Non-monotone proximal gradient methods in infinite-dimensional spaces with applications to non-smooth optimal control problems

Christian Kanzow, Daniel Wachsmuth

Abstract We are interested in solving non-smooth optimization problems in (possibly) infinite-dimensional Hilbert spaces. These problems have a special structure $\min f(x) + g(x)$, where f is assumed to be differentiable, and g is such that the associated prox-operator can be computed efficiently. We prove well-posedness of a spectral gradient method, and survey available convergence results. We study special choices of g, including the L^p-pseudo-norms with $p \in [0, 1)$. Numerical experiments show the applicability of the method.

1 Introduction

We are interested in solving optimization problems of the type: Minimize

$$\phi(x) := f(x) + g(x). \tag{1}$$

Here, $f : X \to \mathbb{R}$ is assumed to be Fréchet differentiable, where X is a Hilbert space, while $g : X \to \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is a given function. Note that we do not assume convexity or differentiability of g. Rather, we assume that the proximal operator to g is available, i.e., for all $x_0 \in X$ and $\gamma > 0$ the optimization problem

$$\min_{x \in X} \frac{1}{2\gamma} \|x - x_0\|_X^2 + g(x) \tag{2}$$

Christian Kanzow
University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany, e-mail: kanzow@mathematik.uni-wuerzburg.de

Daniel Wachsmuth
University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany, e-mail: daniel.wachsmuth@mathematik.uni-wuerzburg.de

is solvable. For the numerical realization, it is important that the solution of this problem can be computed efficiently. In this paper, we apply non-monotone proximal gradient methods. Given a current iterate x_k, the new iterate x_{k+1} is computed as solution of

$$\min f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2\gamma_k}\|x - x_k\|_X^2 + g(x),$$

where the parameter γ_k is determined by a non-monotone line-search method. Here, "non-monotone" refers to the property that the resulting sequence $(\phi(x_k))$ is not necessarily monotonically decreasing. Rather, the iterates are required to satisfy

$$\phi(x_{k+1}) \le \Phi_k - \frac{\sigma}{2\gamma_k}\|x_{k+1} - x_k\|_X^2,$$

where Φ_k is a merit function satisfying $\phi(x_k) \le \Phi_k$. The classical (monotone) proximal gradient method is recovered by setting $\Phi_k := \phi(x_k)$. In the seminal work [10], the choice

$$\Phi_k := \max_{r \in 0,\dots,m} \phi(x_{k-r})$$

was proposed. Another merit function was introduced in [26], which reads

$$\Phi_{k+1} := (1-p)\Phi_k + p \cdot \phi(x_{k+1}), \quad p \in (0,1),$$

so that Φ_{k+1} is a weighted average of all previous function values $\phi(x_0), \dots, \phi(x_{k+1})$. Both methods were initially proposed for unconstrained optimization. Proximal gradient methods with these choices of merit functions were analyzed in [8, 15]. Both references work in the finite-dimensional setting. In this work, we follow [8], which used the weighted approach of [26]. The advantage of this approach over the max-based one [10, 15] is the validity of the following claim

$$\sum_{k=0}^{\infty} \frac{1}{\gamma_k}\|x_{k+1} - x_k\|_X^2 < +\infty,$$

which is of importance for some of the problems we are interested in, see Section 3 below.

In particular, we are interested in applying the proximal gradient method to optimal control problems with non-smooth integral functionals. Therefore, we set $X := L^2(\Omega)$, with $\Omega \subset \mathbb{R}^d$. The function g is induced by a possibly non-convex function h as

$$g(x) := \int_\Omega h(x(\omega)) + \frac{\alpha}{2}|x|^2 d\omega,$$

where $h : \mathbb{R} \to \bar{\mathbb{R}}$ and $\alpha \ge 0$. We discuss the choices $h(x) = |x|_0$ (L^0-pseudo-norm), $h(x) = |x|^p$ with $p \in (0,1)$ (L^p-pseudo-norm), and $h(x) = I_{\mathbb{Z}}$ (indicator function of the integers) below in Section 3. All of them lead to non-smooth and non-convex problems, which got reasonable attention in the recent past. Problems with L^p-pseudo-norms ($p \in [0,1)$) were considered in [1, 11]. Monotone proximal gradient methods with a slightly different descent condition were analyzed in [19, 22]. Here, one has to emphasize that the convergence analysis of such method is much more challenging in infinite-dimensional spaces, as we cannot expect that the sequence of iterates has a (strongly) converging subsequence. Nevertheless some convergence results can be proven for these specific choices of h, we refer to Section 3 below. Optimal control problems with integer-valued controls were considered in, e.g., [7, 17, 19]. Other possible applications are problems with L^0-constraints [23] or switching constraints [6, 16].

While interesting in its own right, these proximal gradients can be used as building block in more complicated settings. In [9, 13] problems of the type $\min \phi(x)$ subject to $c(x) \in K$ were considered. Here, ϕ is as above, $c : X \to Y$ is Fréchet differentiable with values in another Hilbert space Y. In these works, an augmented Lagrange method was analyzed, where the differentiable constraint $c(x) \in K$ was penalized. The arising subproblems were solved by the proximal gradient method. The convergence theory was carried out in finite-dimensional spaces. Future research could be dedicated to generalize these methods to the infinite-dimensional setting, where for instance control problems with state constraints [21] or optimal control of the obstacle problem could be interesting.

2 The proximal gradient method with non-monotone linesearch in Hilbert spaces

Assumption 1 We assume the following properties of f, g:

- X is a real Hilbert space,
- $f : X \to \mathbb{R}$ is continuously Fréchet differentiable with locally Lipschitz continuous ∇f,
- $g : X \to \bar{\mathbb{R}}$ is proper, lower semicontinuous, and bounded from below by a continuous affine function,
- $\phi := f + g$ is bounded from below.

Definition 1 Let $x \in X$ be such that $g(x) < \infty$. We say that x is M-stationary if

$$0 \in \nabla f(x) + \partial g(x),$$

where ∂g is the limiting (or Mordukhovich) subdifferential.

