + s \limsup_{n\to+\infty} g''(u_n) \geq 0.$$

This inequality is a contradiction to (B.3a) and (B.3b) due to the choice of $s > s_0$, and the first claim is proven.

In order to prove the claim (2), we will apply Lemma 3.6. First, assume that (B.3a) or (B.3b) is satisfied, i.e., there is $\epsilon_1 > 0$ such that g is strictly concave on $(0, \epsilon_1]$. By reducing ϵ_1 if necessary, we get $g(\epsilon_1) > 0$. Since $g(0) = 0$, it holds $g(u) \geq \frac{g(\epsilon_1)}{\epsilon_1}|u|$ for all $\boldsymbol{u \in [0, \epsilon_1]}$ by concavity. Due to symmetry, this holds for all u with $|u| \leq \epsilon_1$. Since $g(u) \geq 0$ for all u by (B4), it holds $\frac{1}{2}u^2 + sg(u) \geq \frac{\epsilon_1}{2}|u|$ for all $|u| \geq \epsilon_1$. This proves $\frac{1}{2}u^2 + sg(u) \geq \min(\frac{\epsilon_1}{2}, \frac{sg(\epsilon_1)}{\epsilon_1})|u|$ for all u, and the set appearing in (3.4) is non-empty. Second, if (B3.c) is satisfied, then there are $\epsilon_2, \tau > 0$ such that $g(u) \geq \tau$ for all u with $|u| \in (0, \epsilon_2)$ as g is lower semicontinuous. Therefore, it holds $g(u) \geq \tau \geq \frac{\tau}{\epsilon_2}|u|$ if $|u| \in (0, \epsilon_2)$. Similarly as in the first case, we find that the set in (3.4) is non-empty. By (B2), this set is bounded. Thus, the claim follows with q_0 from (3.4) and Lemma 3.6. □

Remark 3.8 In general, the constant u_0 in Theorem 3.7 depends on s and the structure of g.

Example 3.9 The proximal map of $g(u) := |u|_0$ is given by the hard-thresholding operator, defined by $\operatorname{prox}_{sg}(q) = \begin{cases} 0 & \text{if } |q| \leq \sqrt{2s}, \\ q & \text{otherwise.} \end{cases}$

With the above considerations in mind, let us discuss the minimization problem

$$\min_{u\in\mathbb{R}} g_k u + \frac{L}{2}(u-u_k)^2 + g(u), \tag{3.6}$$

which arises as the pointwise minimization of the integrand in (1.1).

Corollary 3.10 *Let $g_k, u_k \in \mathbb{R}, L > 0$ be given. Then $u \in \mathbb{R}$ is a solution to (3.6) if and only if*

$$u \in \operatorname{prox}_{L^{-1}g}\left(\frac{Lu_k - g_k}{L}\right).$$

If $\frac{1}{L} > s_0$, see Theorem 3.7, then all global solutions u satisfy

$$u = 0 \text{ or } |u| \geq u_0(L^{-1})$$

with some $u_0(L^{-1}) > 0$ as in Theorem 3.7.

Proof Problem (3.6) is equivalent to

$$\min_{u\in\mathbb{R}} \frac{g_k - Lu_k}{L}u + \frac{1}{2}u^2 + \frac{1}{L}g(u)$$

and therefore of the form (3.2). The claim follows from Theorem 3.7. □

4 Analysis of the proximal gradient algorithm

In this section, we will analyze the proximal gradient algorithm. Throughout this section, we assume that f and g satisfy Assumptions A and B.

Algorithm 4.1 (Proximal gradient algorithm) Choose $L > 0$ and $u_0 \in L^2(\Omega)$. Set $k = 0$.

(1) Compute u_{k+1} as solution of

$$\min_{u\in L^2(\Omega)} f(u_k) + \nabla f(u_k)\cdot(u-u_k) + \frac{L}{2}\|u-u_k\|^2_{L^2(\Omega)} + j(u). \tag{4.1}$$

(2) Set $k := k+1$, go to step 1.

The functional to be minimized in (4.1) can be written as an integral functional. In this representation the minimization can be carried out pointwise by using the previous results. The following statements are generalizations of [27, Lemma 3.10, Theorem 3.12].

Lemma 4.2 *Let $u_k \in L^2(\Omega)$ be given. Then*

$$\min_{u\in L^2(\Omega)} f(u_k) + \nabla f(u_k)\cdot(u-u_k) + \frac{L}{2}\|u-u_k\|^2_{L^2(\Omega)} + \int_\Omega g(u(x))\,\mathrm{d}x \tag{4.2}$$

is solvable, and $u_{k+1} \in L^2(\Omega)$ is a global solution if and only if

$$u_{k+1}(x) \in \operatorname{prox}_{L^{-1}g}\left(\frac{1}{L}(Lu_k(x) - \nabla f(u_k)(x))\right) \tag{4.3}$$

for almost all $x \in \Omega$.

Proof Let us show, that we can choose a measurable function satisfying the inclusion (4.3). The set-valued mapping $\operatorname{prox}_{L^{-1}g}$ has a closed graph. Then by [24, Corollary 14.14], the set-valued mapping $x \rightrightarrows \operatorname{prox}_{L^{-1}g}\left(\frac{1}{L}(Lu_k(x) - \nabla f(u_k)(x))\right)$ from Ω to $\mathbb{R}$ is measurable. A well-known result [24, Corollary 14.6] implies the existence of a measurable function u such that $u(x) \in \operatorname{prox}_{L^{-1}g}\left(\frac{1}{L}(Lu_k(x) - \nabla f(u_k)(x))\right)$ for almost all $x \in \Omega$. Due to the growth condition of Lemma 3.4, we have $u \in L^2(\Omega)$, and hence u solves (4.2). If u_{k+1} solves (4.2) then (4.3) follows by a standard argument, see e.g., [27, Theorem 3.10]. □

Remark 4.3 Due to its non-convexity, the minimization problem in Algorithm 4.1 may not have a unique minimizer, and $\operatorname{prox}_{L^{-1}g}\left(\frac{1}{L}(Lu_k(x) - \nabla f(u_k)(x))\right)$ is not a singleton. For the choice $g(u) = |u|_0$ or $g(u) = |u|^p$, $p \in (0,1)$, the image of prox contains zero, and we suggest to choose $u_{k+1}(x) = 0$. For the general case, one can construct a monotonically increasing function $P : \mathbb{R} \to \mathbb{R}$ such that $P(q) \in \operatorname{prox}_{L^{-1}g}(q)$ for all $q \in \mathbb{R}$. Then set $u_{k+1}(x) := P\left(\frac{1}{L}(Lu_k(x) - \nabla f(u_k)(x))\right)$.

We introduce the following notation. For a sequence $(u_k) \subset L^2(\Omega)$ define

$$I_k := \{x \in \Omega : u_k(x) \neq 0\},\ \chi_k := \chi_{I_k}. \tag{4.4}$$

Let us now investigate convergence properties of Algorithm 4.1. The following Lemma will be helpful for what follows. It strongly builds on the sparsity promoting property of g, and uses all conditions of Assumption B via Theorem 3.7.

Lemma 4.4 *Assume $\frac{1}{L} > s_0$ with s_0 from Theorem 3.7. Let $u_k, u_{k+1} \in L^2(\Omega)$ be consecutive iterates of Algorithm 4.1. Then*

$$\|u_{k+1} - u_k\|^p_{L^p(\Omega)} \geq u_0^p \|\chi_k - \chi_{k+1}\|_{L^1(\Omega)}$$

holds for $p \in [1, +\infty)$, where $u_0 := u_0(L^{-1})$ is as in Theorem 3.7.

Proof Since $u_k(x) \neq 0$ and $u_{k+1}(x) = 0$ on $I_k \setminus I_{k+1}$ by (4.4), it holds $|u_{k+1}(x) - u_k(x)| \geq u_0$ for all $x \in I_k \setminus I_{k+1}$ by Corollary (3.10). Hence,

$$\begin{aligned}\|u_{k+1}-u_k\|^p_{L^p(\Omega)} &= \int_\Omega |u_{k+1}(x)-u_k(x)|^p \,\mathrm{d}x\\ &\geq \int_{(I_k\setminus I_{k+1})\cup(I_{k+1}\setminus I_k)} |u_{k+1}(x)-u_k(x)|^p \,\mathrm{d}x \geq u_0^p\|\chi_{k+1}-\chi_k\|_{L^1(\Omega)},\end{aligned}$$

where we have used $\|\chi_{k+1}-\chi_k\|_{L^1(\Omega)} = |(I_k\setminus I_{k+1})\cup(I_{k+1}\setminus I_k)|$. □

Now, we are in the position to prove the first, basic convergence result. This theorem already makes full use of Assumptions A and B.

Theorem 4.5 *For $L > L_f$* let (u_k) *be a sequence of iterates generated by Algorithm 4.1. Then the following statements hold:*

(1) *The sequence $(f(u_k)+j(u_k))$ is monotonically decreasing and converging.*
(2) *The sequences (u_k) and $(\nabla f(u_k))$ are bounded in $L^2(\Omega)$ if $f+j$ is weakly coercive on $L^2(\Omega)$, i.e., $f(u)+j(u)\to+\infty$ as $\|u\|_{L^2(\Omega)}\to+\infty$.*
(3) *It holds $u_{k+1}-u_k\to 0$* in $L^2(\Omega)$ *and pointwise almost everywhere on* Ω.
(4) *Let s_0 be as in Theorem 3.7. If $\frac{1}{L}>s_0$, then the sequence of characteristic functions (χ_k) is converging in $L^1(\Omega)$ and pointwise a.e. to some characteristic function χ.*

Proof *(1)* Due to the Lipschitz continuity of ∇f by (A2) it holds

$$f(u_{k+1}) \leq f(u_k) + \nabla f(u_k)(u_{k+1}-u_k) + \frac{L_f}{2}||u_{k+1}-u_k||^2_{L^2(\Omega)}. \tag{4.5}$$

Using the optimality of u_{k+1}, we find that the inequality

$$f(u_{k+1})+j(u_{k+1}) \leq f(u_k)+j(u_k) - \frac{L-L_f}{2}\|u_{k+1}-u_k\|^2_{L^2(\Omega)} \tag{4.6}$$

holds. Hence, $(f(u_k)+j(u_k))$ is decreasing. Convergence follows because f and j are bounded from below by Assumptions (A2) and (B1).

(2) Weak coercivity of the functional implies that (u_k) is bounded. Furthermore, because of

$$\begin{aligned}\|\nabla f(u_k)\|_{L^2(\Omega)} &\leq \|\nabla f(u_k)-\nabla f(0)\|_{L^2(\Omega)} + \|\nabla f(0)\|_{L^2(\Omega)}\\ &\leq L_f\|u_k\|_{L^2(\Omega)} + \|\nabla f(0)\|_{L^2(\Omega)},\end{aligned}$$

boundedness of $(\nabla f(u_k))$ in $L^2(\Omega)$ follows.

(3) Summation over $k=1,\dots,n$ in (4.6) yields

$$\sum_{k=1}^n (f(u_{k+1})+j(u_{k+1})) \leq \sum_{k=1}^n \left(f(u_k)+j(u_k) - \frac{L-L_f}{2}\|u_{k+1}-u_k\|^2_{L^2(\Omega)}\right)$$

and hence

$$f(u_{n+1}) + j(u_{n+1}) + \sum_{k=1}^{n} \frac{L - L_f}{2} \|u_{k+1} - u_k\|^2_{L^2(\Omega)} \leq f(u_1) + j(u_1) < +\infty.$$

Letting $n \to +\infty$ implies $\sum_{k=1}^{+\infty} \|u_{k+1} - u_k\|^2_{L^2(\Omega)} < +\infty$ and therefore $\|u_{k+1} - u_k\|_{L^2(\Omega)} \to 0$. By the Lemma of Fatou, we have further

$$\int_\Omega \liminf_{n\to+\infty} \sum_{k=0}^{n} |u_{k+1}(x) - u_k(x)|^2 \,\mathrm{d}x \leq \liminf_{n\to+\infty} \sum_{k=0}^{n} \|u_{k+1}(x) - u_k(x)\|^2_{L^2(\Omega)} < +\infty.$$

This implies $\liminf_{n\to+\infty} \sum_{k=0}^{n} |u_{k+1}(x) - u_k(x)|^2 < +\infty$ for almost all $x \in \Omega$, and the second claim follows.

(4) By Lemma 4.4, we get

$$\frac{L - L_f}{2} u_0^2 \sum_{k=1}^{+\infty} \|\chi_k - \chi_{k+1}\|_{L^1(\Omega)} \leq \frac{L - L_f}{2} \sum_{k=1}^{+\infty} \|u_k - u_{k+1}\|_{L^2(\Omega)} < +\infty$$

Hence, (χ_k) is a Cauchy sequence in $L^1(\Omega)$, and therefore also converging in $L^1(\Omega)$, i.e., $\chi_k \to \chi$ for some characteristic function χ. Pointwise a.e. convergence of (χ_k) can be proven by Fatou's Lemma. □

4.1 Stationarity conditions for weak limit points from inclusions

In order to make full use of Theorem 4.5, we assume throughout this section that the proximal parameter L in Algorithm 4.1 satisfies

$$L > L_f \text{ and } \frac{1}{L} > s_0,$$

where s_0 is from Theorem 3.7, see (3.5).

Under a weak coercivity assumption, Theorem 4.5(2) implies that Algorithm 4.1 generates a sequence (u_k) with weak limit point $u^* \in L^2(\Omega)$, i.e., there exists a subsequence of iterates (u_k) converging weakly to u^* in $L^2(\Omega)$. Due to the lack of weak lower semicontinuity in the term $u \mapsto \int_\Omega g(u)\,\mathrm{d}x$, however, we cannot conclude anything about the value of the objective functional in a weak limit point. Unfortunately, we are not able to show

$$f(u^*) + j(u^*) \leq \lim_{k\to+\infty} f(u_k) + j(u_k)$$

along the subsequence, as it was done in [27, Thm. 3.14] for the special choice $g(u) := |u|_0$. Nevertheless, by using results of set-valued analysis we will show that a weak limit point of a sequence (u_k) of iterates satisfies a certain inclusion in almost every point $x \in \Omega$, which can be interpreted as a pointwise stationary condition for weak limit points.

By definition, the iterates satisfy the inclusion

$$u_{k+1}(x) \in \operatorname{prox}_{L^{-1}g}\Big(\frac{1}{L}(Lu_k(x) - \nabla f(u_k)(x))\Big)$$

for almost all $x \in \Omega$, see e.g., (4.3). However, this inclusion seems to be useless for a convergence analysis, as the function u_{k+1} to the left of the inclusion as well as the arguments $Lu_k - \nabla f(u_k)$ only have weakly converging subsequences at best. The idea is to construct a set-valued mapping $\mathcal{G} : \mathbb{R} \rightrightarrows \mathbb{R}$ such that a solution u_{k+1} of (4.2) satisfies the inclusion

$$u_{k+1}(x) \in \mathcal{G}(z_k(x)) \tag{4.7}$$

in almost every point $x \in \Omega$ for some $z_k \in L^2(\Omega)$, where (z_k) converges strongly or pointwise almost everywhere. Here, we will use

$$z_k := -\big(\nabla f(u_k) + L(u_{k+1} - u_k)\big). \tag{4.8}$$

By Theorem 4.5, we have $u_{k+1} - u_k \to 0$ in $L^2(\Omega)$ and pointwise almost everywhere. With the additional assumption that subsequences of $(\nabla f(u_k))$ converge pointwise almost everywhere, the argument of the set-valued mapping converges pointwise almost everywhere. In the context of optimal control problems, such an assumption is not a severe restriction.

If $\nabla f : L^2(\Omega) \to L^2(\Omega)$ is completely continuous, then this assumption is fulfilled. For many control problems, this property of ∇f is guaranteed to hold.

So there is a chance to pass to the limit in the inclusion (4.7).

Corollary 4.6 *Let (u_k) be a sequence of iterates generated by Algorithm 4.1 with weak limit point $u^* \in L^2(\Omega)$, i.e., $u_{k_n} \rightharpoonup u^*$. Assume $\nabla f(u_{k_n})(x) \to \nabla f(u^*)(x)$ for almost every $x \in \Omega$. Then it follows $z_{k_n}(x) \to -\nabla f(u^*)(x)$ for almost every $x \in \Omega$.*

Proof This is a direct consequence of the definition of (z_k) in (4.8) and Theorem 4.5(3). □

Let us now give an equivalent characterization of $\mathcal{G}$ as defined in (4.7).

Lemma 4.7 *Let u_{k+1} be a solution of (4.2). Then*

$$u_{k+1}(x) \in \mathcal{G}(z_k(x)) f.a.a. x \in \Omega,$$

where the set-valued mapping $\mathcal{G} : \mathbb{R} \rightrightarrows \mathbb{R}$ is given by

$$\begin{aligned} u \in \mathcal{G}(z) &\Longleftrightarrow u \in \operatorname*{arg\,min}_{v\in\mathbb{R}} -zv + \frac{L}{2}(v-u)^2 + g(v) \\ &\Longleftrightarrow u \in \operatorname{prox}_{L^{-1}g}\Big(\frac{Lu+z}{L}\Big) \end{aligned} \tag{4.9}$$

Unfortunately, the set-valued map $\mathcal{G}$ is neither monotone nor single-valued in general. If g would be convex, then the optimality condition of the minimization problem in (4.9) implies $z \in \partial g(u)$. Hence, it holds $\mathcal{G} = \partial g^*$, where g^* denotes the convex conjugate of g, and $\mathcal{G}$ would be monotone. If in addition, g is strictly convex, then $\mathcal{G}$ would be single-valued.

As a first direct consequence from the definition of $\mathcal{G}$, we get

Corollary 4.8 *Let $u_0 := u_0(L^{-1})$ and $q_0 := q_0(L^{-1})$ be the positive constants from Theorem 3.7. Let $u, z \in \mathbb{R}$ be such that $u \in \mathcal{G}(z)$. Then we have:* If $u > 0$ *then* $u \geq \max\left(u_0, \frac{Lq_0 - z}{L}\right)$, *and if* $u < 0$ *then* $u \leq \min\left(-u_0, -\frac{Lq_0+z}{L}\right)$. *In case* $u = 0$ *it holds* $|z| \leq Lq_0$.

Proof Here, we will use the sparsity promoting property of $\operatorname{prox}_{L^{-1}g}$ in (4.9). If $u \neq 0$ then by Lemma 3.3 and Theorem 3.7, it follows that $u \geq u_0$ if and only if $\frac{Lu+z}{L} \geq q_0$ and likewise $u < -u_0$ if and only if $\frac{Lu+z}{L} \leq -q_0$. The claim follows for $u > 0$ and $u < 0$, respectively. On the other hand $u = 0$ is a solution if and only if $|\frac{z}{L}| \leq q_0$, which implies the claim for $u = 0$. □

4.2 A convergence result for inclusions

In this section, we will prove a convergence result to be able to pass to the limit in the inclusion (4.7) and to identify the set-valued map that is obtained in this limiting process. First, let us recall a few helpful notions and results from set-valued analysis that can be found in the literature, see e.g., [2, 24].

Definition 4.9 For a sequence of sets $A_n \subset \mathbb{R}^n$ we define the *outer limit* by

$$\limsup_{n\to+\infty} A_n := \{x : \exists (x_{n_k}), x_{n_k} \to x, x_{n_k} \in A_{n_k}\}.$$

Definition 4.10 Let $S : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ be a set-valued map.

(1) The domain and graph of S are defined by

$$\operatorname{dom} S := \{x : S(x) \neq \emptyset\}, \ \operatorname{gph} S := \{(x, y) : y \in S(x)\}.$$

(2) S is called *outer semicontinuous* in $\bar{x}$ if

$$\limsup_{x\to\bar{x}} S(x) \subseteq S(\bar{x}).$$

(3) S is called *locally bounded* at $x \in \mathbb{R}^m$ if there is a neighborhood U of x such that $S(U)$ is bounded.

A set-valued mapping S is outer semicontinuous if and only if it has a closed graph. The following convergence analysis relies on [2, Thm. 7.2.1]. There the local boundedness of $\mathcal{G}$ is a prerequisite, which is not satisfied in general in our situation. Hence, we have to extend this result to set-valued maps into $\mathbb{R}^n$ that are not locally bounded. Let us define the following set-valued map that serves as a generalization of $x \rightrightarrows \operatorname{conv}(F(x))$ for the locally unbounded situation.

Definition 4.11 Let $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ be a set-valued map.

Define the set-valued map $\operatorname{conv}^\infty F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ by

$$(\operatorname{conv}^\infty F)(x) := \limsup_{k\to+\infty} \operatorname{conv}\big(F\big(x + B_{1/k}(0)\big)\big).$$

By definition, it holds $\operatorname{gph} F \subset \operatorname{gph} \operatorname{conv}^\infty F$. In addition, we have $\overline{conv}\,(F(x)) \subset (\operatorname{conv}^\infty F)(x)$ for all $x \in \mathbb{R}^m$. If F is locally bounded in x, then $(\operatorname{conv}^\infty)F(x) = \overline{conv}\,(F(x))$, which can be proven using Carathéodory's theorem. In general, $\operatorname{dom} \operatorname{conv}^\infty F$ is strictly larger than $\operatorname{dom} F$.

Example 4.12 Define $F : \mathbb{R} \rightrightarrows \mathbb{R}$ by

$$\operatorname{gph} F = \{(x, y) : \ yx = 1\}.$$

Then F is not locally bounded near $x = 0$. Here it holds $\operatorname{gph}(\operatorname{conv}^\infty F) = \operatorname{gph} F \cup (\{0\} \times \mathbb{R})$, so that $\operatorname{dom}(\operatorname{conv}^\infty F) = \mathbb{R} \neq \operatorname{dom} F$.

Theorem 4.13 *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ be a set-valued map. Let sequences of measurable functions $(x_n), (y_n)$, $x_n : \Omega \to \mathbb{R}^m$, $y_n : \Omega \to \mathbb{R}^n$, be given such that*

(1) *x_n converges almost everywhere to some measurable function $x : \Omega \to \mathbb{R}^m$,*
(2) *y_n converges weakly to a function y* in $L^1(\Omega;\mathbb{R}^n, \mu)$,
(3) *$y_n(t) \in F(x_n(t))$ for almost all $t \in \Omega$.*

Then for almost all $t \in \Omega$ it holds:

$$y(t) \in (\operatorname{conv}^\infty F)(x(t)).$$

Proof Arguing as in the proof of [2, Thm. 7.2.1], we find

$$y(t) \in \bigcap_{k\in\mathbb{N}} \overline{conv}\,\big(F(x(t) + B_{1/k}(0))\big)$$

for almost all $t \in \Omega$. Note that we can choose $\mathcal{W} = \{0\}$ as our assumption (3) is stronger than the condition (7.1) in [2, Thm. 7.2.1].