Clearly, a local minimum of ϕ is M-stationary. We will see that M-stationarity plays an important role in the analysis of the proximal gradient method.

In this paper, we will use the following non-monotone proximal gradient method as sketched in Algorithm 1. Here, we follow [8] and choose a merit function, which is a weighted mean of the function values at previous iterates. Another popular choice is to use the maximum value of a finite number of function values at previous iterates [12, 13, 15], see also Remark 1 below.

Algorithm 1 Non-monotone proximal gradient method [8, Algorithm 3.1]

1: Choose $\tau \in (0,1)$, $\sigma \in (0,1)$, $0 < \gamma_{\min} < \gamma_{\max}$, $p \in (0,1)$, $x_0 \in \operatorname{dom} g$
2: Set $\Phi_0 := \phi(x_0)$
3: **for** $k \leftarrow 0, 1 \ldots$ **do** ▷ Outer loop
4: Choose $\gamma_{k,0} \in [\gamma_{\min}, \gamma_{\max}]$
5: **for** $i \leftarrow 0, 1 \ldots$ **do** ▷ Inner loop
6: Set $\gamma_{k,i} \leftarrow \gamma_{k,0}\tau^i$
7: Compute $x_{k,i}$ as solution of

$$\min f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2\gamma_{k,i}}\|x - x_k\|_X^2 + g(x)$$

8: **if** $x_{k,i}$ satisfies termination criterion **then**
9: **return**
10: **end if**
11: **if** $\phi(x_{k,i}) \le \Phi_k - \frac{\sigma}{2\gamma_{k,i}}\|x_{k,i} - x_k\|_X^2$ **then**
12: $x_{k+1} \leftarrow x_{k,i}$
13: $\gamma_k \leftarrow \gamma_{k,i}$
14: Set $\Phi_{k+1} \leftarrow (1-p)\Phi_k + p \cdot \phi(x_{k+1})$
15: **break**
16: **end if**
17: **end for**
18: **end for**

The initial step-size $\gamma_{k,0}$ is not required to be constant during the iteration. Hence, this method can use initial guesses provided by a Barzilai-Borwein approach [2, 8, 9, 13].

By a simple inductive proof, one can show that step 14 implies

$$\phi(x_k) \le \Phi_k,$$

so that the descent condition of Algorithm 1 is less restrictive than for the monotone method, which chooses $\Phi_k := \phi(x_k)$.

As termination criterion, [8, 13] propose to use

$$\left\| \nabla f(x_k) + \frac{1}{\gamma_{k,i}}(x_{k,i} - x_k) - \nabla f(x_{k,i}) \right\|_X \le \epsilon$$

with some tolerance $\epsilon \ge 0$, as this quantity measures the violation of M-stationarity for $x_{k,i}$.

Note that in the original publication [8], the Hilbert space X was assumed to be finite-dimensional. However, basic properties of the algorithm carry over directly to the infinite-dimensional case without change of proof. Due to the assumptions, the minimization problem in the inner loop is solvable. For well-posedness, we have the following result.

Lemma 1 *Let Assumption 1 be satisfied. Assume x_k is not M-stationary. Then the inner loop terminates in finitely many steps.*

Proof This is [8, Lemma 4.1], which does not rely on finite-dimensionality. The proof uses the observation that $\Phi_k \geq \phi(x_k)$, so that the condition in the linesearch is easier to satisfy than for the monotone method, which uses $\Phi_k := \phi(x_k)$. Thus finite termination follows from results for the monotone proximal gradient method [15, Lemma 3.1]. □

As in [8], we have the following basic convergence theorem, which is typical for proximal gradient methods.

Lemma 2 *Let Assumption 1 be satisfied. The sequences $(\phi(x_k))$ and (Φ_k) converge to some $\phi^* \geq \inf \phi$, where (Φ_k) is monotonically decreasing. In addition,*

$$\sum_{k=0}^{\infty} \frac{1}{\gamma_k} \|x_{k+1} - x_k\|_X^2 < +\infty,$$

and $\lim_{k\to\infty} \|x_{k+1} - x_k\|_X \to 0$.

Proof This is [8, Lemmas 4.2, 4.3], which does not rely on finite-dimensionality. The first claim is a consequence of the choice of Φ_k, the third claim follows from a telescoping sum argument, which is enabled by the special choice of Φ_k. □

In order to be able to pass to the limit in the optimality conditions of the inner problem, we need that

$$\frac{1}{\gamma_k} \|x_{k+1} - x_k\|_X \to 0$$

at least along the subsequence under consideration. Due to our infinite-dimensional setting, we cannot expect that the sequence of iterates has a strongly converging subsequence. Instead, we will work with weakly converging subsequences. This necessitates stronger assumptions on ∇f than Assumption 1.

Assumption 2 ∇f is completely continuous, i.e., $z_k \rightharpoonup z$ in X implies $\nabla f(z_k) \to \nabla f(z)$ in X.

Similar assumptions were used, e.g., in [19, 22], where also some examples are discussed. These assumptions are satisfied for standard classes of optimal control problems. Note that Assumptions 1 and 2 imply sequentially weak continuity of f, see Remark 3 at the end of this section.

Lemma 3 *Let Assumptions 1 and 2 be satisfied. Let (x_k) be a sequence generated by Algorithm 1. Let $(x_k)_{k\in K}$, $K \subset \mathbb{N}$, be a subsequence such that $x_k \rightharpoonup_K x^*$. Then $\gamma_k^{-1}\|x_{k+1} - x_k\|_X \to_K 0$.*

Proof The proof closely follows that of [8, Lemma 4.4] and [15, Proposition 3.2] It is immediate that the claim follows from Lemma 2 if the sequence $(\gamma_k)_{k\in K}$ is bounded away from zero. Hence, it remains to consider the case $\liminf_{k\in K} \gamma_k = 0$. By choosing another subsequence if necessary, we can assume $\gamma_k \to_K 0$ and $\gamma_k < \gamma_{\min}$ for all $k \in K$. Then the inner loop is performed at least twice, and the step-size γ_k/τ

does not satisfy the decrease condition. Define $\hat\gamma_k := \gamma_k/\tau$, and denote by $\hat x_{k+1}$ the solution of the problem in step 7 to $\hat\gamma_k$. Using standard arguments (e.g., from penalty methods), one can prove that $\|x_{k+1} - x_k\|_X \le \|\hat x_{k+1} - x_k\|_X$ for all $k \in K$. We will prove $\|\hat x_{k+1} - x_{k+1}\|_X \to_K 0$ and $\gamma_k^{-1}\|\hat x_{k+1} - x_k\|_X \to_K 0$.