Take $t \in \Omega$ such that the above inclusion is satisfied. Then there is a sequence (u_k) such that $u_k \to y(t)$, $u_k \in \operatorname{conv}(F(x(t) + B_{1/k}(0)))$. This implies $y(t) \in \limsup_{k\to+\infty} \operatorname{conv}\big(F(x(t) + B_{1/k}(0))\big)$, or equivalently $y(t) \in (\operatorname{conv}^\infty F)(x(t))$. □

Let us close this section with an example that shows that $\mathcal{G}$ is not necessarily locally bounded.

Example 4.14 Let $L > 0$ and define $\boldsymbol{g(u) := \delta_{\mathbb{Z}}(u)}$ the indicator function of integers with the associated map $\mathcal{G}$ defined as in (4.9). Set $U := [-\frac{L}{2}, \frac{L}{2}]$. Then it holds that $\mathcal{G}(z) = \mathbb{Z}$ for all $z \in U$, i.e., $\mathcal{G}$ is clearly not locally bounded in the origin.

4.3 Stationarity conditions for weak limit points

Recall that for iterates (u_k) of Algorithm 4.1 and the corresponding sequence z_k, see (4.8), we have by construction

$$u_{k+1}(x) \in \mathcal{G}(z_k(x)) \text{ f.a.a. } x \in \Omega,$$

with $\mathcal{G}$ is defined as in (4.9). By Theorem 4.13, we could expect the inclusion $u^*(x) \in (\operatorname{conv}{}^\infty \mathcal{G})(-\nabla f(u^*)(x))$ to hold pointwise almost everywhere in the subsequential limit. However, the convexification of $\mathcal{G}$ results in a set-valued map that is very large. In order to obtain a smaller inclusion in the limit, we will employ the result of Corollary 4.8: the graph of $\mathcal{G}$ can be split into three separated components. In the sequel, we will show that we can pass to the limit with each component separately, which leads to a smaller set-valued map in the limit. This observation motivates the following splitting of the map $\mathcal{G}$.

Definition 4.15 We define the following set-valued mappings.

(1) $\mathcal{G}^+ : \mathbb{R} \rightrightarrows \mathbb{R}$ with $u \in \mathcal{G}^+(z) :\Longleftrightarrow u \in \mathcal{G}(z)$ and $u > 0$,
(2) $\mathcal{G}^- : \mathbb{R} \rightrightarrows \mathbb{R}$ with $u \in \mathcal{G}^-(z) :\Longleftrightarrow u \in \mathcal{G}(z)$ and $u < 0$,
(3) $\mathcal{G}^0 : \mathbb{R} \rightrightarrows \mathbb{R}$ with $u \in \mathcal{G}^0(z) :\Longleftrightarrow u \in \mathcal{G}(z)$ and $u = 0$.

Obviously we have by construction

$$u_{k+1}(x) \in (\mathcal{G}^+ \cup \mathcal{G}^- \cup \mathcal{G}^0)(z_k(x)) \quad \text{f.a.a. } x \in \Omega. \tag{4.10}$$

For better illustration, let us give an example of the mappings $\mathcal{G}^+, \mathcal{G}^-$ and $\mathcal{G}^0$.

Example 4.16 Let $g(u) := \frac{\alpha}{2}|u|^2 + \beta|u|^p + \delta_{[-b,b]}(u)$, $p \in (0,1)$, $\boldsymbol{b} > \boldsymbol{0}$. In Fig. 1, the union $\mathcal{G}^0 \cup \mathcal{G}^+ \cup \mathcal{G}^-$ and the convexified map $\mathcal{G}^0 \cup \overline{conv}\, \mathcal{G}^+ \cup \overline{conv}\, \mathcal{G}^-$ is depicted for the special choices of the parameters. A more detailed discussion of this choice is given in Sect. 5.1.

Corollary 4.17 *The mappings $\mathcal{G}, \mathcal{G}^0, \mathcal{G}^+$, and $\mathcal{G}^-$ are outer semicontinuous.*

Proof $\mathcal{G}$ being outer semicontinuous is equivalent to the closedness of its graph. Let $(u_n), (q_n)$ be sequences such that $u_n \to u$, $q_n \to q$ and $u_n \in \mathcal{G}(q_n)$. By definition it holds

$$0 \le -q_n(v - u_n) + (g(v) - g(u_n)) + \frac{L}{2}(v - u_n)^2$$

for all $v \in \mathbb{R}$. Passing to the limit in above inequality yields

$$0 \le -q(v - u) + (g(v) - g(u)) + \frac{L}{2}(v - u)^2$$

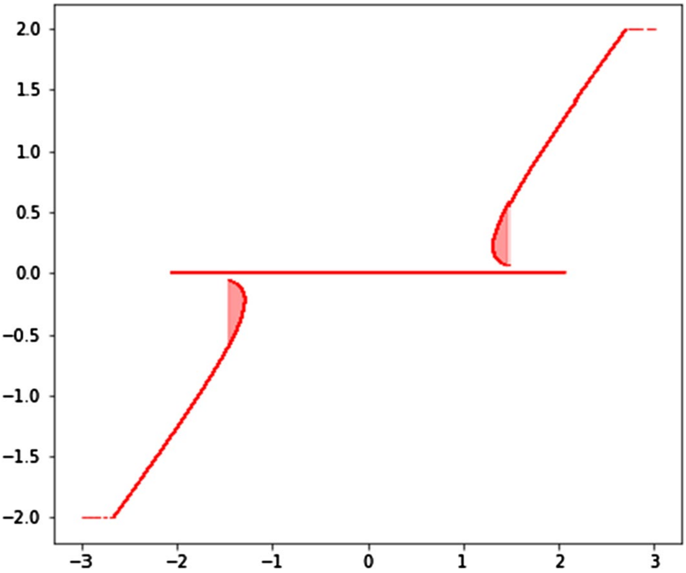


Fig. 1 The union $(\mathcal{G}^0 \cup \mathcal{G}^+ \cup \mathcal{G}^-)$ and its convexification $(\mathcal{G}^0 \cup \overline{conv}\,\mathcal{G}^+ \cup \overline{conv}\,\mathcal{G}^-)$ (filled area) with parameters $(L, \alpha, \beta, b) = (0.01, 0.01, 0.01, 2)$

due to the lower semicontinuity of g. Hence,

$$u \in \underset{v\in\mathbb{R}}{\arg\min}\ -qv + \frac{L}{2}(v-u)^2 + g(v),$$

i.e., $u \in \mathcal{G}(q)$, which is the claim for $\mathcal{G}$. For $\mathcal{G}^+, \mathcal{G}^-, \mathcal{G}^0$ the claim follows as their graphs are intersections of closed sets with $\operatorname{gph}\mathcal{G}$, which follows from Corollary 4.8, where we used $L^{-1} > s_0$ in case of $\mathcal{G}^+, \mathcal{G}^-$. □

In the sequel we want to apply Theorem 4.13 to each of the set-valued maps in (4.10) separately. Let us first show the next helpful result, which gives us finer control of the subsets of Ω, where $u_{k+1}(x) \in \mathcal{G}^+(z_k(x))$ and $u_{k+1}(x) \in \mathcal{G}^-(z_k(x))$. It can be seen as a refinement of claim (4) of Theorem 4.5.

Lemma 4.18 *Let (u_k) be a sequence of iterates generated by Algorithm 4.1. Let real numbers $a < b$ be given. Define*

$$\begin{aligned} A_k^{b+} &:= \{x \in \Omega : \ u_k(x) \geq b\}, \\ A_k^{a-} &:= \{x \in \Omega : \ u_k(x) \leq a\}, \end{aligned}$$

and $\chi_k^{b+} := \chi_{A_k^{b+}}$, $\chi_k^{a-} := \chi_{A_k^{a-}}$. Then it holds

$$\sum_{k=1}^{+\infty} \|\chi_{k+1}^{b+}\chi_k^{a-} + \chi_{k+1}^{a-}\chi_k^{b+}\|_{L^1(\Omega)} < +\infty.$$

If $\chi_k^{b+} + \chi_k^{a-} = 1$ for all k almost everywhere, then there are characteristic functions χ^{b+}, χ^{a-} such that $\chi^{b+} + \chi^{a-} = 1$ almost everywhere, $\chi_k^{b+} \to \chi^{b+}$ and $\chi_k^{a-} \to \chi^{a-}$ strongly in $L^1(\Omega)$ and pointwise almost everywhere.

Proof Let $x \in \Omega$. If $\chi_{k+1}^{b+}(x)\chi_k^{a-}(x) = 1$, then by definition it holds $u_{k+1}(x) - u_k(x) \geq b - a$. This proves $\|\chi_{k+1}^{b+}\chi_k^{a-}\|_{L^1(\Omega)} \leq (b-a)^{-2}\|u_{k+1} - u_k\|_{L^2(\Omega)}^2$. Similarly, we obtain $\|\chi_{k+1}^{a-}\chi_k^{b+}\|_{L^1(\Omega)} \leq (b-a)^{-2}\|u_{k+1} - u_k\|_{L^2(\Omega)}^2$. Since $\sum_{k=1}^{+\infty} \|u_{k+1} - u_k\|_{L^2(\Omega)}^2 < +\infty$, the claim follows. Suppose $\chi_k^{b+} + \chi_k^{a-} = 1$ almost everywhere for all k. Then a simple calculation using the definitions of χ_k^{b+}, χ_k^{a-} yields

$$\chi_{k+1}^{b+}\chi_k^{a-} + \chi_{k+1}^{a-}\chi_k^{b+} = \chi_{k+1}^{b+}(1 - \chi_k^{b+}) + (1 - \chi_{k+1}^{b+})\chi_k^{b+} = |\chi_{k+1}^{b+} - \chi_k^{b+}|,$$

which implies the second claim. □

As a direct consequence, we obtain the convergence of the characteristic functions associated to the positive and negative parts of (u_k). Let us introduce the notation

$$I_k^+ := \{x \in \Omega \, : \, u_k(x) > 0\} \text{ and } I_k^- := \{x \in \Omega \, : \, u_k(x) < 0\} \tag{4.11}$$

with associated characteristic functions χ_k^+, χ_k^-, which we will use in the sequel.

Corollary 4.19 *There are characteristic functions χ^+, χ^- such that $\chi_k^+ \to \chi^+$* and *$\chi_k^- \to \chi^-$* in *$L^1(\Omega)$ and pointwise almost everywhere as $k \to +\infty$.*

Proof Let $u_0 := u_0(L^{-1})$ from Theorem 3.7. By Lemma 4.18 with $a = 0$ and $b = u_0$, we obtain the claim for $\chi_k^+ \to \chi^+$. The proof for $\chi_k^- \to \chi^-$ follows with $a = -u_0$ and $b = 0$. □

Now, we will pass to the limit in the inclusion (4.10). Using characteristic functions, we will split the inclusion into three inclusions for $\mathcal{G}^+, \mathcal{G}^0, \mathcal{G}^-$.

Theorem 4.20 *Let (u_k) be a sequence of iterates generated by Algorithm 4.1 with weak limit point $u^* \in L^2(\Omega)$, i.e., $u_{k_n} \rightharpoonup u^*$. Assume $\nabla f(u_{k_n})(x) \to \nabla f(u^*)(x)$ for almost every $x \in \Omega$.* Let *$\mathcal{G}^0, \mathcal{G}^+, \mathcal{G}^- \, : \, \mathbb{R} \rightrightarrows \mathbb{R}$ be as in Definition 4.15. Then*

$$u^*(x) \in \left(\mathcal{G}_0 \cup \operatorname{conv}^\infty \mathcal{G}^+ \cup \operatorname{conv}^\infty \mathcal{G}^-\right)(-\nabla f(u^*)(x)) \tag{4.12}$$

holds for almost all $x \in \Omega$.

Proof By Corollary 4.6 we have

$$z_{k_n} = -\left(\nabla f(u_{k_n}) + L(u_{k_n+1} - u_{k_n})\right) \to -\nabla f(u^*) := z$$

pointwise almost everywhere on Ω.

Let χ_k^+, χ_k^- as in (4.11). By Corollary 4.19 it holds $\chi_k^+ \to \chi^+$ and $\chi_k^- \to \chi^-$ in $L^1(\Omega)$ and pointwise almost everywhere for characteristic functions χ^+, χ^-.

Let us fix $(u', q') \in \operatorname{gph} \mathcal{G}^+$. Then the inclusion

$$\chi^+_{k_n+1} u_{k_n+1} + (1 - \chi^+_{k_n+1}) u' \in \mathcal{G}^+(\chi^+_{k_n+1} z_{k_n} + (1 - \chi^+_{k_n+1}) q')$$

is satisfied almost everywhere on Ω. By Theorem 4.13, we obtain

$$\chi^+ u^* + (1 - \chi^+) u' \in \operatorname{conv}^\infty \mathcal{G}^+(\chi^+ z + (1 - \chi^+) q')$$

almost everywhere on Ω. Similarly, we obtain for $(u'', q'') \in \operatorname{gph} \mathcal{G}^-$

$$\chi^- u^* + (1 - \chi^-) u'' \in \operatorname{conv}^\infty \mathcal{G}^-(\chi^- z + (1 - \chi^-) q'')$$

and

$$(1 - \chi) u^* \in \mathcal{G}^0((1 - \chi) z)$$

almost everywhere, where χ_k and χ are as in Theorem 4.5 and (4.4). Note that $\operatorname{conv}^\infty \mathcal{G}^0 = \mathcal{G}^0$. By construction, $\chi^+_k + \chi^-_k = \chi_k$, which implies $\chi^+ + \chi^- = \chi$. Then we can combine all the inclusions above into one, which is

$$u^*(x) \in \left(\mathcal{G}_0 \cup \operatorname{conv}^\infty \mathcal{G}^+ \cup \operatorname{conv}^\infty \mathcal{G}^-\right)(-\nabla f(u^*)(x))$$

for almost all $x \in \Omega$. □

Interestingly, we can get rid of the convexification operator $\operatorname{conv}^\infty$ if we assume that the whole sequence $(\nabla f(u_k))$ converges pointwise almost everywhere.

Theorem 4.21 *Let (u_k) be a sequence of iterates generated by Algorithm 4.1 with weak limit point $u^* \in L^2(\Omega)$. Assume $\nabla f(u_k) \to \nabla f(u^*)$ pointwise almost everywhere. Then*

$$u^*(x) \in \mathcal{G}(-\nabla f(u^*)(x))$$

holds for almost all $x \in \Omega$.

Proof Denote $z(x) := -\nabla f(u^*)(x)$. Then $z_k(x) \to z(x)$ almost everywhere, which can be proven similarly to Corollary 4.6.

Let $(\tilde{z}, \tilde{u}) \notin \operatorname{gph} \mathcal{G}$. Since $\operatorname{gph} \mathcal{G}$ is closed, there is $\epsilon > 0$ such that

$$\left(B_\epsilon(\tilde{z}) \times B_\epsilon(\tilde{u})\right) \cap \operatorname{gph} \mathcal{G} = \emptyset.$$

Let $\epsilon' \in (0, \epsilon)$. Set

$$I := \{x \in \Omega : \ |\tilde{z} - z(x)| < \epsilon'\},$$

and

$$I_K := \{x \in I : \ |\tilde{z} - z_k(x)| < \epsilon \quad \forall k > K\}.$$

The sequence (I_K) is monotonically increasing. Since $z_k(x) \to z(x)$ for almost all $x \in \Omega$, we have $\cup_{K\in\mathbb{N}} I_K = I$. Here, the pointwise convergence of the whole sequence (z_k) is needed to conclude $I \subseteq \cup_{K\in\mathbb{N}} I_K$. Define

$$\begin{aligned} A_k^+ &:= \{x \in \Omega :\ u_k(x) \geq \tilde{u} + \epsilon\},\\ A_k^- &:= \{x \in \Omega :\ u_k(x) \leq \tilde{u} - \epsilon\}, \end{aligned}$$

and $\chi_k^+ := \chi_{A_k^+}$, $\chi_k^- := \chi_{A_k^-}$. By Lemma 4.18 above, we have $\sum_{k=1}^{+\infty} \|\chi_{k+1}^+ \chi_k^- + \chi_{k+1}^- \chi_k^+\|_{L^1(\Omega)} < +\infty$ and therefore $\chi_{k+1}^+ \chi_k^- + \chi_{k+1}^- \chi_k^+ \to 0$ in $L^1(\Omega)$ and pointwise almost everywhere.

Let $x \in I$. Then there is K such that $x \in I_K$. This implies $u_k(x) \notin B_\epsilon(\tilde{u})$ for all $k > K$. The sum $\sum_{k=K+1}^{+\infty} (\chi_{k+1}^+ \chi_k^- + \chi_{k+1}^- \chi_k^+)(x)$ counts the number of switches between values larger than $\tilde{u} + \epsilon$ and smaller than $\tilde{u} - \epsilon$ from $u_k(x)$ to $u_{k+1}(x)$. Since this sum is finite for almost all $x \in \Omega$, there is only a finite number of such switches. Then there is $K' > K$ such that either $u_k(x) \geq \tilde{u} + \epsilon$ for all $k > K'$ or $u_k(x) \leq \tilde{u} - \epsilon$ for all $k > K'$. Set

$$\begin{aligned} S_K^+ &:= \{x \in I :\ u_k(x) \geq \tilde{u} + \epsilon \quad \forall k > K\},\\ S_K^- &:= \{x \in I :\ u_k(x) \leq \tilde{u} - \epsilon \quad \forall k > K\}. \end{aligned}$$

The sequences (S_K^+) and (S_K^-) are increasing, and $\cup_{K\in\mathbb{N}}(S_K^+ \cup S_K^-) = I$.

Since $u_{k_n} \rightharpoonup u^*$, this implies $u^* \geq \tilde{u} + \epsilon$ on S_K^+ and $u^* \leq \tilde{u} - \epsilon$ on S_K^-. Since $\cup_{K\in\mathbb{N}}(S_K^+ \cup S_K^-) = I$, this implies

$$u^*(x) \notin B_\epsilon(\tilde{u})$$

for almost all $x \in I$, which implies

$$((z(x), u^*(x)) \notin B_{\epsilon'}(\tilde{z}) \times B_\epsilon(\tilde{u})$$

for almost all $x \in \Omega$. Since we can cover the complement of gph $\mathcal{G}$ by countably many such sets, the claim follows. □

Remark 4.22 If g is convex, the result above implies that u^* is a stationary point. This follows from the equivalence, see (4.9),

$$\begin{aligned} u^*(x) \in \mathcal{G}(-\nabla f(u^*)(x)) &\iff u^*(x) \in \operatorname{prox}_{L^{-1}g}\Big(u^*(x) - \frac{1}{L}\nabla f(u^*)(x)\Big)\\ &\iff u^*(x) \in \underset{v\in\mathbb{R}}{\arg\min}\ \nabla f(u^*)(x)v + \frac{L}{2}(v - u^*(x))^2 + g(v), \end{aligned}$$

which holds for almost all x. By convexity of g, this is equivalent to $-\nabla f(u^*)(x) \in \partial g(u^*(x))$, where ∂g is the convex subdifferential of g, see e.g., [3, Cor. 16.44].

4.4 Pointwise convergence of iterates

So far we were able to show that weak limit points of iterates (u_k) satisfy a certain inclusion in a pointwise sense. However, the resulting set in the limit might still be large or even unbounded in general. Assuming that $\mathcal{G}$ is (locally) single-valued on its components $\mathcal{G}^+, \mathcal{G}^-, \mathcal{G}^0$, we can show local pointwise convergence of a subsequence of iterates (u_{k_n}) to a weak limit point $u^* \in L^2(\Omega)$. In the next result this is illustrated for the map $\mathcal{G}^+$, however, it can be shown for the components $\mathcal{G}^-, \mathcal{G}^0$ similarly.

To this end, recall the definition of χ_k^+ in (4.11) and the fact that $\chi_k^+ \to \chi^+$ in $L^1(\Omega)$ and pointwise almost everywhere by Corollary 4.19.

Theorem 4.23 *Let $\bar{z} \in \operatorname{dom}(\mathcal{G}^+)$. Assume that $\mathcal{G}^+ : \mathbb{R} \to \mathbb{R}$ is single-valued and locally bounded on $B_\epsilon(\bar{z}) \cap \operatorname{dom}(\mathcal{G}^+)$ for some $\epsilon > 0$. Let $u_{k_n} \rightharpoonup u^*$ in $L^2(\Omega)$ and assume $\nabla f(u_{k_n})(x) \to \nabla f(u^*)(x)$ pointwise almost everywhere. Define the set*

$$I_\epsilon := \left\{ x \in \operatorname{supp}(\chi^+) : -\nabla f(u^*)(x) \in B_\epsilon(\bar{z}) \cap \operatorname{dom}(\mathcal{G}^+) \right\}.$$

Then

$$u_{k_n}(x) \to u^*(x)$$

holds for almost all $x \in I_\epsilon$. Furthermore, we have

$$u^*(x) \in \operatorname{prox}_{L^{-1}g}\left(\frac{1}{L}(Lu^*(x) - \nabla f(u^*)(x)) \right) \text{ f.a.a. } x \in I_\epsilon.$$

Proof By Corollary 4.6, we get $z_{k_n}(x) \to z(x) := -\nabla f(u^*)(x)$ pointwise almost every where. In addition, χ_k^+ converges to χ^+ pointwise almost everywhere.

Take $x \in I_\epsilon$ such that $z_{k_n}(x) \to z(x)$ and $\chi_k^+(x) \to \chi^+(x)$. Then there is $K > 0$ such that $|z_{k_n}(x) - \bar{z}| < \epsilon$ for all $k_n > K$. Since $x \in \operatorname{supp}(\chi^+)$, there is $K' > 0$ such that $x \in \operatorname{supp}(\chi_k^+)$ for all $k > K'$. Hence, for k_n sufficiently large we have

$$z_{k_n}(x) \in B_\epsilon(\bar{z}) \cap \operatorname{dom}(\mathcal{G}^+).$$

Since $\mathcal{G}^+$ is single-valued, locally bounded and outer semicontinuous in $B_\epsilon(\bar{z}) \cap \operatorname{dom}(\mathcal{G}^+)$, it is continuous, see also [24, Cor. 5.20].