Since $\hat x_{k+1}$ solves the problem in step 7, we have

$$f(x_k) + \nabla f(x_k)(\hat x_{k+1} - x_k) + \frac{1}{2\hat\gamma_k}\|\hat x_{k+1} - x_k\|_X^2 + g(\hat x_{k+1}) \le f(x_k) + g(x_k) = \phi(x_k). \tag{3}$$

Since $(f(x_k))_{k\in K}$ is convergent due to Remark 3, $(\phi(x_k))$ converges by Lemma 2, and g is bounded from below by an affine function, it follows that $\hat\gamma_k^{-1}\|\hat x_{k+1} - x_k\|_X^2$ is bounded, which implies $\hat x_{k+1} - x_k \to_K 0$.

By the mean-value theorem, there is ξ_k on the line between x_k and $\hat x_{k+1}$ such that

$$f(\hat x_{k+1}) = f(x_k) + \nabla f(\xi_k)(\hat x_{k+1} - x_k).$$

Using this in (3) implies

$$\phi(\hat x_{k+1}) + (\nabla f(x_k) - \nabla f(\xi_k))(\hat x_{k+1} - x_k) + \frac{1}{2\hat\gamma_k}\|\hat x_{k+1} - x_k\|_X^2 \le \phi(x_k).$$

Since $\hat x_{k+1}$ violates the decrease condition of step 11, we have

$$\phi(\hat x_{k+1}) > \Phi_k - \frac{\sigma}{2\hat\gamma_k}\|\hat x_{k+1} - x_k\|_X^2.$$

Combining the latter two inequalities results in

$$(\nabla f(x_k) - \nabla f(\xi_k))(\hat x_{k+1} - x_k) + \frac{1-\sigma}{2\hat\gamma_k}\|\hat x_{k+1} - x_k\|_X^2 \le \phi(x_k) - \Phi_k \le 0,$$

where we used also Lemma 2. Applying Cauchy-Schwarz and $\hat\gamma_k = \gamma_k/\tau$, shows

$$\frac{(1-\sigma)\tau}{2\gamma_k}\|\hat x_{k+1} - x_k\|_X \le \|\nabla f(x_k) - \nabla f(\xi_k)\|_X.$$

Since $x_k \rightharpoonup_K x^*$ and $\hat x_{k+1} - x_k \to_K 0$, it follows $\xi_k \rightharpoonup_K x^*$ and $\nabla f(x_k) - \nabla f(\xi_k) \to_K 0$ by Assumption 2. Hence, $\gamma_k^{-1}\|\hat x_{k+1} - x_k\|_X \to_K 0$, which implies $\gamma_k^{-1}\|x_{k+1} - x_k\|_X \to_K 0$ as $\|x_{k+1} - x_k\|_X \le \|\hat x_{k+1} - x_k\|_X$. □

As we cannot expect strong convergence of iterates, it is hard to prove M-stationarity of weak limit points of (x_k) in the general case. In particular, the assumptions of the corresponding convergence results [8, 15] are not satisfied for the integral functionals we have in mind.

Let us close this section with several remarks.

Remark 1 In [15] the merit function

$$\Phi_k := \max_{r=0,\dots,m_k} \phi(x_{k-r})$$

was used. The convergence properties of that choice are comparable to those of Algorithm 1 in the finite-dimensional case [13, 15]. However it seems impossible to prove the property

$$\sum_{k=1}^{\infty} \frac{1}{\gamma_k} \|x_k - x_{k+1}\|_X^2 < +\infty,$$

which is of some relevance for control problems with sparsity functionals [19, 22], see also Theorem 3 below.

Remark 2 In [13], g was chosen to be the indicator function of a closed set D. There a different termination criterion was used in the inner loop: The inner loop terminated as soon as $\phi(x_{k,i}) \leq \max_{r=0,\dots,m_k} \phi(x_{k-r}) - \sigma \nabla f(x_k)(x_{k,i} - x_k)$. This condition implies that our termination criterion is fulfilled, see [13, eq. (A.5)].

Remark 3 Let us prove that Assumptions 1 and 2 imply sequentially weak continuity of f. Let $x_k \rightharpoonup x$ in X. Due to Fréchet differentiability by Assumption 1, we can use the mean-value theorem to write $f(x_k) - f(x) = \nabla f(\xi_k)(x_k - x)$ with $\xi_k = \lambda_k x_k + (1 - \lambda_k)x$, $\lambda_k \in [0, 1]$. One can prove by a subsequence-subsequence argument that $\xi_k \rightharpoonup x$. Then $\nabla f(\xi_k) \to \nabla f(x)$ and $f(x_k) \to f(x)$.

3 Application to non-smooth optimal control problems

In this section, we will work with $X := L^2(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is an open and bounded set supplied with the Lebesgue measure. We are interested in setting

$$g(x) := \int_\Omega h(x(\omega)) + \frac{\alpha}{2}|x|^2 d\omega, \tag{4}$$

where $h : \mathbb{R} \to \mathbb{R}$ is lower-semicontinuous and possibly non-convex, and $\alpha > 0$. Since we want to use the Pontryagin maximum principle, we work with the Lebesgue measure on $\mathbb{R}^d$. In addition, we require that Assumption 1 is satisfied. Note that for non-convex h, the functional g is not weakly lower semicontinuous, and the existence of minima of $f + g$ cannot be proven in general. Let us emphasize that $\alpha > 0$ does not imply existence of minimizers, so it does not act as a classical regularization term.