This implies

$$\lim_{n\to+\infty} u_{k_n+1}(x) = \lim_{n\to+\infty} \mathcal{G}^+(z_{k_n}(x)) = \mathcal{G}^+(\lim_{n\to+\infty} z_{k_n}(x)) = \mathcal{G}^+(z(x)).$$

The continuity property mentioned above implies $\operatorname{conv}^\infty \mathcal{G}^+(z(x)) = \mathcal{G}^+(z(x))$. Then by Theorem 4.20, $\mathcal{G}^+(z(x)) = \{u^*(x)\}$, and the convergence $u_{k_n}(x) \to u^*(x)$ follows. Since $u_{k_n+1}(x) \to u^*(x)$ as well, we can pass to the limit in the inclusion (4.3), where we used the closedness of the graph of the proximal operator, which completes the proof. □

4.5 Strong convergence of iterates

Many optimal control problems of type (P) include a smooth cost functional of the form $u \to \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}$, $\alpha > 0$. In this section, we will consider the appearence of above regularization term as a special case and treat it explicitly in the convergence analysis. This allows us to obtain almost everywhere and strong convergence of a subsequence. Let $\tilde{g} : \mathbb{R} \to \mathbb{R}$ satisfy Assumption B and consider a sequence of iterates computed by

$$\begin{aligned} u_{k+1} \in \operatorname*{arg\,min}_{u \in L^2(\Omega)} & f(u_k) + \nabla f(u_k) \cdot (u - u_k) + \frac{L}{2}\|u - u_k\|^2_{L^2(\Omega)} \\ & + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} + \int_\Omega \tilde{g}(u(x))\,\mathrm{d}x. \end{aligned} \tag{4.13}$$

Note, that we do not include the term $\frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}$ into the functional f, as the derivative of this term is not completely continuous.

The solution to (4.13) is now given by

$$u_{k+1}(x) \in \operatorname{prox}_{\frac{1}{L+\alpha}\tilde{g}}\left(\frac{1}{L+\alpha}(Lu_k(x) - \nabla f(u_k)(x))\right)$$

for almost every $x \in \Omega$. All the analysis that was done in this section still applies in this case and all results can be transferred except for a possible change of notation. In particular, we adapt the set-valued map $\mathcal{G} : \mathbb{R} \rightrightarrows \mathbb{R}$ from Lemma 4.7 which is now defined by

$$u \in \mathcal{G}(z) :\Longleftrightarrow u \in \operatorname*{arg\,min}_{v \in \mathbb{R}} -zv + \frac{L}{2}(v - u)^2 + \frac{\alpha}{2}v^2 + \tilde{g}(v).$$

In the following it will be essential that $\operatorname{dom}(\tilde{g})$ is convex. Let us for simplicity assume $\operatorname{dom}(\tilde{g}) = [-b, b]$ with $b \in (0, +\infty]$, i.e., the subproblem (4.13) is equivalent to a box constrained optimization problem with the constraint $|u(x)| \le b$ for almost every $x \in \Omega$. To obtain strong convergence of iterates in $L^1(\Omega)$ and a L-stationary condition almost everywhere, we need to put stronger and more restricting assumptions on $\tilde{g}$. To this end, we introduce the following extension of Assumption B.

Assumption B$^+$

(B6) Ω has finite Lebesgue measure.
(B7) $g(u) = \frac{\alpha}{2}u^2 + \tilde{g}(u)$ with $\alpha > 0$ and $\tilde{g}$ satisfies Assumption B.
(B8) $\operatorname{dom}(\tilde{g}) = [-b, b]$ with $b \in (0, +\infty]$.
(B9) For all $s > 0$ there is $u_I := u_I(s) \in (0, b)$ such that $u \mapsto \frac{1}{2}u^2 + s\tilde{g}(u)$ is strictly convex on $[u_I, b]$.

In the rest of this section, we will assume that Assumption B$^+$ is satisfied. The goal is to express $\mathcal{G}$ in terms of single-valued continuous mappings $G^+, G^- : L^2(\Omega) \to L^2(\Omega)$, which will be derived below. Here, (B7)–(B9) are used to prove a corresponding result for the scalar-valued case, while (B6) is necessary to lift this to the $L^2(\Omega)$-case. Let us start with the following observation.

Lemma 4.24 *Let s_0, $u_0(\frac{1}{L+\alpha})$, $q_0(\frac{1}{L+\alpha})$ be as in Theorem 3.7, and let $u_I(\frac{1}{\alpha})$ be as in (B9). Assume $u_0(\frac{1}{L+\alpha}) \geq u_I(\frac{1}{\alpha})$ and $\frac{1}{L+\alpha} > s_0$. Then the following statements are true:*

(1) *If $u > 0$ and $u \in \mathcal{G}(z)$ for some $z \in \mathbb{R}$ then*

$$u = \underset{v\in[u_I(\frac{1}{\alpha}),b]}{\arg\min} -zv + \frac{\alpha}{2}v^2 + \tilde{g}(v).$$

(2) *Similarly, if $u < 0$ and $u \in \mathcal{G}(z)$ for some $z \in \mathbb{R}$ then*

$$u = \underset{v\in[-b,-u_I(\frac{1}{\alpha})]}{\arg\min} -zv + \frac{\alpha}{2}v^2 + \tilde{g}(v).$$

Proof Set $u_0 := u_0(\frac{1}{L+\alpha})$ and $u_I := u_I(\frac{1}{\alpha})$. Let us discuss the case $u \geq u_0$ only. If $u \in \mathcal{G}(z)$ for some $z \in \mathbb{R}$, then by definition of $\mathcal{G}$ and Theorem 3.7, $u \geq u_0 \geq u_I$. This implies

$$\begin{aligned} u &= \underset{v\in\mathbb{R}}{\arg\min} -zv + \frac{L}{2}(v-u)^2 + \frac{\alpha}{2}v^2 + \tilde{g}(v) \\ &= \underset{v\in[u_I,b]}{\arg\min} -zv + \frac{L}{2}(v-u)^2 + \frac{\alpha}{2}v^2 + \tilde{g}(v). \end{aligned} \tag{4.14}$$

Due to (B9), $u \mapsto \frac{1}{2}u^2 + \frac{1}{\alpha}\tilde{g}(u)$ is strictly convex on $[u_I, b]$ and therefore also (4.14) is a strictly convex optimization problem. In addition, all involved functions are continuous on (u_I, b) by Assumption B$^+$. Hence, u can be characterized in terms of the convex subdifferential by

$$0 \in -z + \partial\tilde{g}_\alpha(u),$$

where $\tilde{g}_\alpha(u) := \frac{\alpha}{2}u^2 + \tilde{g}(u)$.

Using again the convexity (B9), u is optimal to

$$\min_{v\in[u_I,b]} -zv + \tilde{g}_\alpha(v),$$

which is the claim. □

Let χ_k^+, χ_k^- be as in (4.11). Recall the result of Corollary 4.19: $\chi_k^+ \to \chi^+$ and $\chi_k^- \to \chi^-$ in $L^1(\Omega)$ and pointwise almost everywhere for some characteristic functions χ^+, χ^-.

Lemma 4.25 *Let s_0, $u_0(\frac{1}{L+\alpha})$, $q_0(\frac{1}{L+\alpha})$ as in Theorem 3.7. Suppose $u_0(\frac{1}{L+\alpha}) \geq u_I(\frac{1}{\alpha})$ and $\frac{1}{L+\alpha} > s_0$.*

Assume u_{k+1} is a global solution to (4.13) *with $|u_{k+1}(x)| \geq u_0(\frac{1}{L+\alpha})$ for almost all $x \in I_{k+1}$. Then there are continuous mappings $G^+, G^- : L^2(\Omega) \to L^2(\Omega)$ such that*

$$u_{k+1} = \chi^+_{k+1} G^+\left(\frac{z_k}{\alpha}\right) + \chi^-_{k+1} G^-\left(\frac{z_k}{\alpha}\right),$$

with $z_k \in L^2(\Omega)$ defined as in (4.8). *These mappings G^+, G^- are independent of L.*

Proof Throughout the proof, we use $u_I := u_I(\frac{1}{\alpha})$. Note that $\alpha > 0$ by (B7). Let us consider the case $u_{k+1}(x) > 0$ first. By Lemma 4.24, we have $u_{k+1}(x) \in \operatorname{prox}^+_{\alpha^{-1}\tilde{g}}\left(\frac{z_k(x)}{\alpha}\right)$, where we define the set $\operatorname{prox}^+_{\alpha^{-1}\tilde{g}}(z)$ for $z \in \mathbb{R}$

$$u \in \operatorname{prox}^+_{\alpha^{-1}\tilde{g}}(z) :\Longleftrightarrow u \in \operatorname*{arg\,min}_{v \in [u_I, b]} -zv + \frac{1}{2}v^2 + \alpha^{-1}\tilde{g}(v).$$

The latter optimization problem is convex by (B9) and therefore uniquely solvable. Thus, $\operatorname{prox}^+_{\alpha^{-1}\tilde{g}}(z)$ is single-valued for all $z \in \mathbb{R}$. By Lemma 3.2, $\operatorname{prox}^+_{\alpha^{-1}\tilde{g}}$ is outer semicontinuous on $\mathbb{R}$ and monotonically increasing. Let us prove its local boundedness. We have for $u \in \operatorname{prox}^+_{\alpha^{-1}\tilde{g}}(z)$ by optimality

$$-zu + \frac{1}{2}u^2 + \alpha^{-1}\tilde{g}(u) \leq -zu_I + \frac{1}{2}u_I^2 + \alpha^{-1}\tilde{g}(u_I).$$

Using $\tilde{g}(u) \geq 0$ and $-zu \geq -\frac{1}{4}u^2 - z^2$, we find

$$\frac{1}{4}u^2 \leq z^2 - zu_I + \frac{1}{2}u_I^2 + \alpha^{-1}\tilde{g}(u_I), \tag{4.15}$$

i.e., $\operatorname{prox}^+_{\alpha^{-1}\tilde{g}}$ is locally bounded. Then by [24, Corollary 5.20], $\operatorname{prox}^+_{\alpha^{-1}\tilde{g}}$ is a single-valued and continuous function on $\mathbb{R}$ and it holds

$$\chi^+_{k+1}(x)u_{k+1}(x) = \chi^+_{k+1}(x) \operatorname{prox}^+_{\alpha^{-1}\tilde{g}}\left(\frac{1}{\alpha}z_k(x)\right).$$

Similarly, we obtain

$$\chi^-_{k+1}(x)u_{k+1}(x) = \chi^-_{k+1}(x) \operatorname{prox}^-_{\alpha^{-1}\tilde{g}}\left(\frac{1}{\alpha}z_k(x)\right),$$

where the set $\operatorname{prox}^-_{\alpha^{-1}\tilde{g}}(z)$ for $z \in \mathbb{R}$ is defined by

$$u \in \operatorname{prox}^-_{\alpha^{-1}\tilde{g}}(z) :\Longleftrightarrow u \in \operatorname*{arg\,min}_{v \in [-b, -u_I]} -zv + \frac{1}{2}v^2 + \alpha^{-1}\tilde{g}(v).$$

Let us define $G^+, G^- : L^2(\Omega) \to L^2(\Omega)$ by

$$G^+(z)(x) := \operatorname{prox}^+_{\alpha^{-1}\tilde{g}}(z(x)) \text{ and } G^-(z)(x) := \operatorname{prox}^-_{\alpha^{-1}\tilde{g}}(z(x))$$

for $z \in L^2(\Omega)$, respectively. Then by a well-known result, see e.g. [1, Theorem 3.1], the superposition operators G^+ and G^- are continuous from $L^2(\Omega) \to L^2(\Omega)$ due to the growth condition (4.15), where we have used that Ω has finite measure by (B6). □

Now, we are able to prove strong convergence of a subsequence of (u_k) following the proof of [27, Thm. 3.17].

Theorem 4.26 *Suppose complete continuity of* ∇f and let $(u_k) \subset L^2(\Omega)$ *be a sequence generated by solving the corresponding subproblems* 4.13 *with weak limit point* u^*. *Under the same assumptions as in Lemma* 4.25, u^* *is a strong sequential limit point of* (u_k) in $L^1(\Omega)$.

Proof By Lemma 4.25 there exist continuous mappings $G^+, G^- : L^2(\Omega) \to L^2(\Omega)$ such that $u_{k+1} = \chi^+_{k+1} G^+\left(\frac{1}{\alpha} z_k\right) + \chi^-_{k+1} G^-\left(\frac{1}{\alpha} z_k\right)$. Let $u_{k_n} \rightharpoonup u^*$ in $L^2(\Omega)$. By Theorem 4.5 and complete continuity of ∇f, we obtain strong convergence of the sequence

$$z_{k_n} := -\big(\nabla f(u_{k_n}) + L(u_{k_n+1} - u_{k_n})\big) \to -\nabla f(u^*) =: z^*$$

in $L^2(\Omega)$ and $u_{k_n+1} \rightharpoonup u^*$ in $L^2(\Omega)$. In addition, we have by 4.19, $\chi^+_{k_n} \to \chi^+$ and $\chi^-_{k_n} \to \chi^-$ in $L^p(\Omega)$ for all $p < +\infty$, respectively. Hence, the convergence

$$u_{k_n+1} = \chi^+_{k_n+1} G^+\left(\frac{1}{\alpha} z_{k_n}\right) + \chi^-_{k_n+1} G^-\left(\frac{1}{\alpha} z_{k_n}\right) \to \chi^+ G^+\left(\frac{1}{\alpha} z^*\right) + \chi^- G^-\left(\frac{1}{\alpha} z^*\right)$$

in $L^1(\Omega)$ follows by Hölder's inequality. Since strong and weak limit points coincide, it follows $u_{k_n} \to u^*$ in $L^1(\Omega)$ and

$$u^* = \chi^+ G^+\left(\frac{z^*}{\alpha}\right) + \chi^- G^-\left(\frac{z^*}{\alpha}\right).$$

□

With the assumptions in Theorem 4.26 we can find an almost everywhere converging subsequence of iterates, i.e., $u_{k_n}(x) \to u^*(x)$ for almost every $x \in \Omega$. By the closedness of the mapping $\operatorname{prox}_{s\tilde{g}}$, we get

$$u^*(x) \in \operatorname{prox}_{\frac{1}{L+\alpha}\tilde{g}}\left(\frac{1}{L+\alpha}(Lu^*(x) - \nabla f(u^*)(x))\right) \text{f.a.a. } x \in \Omega, \tag{4.16}$$

i.e., u^* is L-stationary to the problem in almost every point. In the case $L = 0$, (4.16) is equivalent to

$$u^*(x) \in \operatorname*{arg\,min}_{u \in \mathbb{R}} f(u_k)(x)u(x) + \frac{\alpha}{2}|u(x)|^2 + \tilde{g}(u(x)) \text{ f.a.a. } x \in \Omega.$$

Hence, in this case u^* satisfies the Pontryagin maximum principle.

4.6 The proximal gradient method with variable step-size

The convergence results of this section require the knowledge of the Lipschitz modulus L_f of ∇f. This can be overcome by replacing the assumption $L > L_f$ by a suitable decrease condition.

Here we use the back-tracking algorithm from [5, Section 10.3.3]. Let us define (compare (4.3)) the set-valued map $T_L : L^2(\Omega) \rightrightarrows L^2(\Omega)$ by

$$T_L(u)(x) := \operatorname{prox}_{L^{-1}g}\left(\frac{1}{L}(Lu(x) - \nabla f(u)(x))\right).$$

Algorithm 4.27 (Proximal gradient with variable step-size) Choose $\eta > 0$, $L_0 > 0$, $\theta > 1$, and $u_0 \in L^2(\Omega)$. Set $k = 0$.

(1) Set $L_k := L_0\theta^j$ and $u_{k+1} := u_{k+1,j}$, where j is the smallest non-negative integer, for which with $u_{k+1,j} \in T_{L_0\theta^j}(u_k)$ the decrease condition

$$\eta\|u_{k+1,j} - u_k\|^2_{L^2(\Omega)} \le (f(u_k) + j(u_k)) - (f(u_{k+1,j}) + j(u_{k+1,j})) \tag{4.17}$$

is satisfied.

(2) Set $k := k + 1$, repeat.

Under Assumption A, the back-tracking strategy in step 1 of the algorithm above terminates after finitely many steps, as $L_0\theta^j - L_f \ge \eta$ for j large enough, compare also (4.6).

For the question on how to choose the minimizers $u_{k+1,j} \in T_{L_0\theta^j}(u_k)$ in the case of non-uniqueness, we refer to Remark 4.3.

The basic convergence result of Theorem 4.5 carries over to the variable step-size situation with the following modifications: The assumption $1/L > s_0$ has to be replaced by $(\limsup L_k)^{-1} > s_0$. The assumption $L > L_f$ is no longer necessary, it was used in the proof of Theorem 4.5 to prove (4.6), which has to be replaced by the decrease condition (4.17). The remaining convergence theory of Chapter 4 is much more technical to transfer, and will be discussed in a future publication.

5 Applications of the proximal gradient method

5.1 Optimal control with L^p control cost, $p \in (0, 1)$

In [27], the discussed proximal method was analyzed and applied to optimal control problems with L^0 control cost, i.e., $g(u) := \frac{\alpha}{2}u^2 + |u|_0$. In this section, we discuss the problem with $g(u) := \frac{\alpha}{2}u^2 + \beta|u|^p + \delta_{[-b,b]}(u)$, $p \in (0, 1)$, $b \in (0, +\infty]$ and consider

$$\min_{u \in L^2(\Omega)} f(u) + \frac{\alpha}{2}\|u\|_{L^2(\Omega)} + \beta \int_\Omega |u(x)|^p \,\mathrm{d}x \tag{5.1}$$

s.t.

$$u \in U_{ad} := \{u \in L^2(\Omega) : \ |u(x)| \le b \text{ a.e. in } \Omega\}$$

with $\alpha \ge 0$, $\beta > 0$. Ω is assumed to have finite Lebesgue measure. In terms of (4.13) with $\tilde{g}(u) := |u|^p + \delta_{[-b,b]}(u)$, the subproblem

$$\min_{u \in U_{ad}} f(u_k) + \nabla f(u_k)(u - u_k) + \frac{L}{2}\|u - u_k\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|_{L^2(\Omega)} + \beta \int_\Omega |u(x)|^p \,\mathrm{d}x \tag{5.2}$$

has to be solved in every iteration of Algorithm 4.1.

Similarly to Lemma 4.2, one can prove that u_{k+1} is a solution to (5.2) if and only if

$$u_{k+1}(x) \in \operatorname{prox}_{\frac{\beta}{L+\alpha}\tilde{g}}\left(\frac{1}{L+\alpha}(Lu_k(x) - \nabla f(u_k)(x))\right) f.a.a.x \in \Omega. \tag{5.3}$$

A visualization of the prox-map of $\tilde{g}$ is given below in Fig. 2. Due to Theorem 3.7 it holds $u_{k+1}(x) = 0$ or $|u_{k+1}(x)| \ge u_0$ for all k.

The particular choice of g allows us to compute the constant u_0 explicitly as a consequence of Lemma 3.6. By solving $\min_{u \neq 0} \frac{u}{2} + \frac{\beta}{L+\alpha}\frac{\tilde{g}(u)}{u}$ we get

$$u_0\left(\frac{\beta}{\alpha + L}\right) = \min\left(b, \left(\frac{\alpha + L}{2\beta(1-p)}\right)^{\frac{1}{p-2}}\right).$$

Moreover, we observe that with a suitable choice of parameters L and α, Assumption B$^+$ is satisfied such that we are able to apply Theorem 4.26 to the L^p problem to obtain a strongly convergent subsequence.

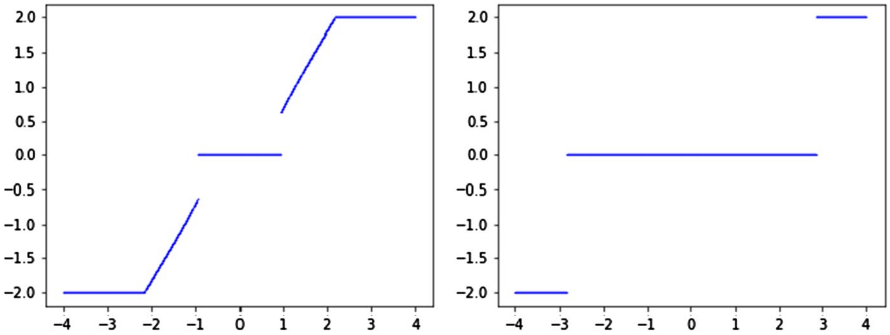


Fig. 2 The mapping $\operatorname{prox}_{s\tilde{g}}(q)$ for parameters $(s, b, p) = (0.5, 2, 0.5)$ (left) and $(s, b, p) = (3, 2, 0.3)$ (right) with $\tilde{g}(u) := |u|^p + \delta_{[-b,b]}$

Corollary 5.1 *Let* $\alpha > 0$ and (u_k) *a sequence of iterates. Furthermore, assume* $L \le (\frac{2}{p} - 1)\alpha$. *Then the assumptions of Theorem 4.26 are satisfied. If in addition* ∇f *is completely continuous from* $L^2(\Omega)$ *to* $L^2(\Omega)$, *then every weak sequential limit point* $u^* \in L^2(\Omega)$ *is a strong sequential limit point in* $L^1(\Omega)$.

Proof Let $k \in \mathbb{N}$. It holds $|u_{k+1}(x)| \ge u_0$ with $u_0 := \min\left(b, \left(\frac{\alpha+L}{2\beta(1-p)}\right)^{\frac{1}{p-2}}\right)$ on I_{k+1}. A short calculation yields that the assumptions on the parameters imply

$$\left(\frac{\alpha + L}{2\beta(1-p)}\right)^{\frac{1}{p-2}} \ge \left(\frac{\alpha}{\beta p(1-p)}\right)^{\frac{1}{p-2}} =: u_I.$$

Here, u_I is the positive point of inflection of

$$\min_{u:|u|\le b} -z_k(x)u + \frac{\alpha}{2}u^2 + \beta|u|^p$$

and it holds that $u \mapsto \frac{1}{2}u^2 + \frac{\beta}{\alpha}|u|^p$ is convex for all $q \in \mathbb{R}$ on $[u_I, +\infty)$ and $(-\infty, u_I)$, respectively, which corresponds to (B9). The claim now follows by Lemma 4.25 and Theorem 4.26. □

5.2 Optimal control with discrete-valued controls

Let us investigate the optimization problem with optimal control taking discrete values. That is, we choose $g(u)$ as the indicator function of integers, i.e., $g := \delta_{\mathbb{Z}}$. The problem (P) now reads

$$\min_{u\in L^2(\Omega)} f(u) + \int_\Omega \delta_{\mathbb{Z}}(u(x))\,\mathrm{d}x. \tag{5.4}$$

Note, this choice satisfies (B3.c). Applying Algorithm 4.1, the subproblem to solve is given by

$$\min_{u\in L^2(\Omega)} f(u_k) + \nabla f(u_k)(u - u_k) + \frac{L}{2}\|u - u_k\|^2_{L^2(\Omega)} + \int_\Omega \delta_{\mathbb{Z}}(u(x))\,\mathrm{d}x \tag{5.5}$$

and can be solved pointwise and explicitly. The analysis carried out in Chapter 4 is applicable. The special choice of g comes along with the following desirable result.