Let now $\bar{x}$ be a local minimum of

$$\min_x f(x) + g(x).$$

Under additional assumptions on f, one can prove that the Pontryagin maximum principle is satisfied [19], which reads in our case: for all $x \in \mathbb{R}$ it holds

$$\bar{x}(\omega) \in \operatorname{argmin}_{x \in \mathbb{R}} \nabla f(\bar{x})(\omega) \cdot x + h(x) + \frac{\alpha}{2}|x|^2 \tag{5}$$

for almost all $\omega \in \Omega$. Since h is non-convex in general, the global minimum of the function to be minimized in (5) might be non-unique, hence the condition is written as an inclusion. If $\alpha > 0$ and h is bounded from below by an affine function then the minimization problem in (5) is always solvable.

Let now (x_k) be a sequence generated by Algorithm 1. Then x_{k+1} satisfies the Pontryagin maximum principle for the inner problem of step 7, which reads

$$x_{k+1}(\omega) \in \operatorname{argmin} \nabla f(x_k)(\omega) \cdot x + \frac{1}{2\gamma_k}(x - x_k(\omega))^2 + h(x) + \frac{\alpha}{2}|x|^2$$

for almost all $\omega \in \Omega$. This can be proven using the celebrated Lebesgue differentiation theorem, see also [19].

We will now reformulate these conditions in terms of inclusions. Define the set $H_{\mathrm{PMP}}(\alpha) \subset \mathbb{R}^2$ by:

$$(x, q) \in H_{\mathrm{PMP}}(\alpha) \;\Leftrightarrow x \in \operatorname{argmin} q \cdot x + h(x) + \frac{\alpha}{2}|x|^2. \tag{6}$$

If h is bounded from below by an affine function and $\alpha > 0$ then $H_{\mathrm{PMP}}(\alpha)$ is non-empty. If $\alpha > 0$ then $H_{\mathrm{PMP}}(\alpha)$ can be characterized as: $(x, q) \in H_{\mathrm{PMP}}(\alpha)$ if and only if $x \in \operatorname{prox}_{\alpha^{-1}g}(-\alpha^{-1}q)$. Then $x \in X$ satisfies the Pontryagin maximum principle if and only if

$$(x,\ \nabla f(x))(\omega) \in H_{\mathrm{PMP}}(\alpha)$$

for almost all $\omega \in \Omega$. Using the definition of H_{PMP} again, we see that the iterates of Algorithm 1 satisfy

$$(x_{k+1},\ \nabla f(x_k) - \gamma_k^{-1} x_k)(\omega) \in H_{\mathrm{PMP}}(\alpha + \gamma^{-1}) \tag{7}$$

In the sequel, we will analyze several choices of h. We will prove that under suitable assumptions on ∇f, M-stationary points are fixed points of the iteration, In the light of Lemma 1 this implies that the inner loop always terminates, as the termination criterion of step 11 is satisfied if $x_{k,i} = x_k$, see Lemma 2. Motivated by this observation, let us define the following notion of stationarity, see also [3].

Definition 2 Let $x \in L^2(\Omega)$ be given such that $x \in \operatorname{dom}(g)$. Let $\gamma > 0$. Then x is called γ-stationary if and only if

$$x \in \operatorname{argmin}_{y \in X} \nabla f(x)(y - x) + \frac{1}{2\gamma}\|y - x\|_X^2 + g(y).$$

In addition, we have the following simple consequences:

Lemma 4 *Let $x \in X$ with $x \in \operatorname{dom} g$ be given. Then it holds:*

1. *If x is γ-stationary for some $\gamma > 0$ then it is M-stationary and γ'-stationary for all $\gamma' \in (0, \gamma)$.*
2. *If x is γ-stationary for all $\gamma > 0$ then it satisfies the Pontryagin maximum principle.*

If x_k is a γ-stationary iterate of Algorithm 1 for some $\gamma > 0$, then $x_{k,i}$ can be chosen equal to x_k if $\gamma_{k,i} \leq \gamma$. Moreover, if g is convex then γ-stationarity implies global optimality. In addition, we can express γ-stationarity in terms of H_{PMP} as follows: x is γ-stationary if and only if

$$(x,\ \nabla f(x) - \gamma^{-1}x)(\omega) \in H_{\text{PMP}}(\alpha + \gamma^{-1}). \tag{8}$$

3.1 L^0-cost

First, we investigate the choice

$$h(x) := |x|_0 := \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \neq 0. \end{cases} \tag{9}$$

This choice of h penalizes the size of the support of x, i.e., the measure of $\{\omega : x(\omega) \neq 0\}$. Hence solutions are expected to have small support, which is important for some control applications [11, 14].

Let us investigate the notion of M-stationarity for this problem. The limiting subdifferential of h was computed in [18, Theorem 3.7].

Lemma 5 *The limiting subdifferential of $x \mapsto h(x) := \int_\Omega |x(\omega)|_0 d\omega$ is given by*

$$\partial h(x) = \{\eta \in L^2(\Omega) :\ \eta(\omega) = 0 \textit{ f.a.a. } \omega \textit{ with } x(\omega) \neq 0\}.$$

Using the chain rule for the limiting subdifferential yields

$$\partial g(x) = \alpha x + \partial h(x).$$

We can write M-stationarity as an inclusion as follows: x is M-stationary if and only if $-(\nabla f(x) + \alpha x) \in \partial h(x)$, or, equivalently,

$$(x,\ \nabla f(x))(\omega) \in H_M := \{(x, q) :\ x \cdot (\alpha x + q) = 0\}.$$

Interestingly, the limiting subdifferential gives no information on $\nabla f(x)(\omega)$ for points $\omega \in \Omega$ where $x(\omega) = 0$. This is not the case for the Pontryagin maximum principle and not for γ-stationarity. Let us compute the set $H_{\text{PMP}}(\alpha)$ as defined in (6). By elementary computations [11, Section 2.2][22, Lemma 3.5], we find for $\alpha > 0$

$$H_{\text{PMP}}(\alpha) := \left\{(x, q) :\ \alpha x + q = 0,\ |x| \geq \sqrt{\frac{2}{\alpha}}\right\} \cup \left\{(0, q) :\ |q| \leq \sqrt{2\alpha}\right\}. \tag{10}$$