Lemma 5.2 *Let* $u_k, u_{k+1} \in U_{ad}$ *be consecutive iterates of Algorithm 4.1. Then*

$$\|u_{k+1} - u_k\|^p_{L^p(\Omega)} \ge \|u_{k+1} - u_k\|_{L^1(\Omega)}$$

holds for all $p \in [1, +\infty)$.

Proof The claim follows directly, since either $|u_{k+1}(x) - u_k(x)| = 0$ or $|u_{k+1}(x) - u_k(x)| \ge 1$ as the iterates are integer-valued in almost every point. □

Lemma 5.2 implies strong convergence of iterates (u_k) in $L^1(\Omega)$.

Theorem 5.3 *Let (u_k) be a sequence generated by Algorithm 4.1 with weak limit point u^*. Then $u_k \to u^*$* in $L^1(\Omega)$.

Proof As in the proof of Theorem 4.5, we get

$$\sum_{k=1}^{+\infty} \|u_{k+1} - u_k\|^2_{L^2(\Omega)} < +\infty$$

and therefore by Lemma 5.2

$$\sum_{k=1}^{+\infty} \|u_{k+1} - u_k\|_{L^1(\Omega)} \leq \sum_{k=1}^{+\infty} \|u_{k+1} - u_k\|^2_{L^2(\Omega)} < +\infty$$

Thus, (u_k) is a Cauchy sequence in $L^1(\Omega)$ and therefore convergent in $L^1(\Omega)$ and it holds $u_k \to u^*$. □

Corollary 5.4 *Let $g = \delta_{\mathbb{Z}}$. Then under the assumptions of Theorem 4.20, $u^*(x) \in \mathcal{G}(-\nabla f(u^*)(x))$ holds for almost all $x \in \Omega$, i.e., u^* is L-stationary.*

Proof The proof follows by passing to the limit in the inclusion (4.7). □

Let us mention some references that address optimal control problems with discrete valued controls. In [20], the constraint $u(x) \in \mathbb{Z} \cap [0, N]$ for some $N > 0$ is replaced by $u(x) \in [0, N]$. Then the resulting convex optimization problem is solved globally. Using a special algorithm, a sequence of discrete-valued controls is constructed that converges weakly to the solution of the relaxed problem. In [13] a penalization of the constraint $u(x) \in \mathbb{Z} \cap [0, N]$ is used and convexified. In both references [13, 20], the underlying partial differential equation is assumed to be linear. In [13], the function that corresponds to f in our paper is assumed to be quadratic in [13], while in [20] it is assumed that the global solution of the relaxed problem can be computed. These assumptions rule out the situation that f is non-convex. Another approach is taken in [10]. There the cost functional is assumed to be linear with a semilinear elliptic state equation such that the control-to-state map is concave, which is a very restrictive setting, too. The control is assumed to be in a bounded subset of $\mathbb{Z}^n$. In these references, finiteness of the admissible set enters the algorithms in an essential way, so that it is not clear how these results can be generalized to $u(x) \in \mathbb{Z}$.

6 Numerical experiments

In this section, we apply the proximal gradient method with variable step-size, Algorithm 4.27, to optimal control problems of type (P) and carry out numerical experiments for cost functionals with different g.

Let us introduce the reduced tracking-type functional

$$f_l(u) := \|S_l u - y_d\|^2_{L^2(\Omega)}, \tag{6.1}$$

where S_l is the weak solution operator of the linear Poisson equation

$$-\Delta y = u \text{ in } \Omega, y = 0 \text{ on } \partial\Omega. \tag{6.2}$$

Clearly, f is Fréchet differentiable, and $\nabla f(u)$ is linear in u. So Assumption A is satisfied.

We choose $\Omega := (0,1)^2$ to be the underlying domain in all following examples. To solve the partial differential equation, the domain is divided into a regular triangular mesh, and the PDE (6.2) is discretized with piecewise linear finite elements. The controls are discretized with piecewise constant functions on the triangles. The finite-element matrices were created with FEnicCS [19]. If not mentioned otherwise, the mesh-size is approximately $h = \sqrt{2}/160 \approx 0.00884$.

To determine the parameter L_k in each iteration, we use Algorithm 4.27 introduced in Sect. 4.6. This ensures decreasing objective values during the iterations. For all our tests we choose

$$\eta = 10^{-4}, \quad \theta = 2.$$

The stopping criterion is as follows:

If $|f(u_{k+1}) + j(u_{k+1}) - (f(u_k) + j(u_k))| \leq 10^{-12}$: STOP.

First, we consider control problems with L^p control cost, which were investigated in Sect. 5.1, i.e., $g(u) := |u|^p + \delta_{[-b,b]}$ with $p \in (0,1)$.

Example 1 Let $g(u) := |u|^p + \delta_{[-b,b]}$ for $p \in (0,1)$ and find

$$\min_{u \in L^2(\Omega)} J(u) := f_l(u) + \|u\|^2_{L^2(\Omega)} + \beta \int_\Omega g(u(x))\,\mathrm{d}x.$$

Setting $U_{ad} := \{u \in L^2(\Omega) : |u(x)| \leq b \text{ a.e. on } \Omega\}$ the problem is equivalent to

$$\min_{u \in U_{ad}} f_l(u) + \|u\|^2_{L^2(\Omega)} + \beta \int_\Omega |u(x)|^p\,\mathrm{d}x.$$

The first example is taken from [27], where the proximal gradient algorithm was investigated for (sparse) optimal control problems with $L^0(\Omega)$ control cost. Since $\int_\Omega |u|^p\,\mathrm{d}x \to \int_\Omega |u|^0\,\mathrm{d}x$ as $p \searrow 0$, we expect similar solutions. We choose the same problem data as in [18, 27]. That is, if not mentioned otherwise,

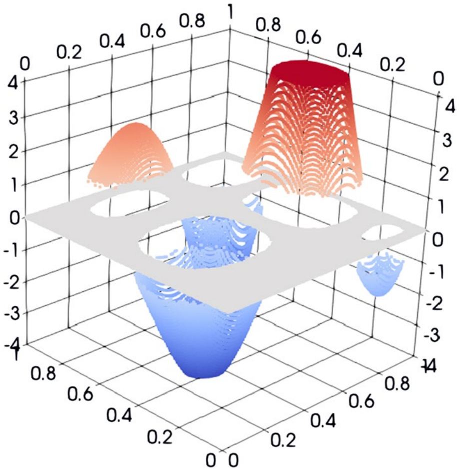
$$y_d(x,y) = 10x\sin(5x)\cos(7y)$$

and $\alpha = 0.01, \beta = 0.01, b = 4$.

A computed solution for $p = 0.8$ is shown in Fig. 3.

Let us comment on how we solved the subproblems (5.2) in practice to obtain the next iterate u_{k+1}. Recall that $u_{k+1}(x)$ is given by the global solution to the scalar-valued problem

Fig. 3 Solution u



$$\min_{|u|\le b} \nabla f(u_k)(x)u + \frac{L}{2}(u - u_k(x))^2 + \frac{\alpha}{2}u^2 + \beta|u|^p,$$

which cannot be computed explicitly in general for $p \in (0,1)$. In our tests, we used a simple gradient method. Its convergence can be guaranteed, since it can be easily determined on which interval the function above is convex, and whether its global minimum is at zero. Within the gradient method we use the standard Armijo backtracking to determine the step-size. It is stopped if the associated gradient of the minimization problem falls below 10^{-6}. In case that $\operatorname{prox}_{\frac{\beta}{L+\alpha}\tilde{g}}\left(\frac{1}{L+\alpha}(Lu_k(x) - \nabla f(u_k)(x))\right)$ is multivalued, we choose $u_{k+1}(x) = 0$. Note that, due to the monotonicity of $\tilde{g}$, this is only the case if $\left|\frac{1}{L+\alpha}(Lu_k(x) - \nabla f(u_k)(x))\right| = q_0$ for some $q_0 > 0$, compare also Lemma 3.6 and Theorem 3.7. Here, it holds $\operatorname{prox}_{\frac{\beta}{L+\alpha}\tilde{g}}(q_0) = \{0, u_0\}$. In the discretized problem, control functions are piecewise constant. Then the minimization in (5.2) decouples into independent minimization problems for each coefficient of the discretized control.

Convergence for decreasing p values In the following we consider solutions for different values of p. We use the same data and discretization as above. We set $L_0 = 0.0001$.

In Table 1 it can be seen that $J(u^*)$ and $j_p(u) := \int_\Omega |u^*|^p dx$ converge for decreasing values of p. The last row in Table 1 shows the result of applying the iterative hard-thresholding algorithm IHT-LS from [27] to the problem with $p = 0$, which is in agreement with our expectation. In the implementation we used a mesh-size of $h = \sqrt{2}/500 \approx 0.0028$.

In this table as well as the following ones, the column "#pde solves" refers to the number of pde solves, which where performed during all iterations. It includes additional pde solves due to the backtracking procedure. Hence, it can be interpreted as a measure of the computational effort.

Table 1 Decreasing values of p

p	$J(u^*)$	$j_p(u^*)$	#pde solves
0.5	5.3831	0.6711	15
0.3	5.3819	0.5725	15
0.1	5.3808	0.4841	15
0.01	5.3804	0.4482	15
0.001	5.3804	0.4448	15
0	5.38034	0.4445	15

Discretization Next, we solved the problem on different levels of discretization to investigate their influence. Here, we used $p = 0.5$ and $L_0 = 0.0001$. As can be seen in Table 2 the algorithm appears to be mesh independent.

Convergence in the case $L > (2/p - 1)\alpha$ So far, in every experiment the assumption on the parameters was naturally satisfied, such that strong convergence of iterates can be proven according to Theorem 5.1. The numerical results confirmed the theory. We will now investigate the case where the assumption is not satisfied, i.e., we choose parameters such that $L > (2/p - 1)\alpha$. In the following we present the result for the problem parameters

$$\alpha = 0.001, p = 0.9, L_0 = 0.005.$$

Furthermore, we set $b = 6$. In our computations the algorithm needed very long to reach the stopping criteria $|J(u_{k+1}) - J(u_k)| \leq 10^{-12}$ as can be seen in Table 3. This might be due to the parameter choice and the step-size strategy. For smaller mesh-sizes more iterations are needed.