Note that $H_{\text{PMP}}(0) = \{(0, 0)\}$. As argued above, see (8), x is γ-stationary if and only if

$$(x,\ \nabla f(x) - \gamma^{-1}x)(\omega) \in H_{\text{PMP}}(\alpha + \gamma^{-1}),$$

which can be rewritten as

$$(x,\ \nabla f(x))(\omega) \in H_\gamma := \\ \left\{(x,q):\ \alpha x + q = 0,\ |x| \geq \sqrt{\frac{2}{\alpha+\gamma^{-1}}}\right\} \cup \left\{(0,q):\ |q| \leq \sqrt{2(\alpha+\gamma^{-1})}\right\},$$

see also [22, Lemma 3.19]. Note that Lemma 4 implies $H_{\mathrm{PMP}} = \bigcap_{\gamma>0} H_\gamma$, $H_\gamma \subsetneq H_{\gamma'}$ for $0 < \gamma' < \gamma$, and

$$H_{\mathrm{PMP}}(\alpha) \subsetneq H_\gamma \subsetneq H_M$$

for $\gamma > 0$ and $\alpha \geq 0$. The graphs of these mappings can be seen in Figure 1.

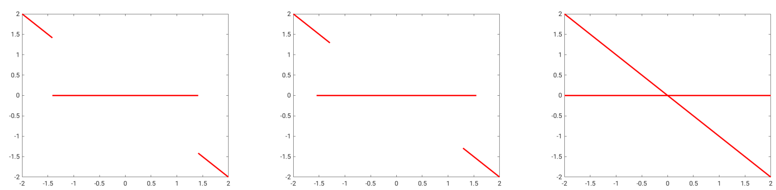

Fig. 1 Graphs of the set-valued maps H_{PMP}, H_γ, H_M (from left to right) for $\alpha = 1$, $\gamma = 5$ (Section 3.1)

Lemma 6 *Let g and h be given by (4) and (9). Let $x_k \in X$ be an iterate of Algorithm 1 such that $\nabla f(x_k) \in L^\infty(\Omega)$. Suppose that x_k is M-stationary, i.e., it satisfies $-(\nabla f(x_k) + \alpha x_k) \in \partial h(x_k)$. Then there is $\gamma > 0$ such that x_k is γ-optimal.*

Proof Since x_k is an iterate of Algorithm 1, it satisfies the Pontryagin maximum principle for the inner problem (7), which is equivalent to

$$(x_k,\ \nabla f(x_{k-1}) - \gamma_{k-1}^{-1} x_{k-1})(\omega) \in H_{\mathrm{PMP}}(\alpha + \gamma^{-1})$$

see also [19]. In particular, $x_k(\omega) \neq 0$ implies $|x_k(\omega)| \geq s > 0$ by (10), where $s := \sqrt{\frac{2}{\alpha+\gamma_{k-1}^{-1}}}$. By M-stationarity, $x_k(\omega) \neq 0$ implies $\alpha x_k(\omega) + \nabla f(x_k)(\omega) = 0$. Now set

$$\gamma := \min\left(\gamma_{k-1},\ \frac{2}{\|\nabla f(x_k)\|^2_{L^\infty(\Omega)}}\right).$$

Then

$$s \geq \sqrt{\frac{2}{\alpha+\gamma^{-1}}}, \quad \|\nabla f(x_k)\|_{L^\infty(\Omega)} \leq \sqrt{2(\alpha+\gamma^{-1})},$$

which implies $(x,\ \nabla f(x))(\omega) \in H_\gamma$ for almost all ω. $\square$

Theorem 3 *Let Assumption 1 be satisfied. Let g and h be given by (4) and (9) with $\alpha > 0$. Suppose that ∇f is completely continuous from $L^2(\Omega)$ to $L^q(\Omega)$ with $q > 2$, i.e., $z_k \rightharpoonup z$ in X implies $\nabla f(z_k) \to \nabla f(z)$ in $L^q(\Omega)$.*

Let (x_k) be a sequence generated by Algorithm 1 together with the sequence of parameters (γ_k). Suppose (γ_k) is bounded. Let $(x_k, \gamma_k)_{k\in K}$, $K \subset \mathbb{N}$, be a subsequence such that $x_k \rightharpoonup_K x^*$ and $\gamma_k \to_K \gamma > 0$.

Then $x_k \to_K x^*$ in $L^r(\Omega)$ for all $r < 2$, and x^* is γ-optimal.

Note that we assume that the whole sequence (γ_k) is bounded, although only converging subsequences are considered. This assumption is fulfilled, if we assume that ∇f is Lipschitz continuous on the set $\{x_k : k \in \mathbb{N}\}$, see [8, Corollary 4.5].

Proof (of Theorem 3) The proof follows along the lines of the proof of [22, Theorem 3.24]. Here, one has to replace the result of [22, Theorem 3.13] by that of Lemma 3 in the following way: Since (γ_k) is bounded by assumption, there is $M > 0$ such that $\gamma_k \leq M$, or equivalently, $\frac{1}{\gamma_k} \geq \frac{1}{M}$ for all $k \in \mathbb{N}$. Then Lemma 3 implies $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\|_X^2 < \infty$. Now the proof continues as in the proof of [22, Theorem 3.24]. □

The assumption of this convergence theorem are those of [22], where a monotone proximal gradient method was investigated using a different decrease condition. In the context of optimal control problems the condition of complete continuity of ∇f is not a severe restriction, see the examples discussed in [19, 22, 23].

3.2 L^p-cost, $p \in (0, 1)$

Now let us investigate the choice

$$h(x) := |x|^p, \tag{11}$$

where $p \in (0, 1)$, which is another method to promote solutions with small support [11]. The function h is lower semicontinuous and non-convex, so that the resulting integral functional g is lower semicontinuous but not weakly lower semicontinuous on $X = L^2(\Omega)$.

The limiting subdifferential of h was computed in [18, Theorem 4.6].

Lemma 7 *The limiting subdifferential of $x \mapsto h(x) := \int_\Omega |x(\omega)|^p \, d\omega$ is given by*

$$\partial h(x) = \{\eta \in L^2(\Omega) : \ \eta(\omega) = p|x(\omega)|^{p-2}x(\omega) \text{ f.a.a. } \omega \text{ with } x(\omega) \neq 0\}.$$

As in the previous section, we can express the different stationarity conditions using inclusions. First, we see that x is M-stationary if and only if

$$(x, \nabla f(x))(\omega) \in H_M := \{(x, q) : \ x \cdot (\alpha x + p|x|^{p-2}x + q) = 0\}$$

for almost all $\omega \in \Omega$. The point x satisfies the Pontryagin maximum principle if and only if [19, Section 5.1]

$$
\begin{aligned}
(x,\ \nabla f(x))(\omega) \in H_{\mathrm{PMP}}(\alpha) :=& \\
&\left\{(x,q) :\ x \neq 0,\ \alpha x + p|x|^{p-2}x + q = 0,\ |x| \geq x_0(\alpha)\right\} \\
&\qquad \cup \{(0,q) :\ |q| \leq q_0(\alpha)\}\,. \qquad (12)
\end{aligned}
$$

Here, $q_0(\alpha) > 0$ and $x_0(\alpha) > 0$ are such that the minimization problem

$$
\min_{x\in\mathbb{R}} -q_0(\alpha)\cdot x + h(x)
$$

has two global minima $x = 0$ and $x = x_0(\alpha)$. They are given by

$$
x_0(\alpha) := \left(\frac{2(1-p)}{\alpha}\right)^{\frac{1}{2-p}}, \quad q_0(\alpha) := \alpha x_0(\alpha) + p x_0(\alpha)^{p-1}.
$$

A point x is γ-stationary, see (8), if and only if $(x,\ \nabla f(x) - \gamma^{-1}x)(\omega) \in H_{\mathrm{PMP}}(\alpha + \gamma^{-1})$ for almost all ω. This can be equivalently written as

$$
\begin{aligned}
(x,\ \nabla f(x))(\omega) \in H_\gamma :=& \\
&\left\{(x,q) :\ x \neq 0,\ \alpha x + p|x|^{p-2}x + q = 0,\ |x| \geq x_0(\alpha+\gamma^{-1})\right\} \\
&\qquad \cup \left\{(0,q) :\ |q| \leq q_0(\alpha+\gamma^{-1})\right\}.
\end{aligned}
$$

The graphs of these mappings can be seen in Figure 2.

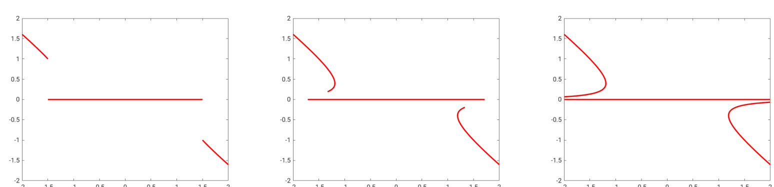

Fig. 2 Graphs of the set-valued maps H_{PMP}, H_γ, H_M (from left to right) for $\alpha = 1$, $p = 0.5$, $\gamma = 0.5$ (Section 3.2)

Again, we have for $\gamma > 0$

$$
H_{\mathrm{PMP}} \subsetneq H_\gamma \subsetneq H_M.
$$

Let us prove that M-stationary iterates are also γ-optimal.

Lemma 8 *Let g and h be given by (4) and (9). Let $x_k \in X$ be an iterate of Algorithm 1 such that $\nabla f(x_k) \in L^\infty(\Omega)$. Suppose that x_k is M-stationary, i.e., it satisfies $-(\nabla f(x_k) + \alpha x_k) \in \partial h(x_k)$. Then there is $\gamma > 0$ such that x_k is γ-optimal.*

Proof The proof is identical to that of Lemma 6. Since $q_0(\alpha+\gamma^{-1}) \to \infty$ for $\gamma \searrow 0$, there is $\tilde{\gamma} > 0$ such that $\|\nabla f(x_k)\|_{L^\infty(\Omega)} \le q_0(\alpha + \tilde{\gamma}^{-1})$. Then $\gamma := \min(\gamma_{k-1}, \tilde{\gamma})$ satisfies the claim. □

In contrast to the case $p = 0$, we do not have a convergence result like Theorem 3. Weak limit points of iterates of the monotone proximal gradient method satisfy a condition that is weaker than γ-stationarity, and that does not imply M-stationarity, see [19]. This is due to the convexifying effect of weak convergence and to the non-convexity of the images of the set-valued map

$$q \mapsto \left\{x : \ x > 0, \ \alpha x + p|x|^{p-2}x + q = 0, \ |x| \ge x_\gamma\right\},$$

for small γ. The graph of this mapping is a subset of H_γ, and its convex hull plays a role in the inclusion satisfied by weak limit points, see [19, Theorem 4.20]. We expect that a similar convergence result is true for the non-monotone method.

3.3 Integer-valued controls

As last example, we look at

$$h(x) = I_{\mathbb{Z}}(x) := \begin{cases} 0 & \text{if } x \in \mathbb{Z} \\ +\infty & \text{if } x \notin \mathbb{Z} \end{cases}, \tag{13}$$

which is the indicator function of the integers. If $g(x) < +\infty$ then $x(\omega) \in \mathbb{Z}$ for almost all ω, Again, we will need $\alpha > 0$.

Let us first compute the limiting subdifferential of h.

Lemma 9 *Let $r \in [1, \infty)$. Define $h : L^r(\Omega) \to \bar{\mathbb{R}}$ by*

$$h(x) := \int_\Omega I_{\mathbb{Z}}(x(\omega))d\omega$$

Let $x \in L^r(\Omega)$ with $h(x) < +\infty$. Then the following claims hold:

1. *If $r \in (1, \infty)$ then the limiting subdifferential satisfies*

$$\partial h(x) = L^{r'}(\Omega) = L^r(\Omega)^*,$$

where r' is such that $\frac{1}{r} + \frac{1}{r'} = 1$.