Recall, the problem in the analysis that comes with this choice of parameters is that the map $\mathcal{G}$ in Lemma 4.7 is not necessarily single-valued anymore on the set of points where an iterate is not vanishing, see also Fig. 1. Let $u_I := u_I(\beta/\alpha) > 0$ denote the constant from (B9) and define the set

Table 2 Influence of mesh-size

h	$J(u^*)$	$j_p(u^*)$	#pde solves
0.071	5.2239	0.6371	13
0.035	5.3429	0.6581	15
0.0177	5.3732	0.6686	15
0.00884	5.3808	0.6704	15
0.00442	5.3827	0.6710	15
0.00221	5.3832	0.6711	15

Table 3 Performance for bad choice of parameters across different mesh-sizes

h	$J(u^*)$	$j_p(u^*)$	#pde solves
0.00884	5.3567	1.1246	395
0.00442	5.3567	1.1247	601
0.00221	5.3567	1.1253	821

$$\Omega_{I,k} := \{x \in \Omega : 0 < |u_k(x)| < u_I\}.$$

Then $\Omega_{I,k}$ is the set of points for which the crucial assumption in Lemma 4.25, which implies the decomposition of $\mathcal{G} \setminus \{0\}$ into single-valued mappings G^+, G^-, is not satisfied. In our numerical experiments, however, we made the observation that the measure of the set $\Omega_{I,k}$ is decreasing as $k \to +\infty$, see Fig. 4. Across different mesh-sizes h the measure decreases and tends to zero along the iterations.

Unfortunately, we were not able to prove such a behavior in the analysis and have no theoretical evidence whether this can be expected in general. However, assuming

$$|\Omega_{I,k}| \to 0$$

based on our numerical result, strong convergence of the sequence (u_k) can be concluded similar to Theorem 4.26.

Example 2 Let us now consider the problem

$$\min_{u \in U_{ad}} f_{sl}(u) + \|u\|^2_{L^2(\Omega)} + \beta \int_\Omega g(u(x))\,\mathrm{d}x$$

with $g(u) = |u|^p$, $p \in (0,1)$. with a semilinear PDE. Here, f_{sl} is given by the standard tracking type functional $u \mapsto \|y_u - y_d\|^2_{L^2(\Omega)}$, where y_u is the solution of the semilinear elliptic state equation

$$-\Delta y + y^3 = u \quad \text{in } \Omega, \qquad y = 0 \quad \text{on } \partial\Omega.$$

As argued in Sect. 2.2, Assumption A is satisfied for this example as well. This example can be found in [12] for semilinear control problems with L^1-cost. The data

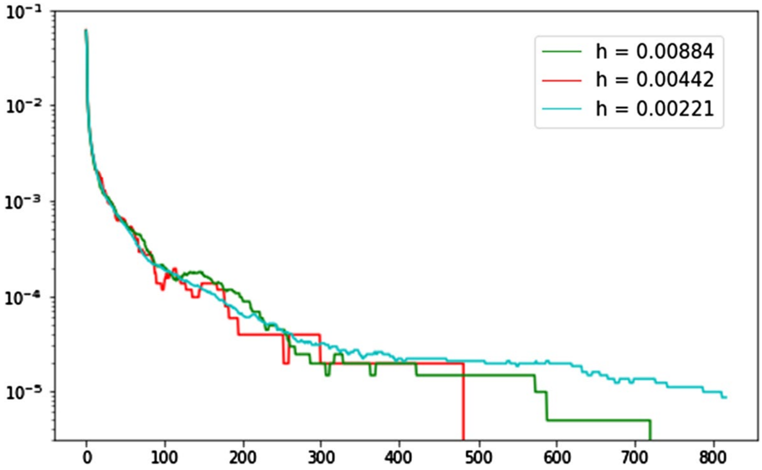


Fig. 4 Measure of $\Omega_{I,k}$ at iteration k for different discretization levels

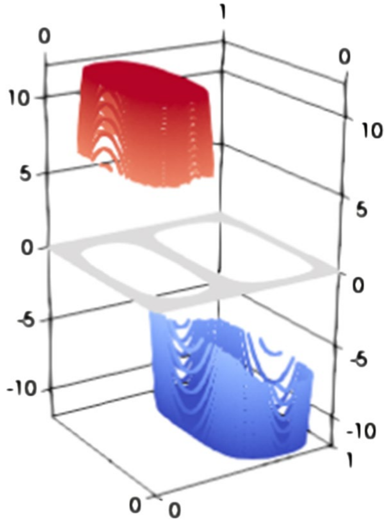


Fig. 5 Solution u of the semi-linear optimal control problem with $g(u) := |u|^{0.5}$

is given by $\alpha = 0.002$, $\beta = 0.03$, $b = 12$ and $y_d = 4\sin(2\pi x_1)\sin(\pi x_2)e^{x_1}$. We use the parameter $L_0 = 0.001$ (Fig. 5).

We made similar observations as in the case of a linear PDE concerning the influence of discretization and different values of p.

Example 3 In this last test, we consider the following linear elliptic optimal control problem with discrete-valued controls:

$$\min_{u \in U_{ad}} f_l(u) + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} + \int_\Omega g(u)\,\mathrm{d}x \tag{6.3}$$

with f_l given in (6.1) and $g(u) := \delta_{\mathbb{Z}}(u)$. The subproblem in Algorithm 4.27 can be solved pointwise and explicitly by computing

$$u_{k+1}(x) = \operatorname*{arg\,min}_{u \in [-b,b]\cap\mathbb{Z}} \nabla f(u_k)(x) + \frac{L}{2}(u - u_k(x))^2 + \frac{\alpha}{2}u^2,$$

where $\operatorname{prox}_{\delta_{\mathbb{Z}}}$ is given by rounding. In the case that above minimization problem is not uniquely solvable, we choose $u_{k+1}(x)$ as the minimizer with the smallest absolute value, i.e., we round towards zero. In Fig. 6, a solution plot of the optimal control is displayed. We adapted again the setting from Example 1 and used exactly the same problem data as before in Example 1, but set $b = 2$ and $L_0 = 0.001$. Again, we find the algorithm is robust with respect to the discretization in the sense of Table 2.

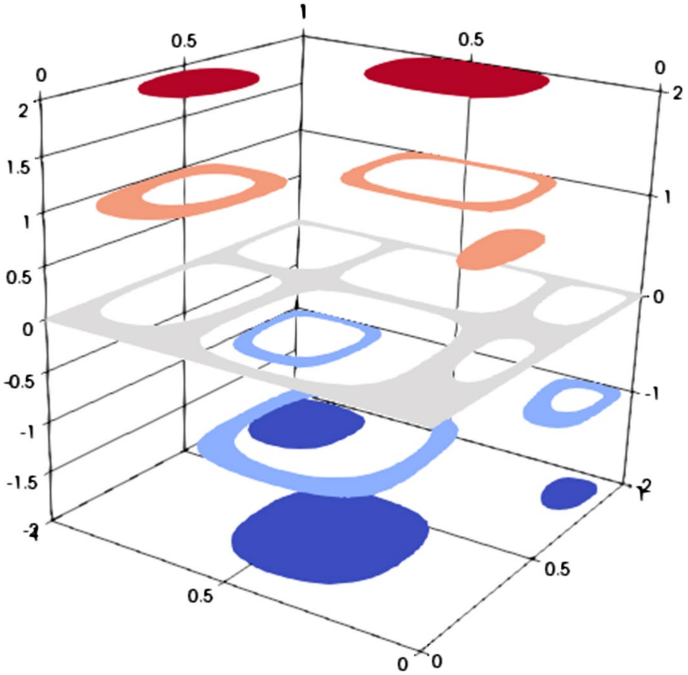


Fig. 6 Optimal control with discrete values

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Appell, J., Zabrejko, P.P.: Nonlinear Superposition Operators, Volume 95 of Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge (1990)
2. Aubin, J.-P., Frankowska, H.: Frankowska. Set-Valued Analysis, volume 2 of Systems & Control: Foundations & Applications. Birkhäuser Boston Inc, Boston (1990)
3. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York (2011)
4. Beck, A.: Introduction to Nonlinear Optimization, Volume 19 of MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Mathematical Optimization Society, Philadelphia (2014).. (Theory, algorithms, and applications with MATLAB)

5. Beck, A.: First-Order Methods in Optimization, Volume 25 of MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Mathematical Optimization Society, Philadelphia (2017)
6. Beck, A., Eldar, Y.C.: Sparsity constrained nonlinear optimization: optimality conditions and algorithms. SIAM J. Optim. **23**(3), 1480–1509 (2013)
7. Bonnans, J.F.: On an algorithm for optimal control using Pontryagin's maximum principle. SIAM J. Control Optim. **24**(3), 579–588 (1986)
8. Bredies, K., Lorenz, D.A., Reiterer, S.: Minimization of non-smooth, non-convex functionals by iterative thresholding. J. Optim. Theory Appl. **165**(1), 78–112 (2015)
9. Breitenbach, T., Borzì, A.: A sequential quadratic Hamiltonian method for solving parabolic optimal control problems with discontinuous cost functionals. J. Dyn. Control Syst. **25**(3), 403–435 (2019)
10. Buchheim, C., Kuhlmann, R., Meyer, C.: Combinatorial optimal control of semilinear elliptic PDEs. Comput. Optim. Appl. **70**(3), 641–675 (2018)
11. Casas, E.: Pontryagin's principle for optimal control problems governed by semilinear elliptic equations. In: Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena (Vorau, 1993), Volume 118 of International Series of Numerical Mathematics, pp. 97–114. Birkhäuser, Basel (1994)
12. Casas, E., Herzog, R., Wachsmuth, G.: Optimality conditions and error analysis of semilinear elliptic control problems with L^1 cost functional. SIAM J. Optim. **22**(3), 795–820 (2012)
13. Clason, C., Kunisch, K.: Multi-bang control of elliptic systems. Ann. Inst. H. Poincaré Anal. Non Linéaire **31**(6), 1109–1130 (2014)
14. Dunn, J.C.: On L^2 sufficient conditions and the gradient projection method for optimal control problems. SIAM J. Control Optim. **34**(4), 1270–1290 (1996)
15. Ekeland, I., Témam, R.: Convex Analysis and Variational Problems, Volume 28 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition (1999). (Translated from the French)
16. Geiersbach, C., Scarinci, T.: Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces. Comput. Optim. Appl. **78**(3), 705–740 (2021)
17. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints, Volume 23 of Mathematical Modelling: Theory and Applications. Springer, New York (2009)
18. Ito, K., Kunisch, K.: Optimal control with $L^p(\Omega)$, $p \in [0, 1)$ control cost. SIAM J. Control Optim. **52**(2), 1251–1275 (2014)
19. Langtangen, H.P., Logg, A.: Solving PDEs in Python, Volume 3 of Simula SpringerBriefs on Computing. Springer, Cham (2016). (The FEniCS tutorial I)
20. Manns, P., Kirches, C.: Multidimensional sum-up rounding for elliptic control systems. SIAM J. Numer. Anal. **58**(6), 3427–3447 (2020)
21. Nikolova, M.: Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. Multiscale Model. Simul. **4**(3), 960–991 (2005)
22. Nikolova, M., Ng, M.K., Zhang, S., Ching, W.-K.: Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. SIAM J. Imaging Sci. **1**(1), 2–25 (2008)
23. Qui, N.T., Wachsmuth, D.: Stability for bang–bang control problems of partial differential equations. Optimization **67**(12), 2157–2177 (2018)
24. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, Volume 317 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Berlin (1998)
25. Sakawa, Y., Shindo, Y.: On global convergence of an algorithm for optimal control. IEEE Trans. Autom. Control **25**(6), 1149–1153 (1980)
26. Tröltzsch, F.: Optimal Control of Partial Differential Equations. Theory, Methods and Applications. American Mathematical Society, Providence (2010)
27. Wachsmuth, D.: Iterative hard-thresholding applied to optimal control problems with $L^0(\Omega)$ control cost. SIAM J. Control Optim. **57**(2), 854–879 (2019)

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