2. *In case $r = 1$, we have*

$$\partial h(x) = \{0\}.$$

Proof Let us denote the Fréchet subdifferential of j by $\hat\partial j$. Let $r > 1$. We show that $L^\infty(\Omega) \subset \hat\partial h(x)$. Let $\eta \in L^\infty(\Omega)$. Let $y \in L^r(\Omega)$ with $h(y) = 0$ be given. Since y and x are integer-valued almost everywhere, it follows

$$\|x-y\|^r_{L^r(\Omega)} \geq 1^{r-1}\|x-y\|_{L^1(\Omega)}.$$

Then

$$\int_\Omega |\eta(y-x)|d\omega \leq \|\eta\|_{L^\infty(\Omega)}\|y-x\|_{L^1(\Omega)} \leq \|\eta\|_{L^\infty(\Omega)}\|y-x\|^r_{L^r(\Omega)},$$

which implies

$$\lim_{\|y-x\|_{L^r(\Omega)}\to 0} \frac{h(y)-h(x)-\int_\Omega \eta(y-x)d\omega}{\|y-x\|_{L^r(\Omega)}} = 0,$$

and hence $\eta \in \hat\partial h(x)$. Since $L^\infty(\Omega)$ is dense in $L^{r'}(\Omega)$, it follows that $\partial h(x) = L^{r'}(\Omega)$.

Now consider the case $r = 1$. As $h(y) \geq 0$ for all $y \in L^r(\Omega)$, it follows $0 \in \hat\partial h(x)$. Take $\eta \in L^\infty(\Omega) \setminus \{0\}$. Then there is $\sigma > 0$ and $A \subset \Omega$ of positive measure such that $|\eta| \geq \sigma > 0$ almost everywhere on A. Now chose $B_k \subset A$ such that $|B_k| \to 0$. Define $x_k := x + \chi_{B_k}\operatorname{sign}(\eta)$. Note that $x_k \in L^1(\Omega)$, $x_k \to x$ in $L^1(\Omega)$, $x_k(\omega) \in \mathbb{Z}$ for almost all t, and hence $h(x_k) = 0$. In addition,

$$\int_\Omega \eta(x_k - x)d\omega = \int_{B_k} |\eta| d\omega \geq \sigma|B_k| = \sigma\|x_k - x\|_{L^1(\Omega)}$$

This implies

$$\liminf_{\|y-x\|_{L^1(\Omega)}\to 0} \frac{h(y)-h(x)-\int_\Omega \eta(y-x)d\omega}{\|y-x\|_{L^1(\Omega)}} \leq -\sigma < 0,$$

and $\eta \notin \hat\partial h(x)$. This shows $\hat\partial h(x) = \{0\}$. By [5, Theorem 3.2], we get $\hat\partial h(x) = \partial h(x)$, which is the claim. □

Hence, M-stationarity does not give any information, as all $x \in \operatorname{dom} g$ and hence all iterates of Algorithm 1 are M-stationary. With the notation of the previous sections, we have $H_M = \mathbb{Z} \times \mathbb{R}$.

The Pontryagin maximum principle can be rephrased as: x satisfies the maximum principle if and only if

$$(x,\, \nabla f(x))(\omega) \in H_{\text{PMP}}(\alpha) := \left\{ (x,q) :\ x \in \operatorname{round}\left(-\frac{1}{\alpha}q\right)\right\},$$

where $\operatorname{round}(x)$ is the set of nearest integers to $x \in \mathbb{R}$, i.e., $\operatorname{round}(0.5) = \{0,1\}$. Similarly to the computations in the previous section, we can characterize a γ-stationary point x by the inclusion

$$(x,\ \nabla f(x) - \gamma^{-1}x)(\omega) \in \left\{(x, q) :\ x \in \operatorname{round}\left(-\frac{1}{\alpha+\gamma^{-1}}(q-\gamma^{-1}x)\right)\right\},$$

which is equivalent to

$$(x,\ \nabla f(x))(\omega) \in H_\gamma := \left\{(x, q) :\ x \in \mathbb{Z},\ \frac{\alpha}{\alpha+\gamma^{-1}}\left|x - \frac{1}{\alpha}q\right| \le \frac{1}{2}\right\}.$$

Hence, every iterate x_k of Algorithm 1 with $x_k, \nabla f(x_k) \in L^\infty(\Omega)$ is γ-optimal if

$$\alpha\left(\left|x_k(\omega) - \alpha^{-1}\nabla f(x_k)(\omega)\right| - \frac{1}{2}\right) \le \frac{1}{2\gamma}$$

for almost all ω, i.e., if $\gamma > 0$ is small enough. In addition, following [19] one can prove that if (γ_k) is bounded then the iterates (x_k) converge strongly in $L^1(\Omega)$:

Lemma 10 *Let Assumption 1 be satisfied. Let g and h be given by (4) and (13) with $\alpha > 0$. Let (x_k) be a sequence generated by Algorithm 1 together with the sequence of parameters (γ_k). Suppose (γ_k) is bounded. Then $x_k \to_K x^*$ in $L^1(\Omega)$.*

Proof Let $M > 0$ such that $\gamma_k \le M$ for all k. Due to Lemma 2, we have

$$\frac{1}{M}\sum_{k=0}^{\infty}\|x_{k+1} - x_k\|_X^2 \le \sum_{k=0}^{\infty}\frac{1}{\gamma_k}\|x_{k+1} - x_k\|_X^2 < +\infty,$$

The iterates (x_k) are integer valued, which implies $|x_{k+1}(\omega) - x_k(\omega)| \in \mathbb{Z}$ for almost all ω, and $\|x_{k+1} - x_k\|_{L^1(\Omega)} \le \|x_{k+1} - x_k\|^2_{L^2(\Omega)}$. Using this in the above inequality, we obtain $\sum_{k=0}^{\infty}\|x_{k+1} - x_k\|_{L^1(\Omega)} < \infty$. Hence (x_k) is a Cauchy-sequence in $L^1(\Omega)$, and thus converging. □

4 Numerical experiments

We conducted some numerical experiments for the following problem:

$$\min \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}$$

where $y \in H_0^1(\Omega)$ is the weak solution of the partial differential equation (pde)

$$-\Delta y = u \text{ in } \Omega$$

and u is integer-valued, i.e.,

$$u(\omega) \in \mathbb{Z} \ \text{ for almost all } \omega \in \Omega.$$

Here, $\Omega \subset \mathbb{R}^d$ is a bounded domain, $y_d \in L^2(\Omega)$ is a given desired state, and $\alpha > 0$. For the computations, we use $\Omega = (0,1)^2$ and $\alpha = 10^{-3}$.

The partial differential equation has a unique weak solution $y_u \in H_0^1(\Omega)$ for all $u \in L^2(\Omega)$, so with the choices $f(u) := \frac{1}{2}\|y - y_d\|_{L^2(\Omega)}$ and $g(u) := \frac{\alpha}{2}\|u\|_{L^2(\Omega)} + \int_\Omega I_{\mathbb{Z}}(u(\omega))d\omega$ this problem fits into the problem class considered in Section 3.3. It is well-known [21] that $\nabla f(u)$ can be computed as $\nabla f(u) = p$, where the so-called adjoint state $p \in H_0^1(\Omega)$ is the unique weak solution of

$$-\Delta p = y - y_d.$$

Hence, the computation of $f(u)$ requires one pde solve, while the joint computation of $(f(u), \nabla f(u))$ requires two pde solves.

State and adjoint variables were discretized using continuous piecewise linear functions on a uniform grid with $h = 1.41 \cdot 10^{-3}$, while the control variable was discretized using piecewise constant functions. Let us remark that for the finest discretization, the control functions have 2, 000, 000 degrees of freedom. Hence, the discretization results in an optimization problem with 2, 000, 000 integer variables.

We used Algorithm 1 with the parameter choice from [8]:

$$\tau = 0.5,\ \sigma = 0.999,\ \gamma_{\min} = 10^{-12},\ \gamma_{\max} = 10^{+12},\ p = 0.2.$$

We tested two different choices to compute the initial step-size $\gamma_{0,k}$: the first choice uses the previous step-size with enlargement:

$$\gamma_{0,k+1} := \mathrm{proj}_{[\gamma_{\min},\gamma_{\max}]}(\gamma_k \cdot \tau^{-1}), \tag{14}$$

the second choice is a spectral Barzilai-Borwein step-size as in [8, Section 5]:

$$\gamma_{0,k+1} := \mathrm{proj}_{[\gamma_{\min},\gamma_{\max}]}\left(\frac{\|x_k - x_{k-1}\|_X^2}{\langle \nabla f(x_k) - \nabla f(x_{k-1}),\ x_k - x_{k-1}\rangle_X}\right). \tag{15}$$

In addition, we implemented the max-strategy to define a new threshold Φ_k [13], where

$$\Phi_k := \max_{r=0,\dots,\min(k,m)} \phi(x_{k-r}) \tag{16}$$

with $m = 5$. And we compared the non-monotone schemes with the monotone one obtained by the choice

$$\Phi_k := \phi(x_k). \tag{17}$$

The termination criterion in step 8 of Algorithm 1 was:

$$\textbf{if } \left\|\frac{1}{\gamma_{k,i}}(x_{k,i} - x_k)\right\|_X \le 2\epsilon \textbf{ and } \left\|\nabla f(x_k) + \frac{1}{\gamma_{k,i}}(x_{k,i} - x_k) - \nabla f(x_{k,i})\right\|_X \le \epsilon \textbf{ return}.$$

where we set $\epsilon := 10^{-4}$. We used this modification of the termination criterion in [8, 13] to save (costly) computations of $\nabla f(x_{k,i})$ in the inner loop when the difference between x_k and $x_{k,i}$ is too large.

We tested these algorithms with different initial guesses and choices of y_d. For each initial value and desired state, we run each of these different methods. The inital choice was generated using `x0 = 50*(2*rand(N,1)-1);`, where N is degrees of freedom of the control variable, while `rand` is Matlab's built-in. For the desired state, we took perturbations of

$$\hat{y}_d(x) := 10x_1 \sin(5x_1) \cos(7x_2).$$

That is, we set $y_d(x_i) := \hat{y}_d(x_i) + \texttt{rand}$ for each node x_i of the triangulation. In this way, we generated a set of 100 random initial guesses and desired states. Then each method was run for each of these choices.

In the performance plots in Figure 3 below, we will use the following labels to distinguish the different choices for initial step-sizes and merit function update:

- **weighted:** Φ_k as in Algorithm 1, $\gamma_{0,k+1}$ as in (14),
- **weighted, spectral:** Φ_k as in Algorithm 1, $\gamma_{0,k+1}$ as in (15),
- **max:** Φ_k as in (16), $\gamma_{0,k+1}$ as in (14),
- **max, spectral:** Φ_k as in (16), $\gamma_{0,k+1}$ as in (15),
- **mono:** Φ_k as in (17), $\gamma_{0,k+1}$ as in (14),
- **mono, spectral:** Φ_k as in (17), $\gamma_{0,k+1}$ as in (15),

The results can be seen in Figure 3. We plotted performance graphs to show

- the best function value, i.e., $\min_k \phi(x_k)$,
- the iteration, when the best value was obtained, i.e., $\operatorname{argmin}_k \phi(x_k)$,
- the number of outer iterations,
- the number of pde solves, which serves as indication of running time, as these pde solves are the most expensive parts of the iterations.

Interestingly, all spectral variants of the algorithms reach the same best function value in all the test cases. In fact, they reach their minimal value after a small number of iterations, while spending some more iterations until the termination criterion is reached. Of the three spectral variants, the weighted versions spent many additional inner iterations until the termination criterion was reached, which resulted in a much higher number of pde solves than the other spectral versions. Here, more research is needed to investigate better termination criteria.

Acknowledgements

This research was partially funded by German Research Foundation (DFG) within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962) under grant numbers KA 1296/24-2 and WA 3626/3-2, respectively.

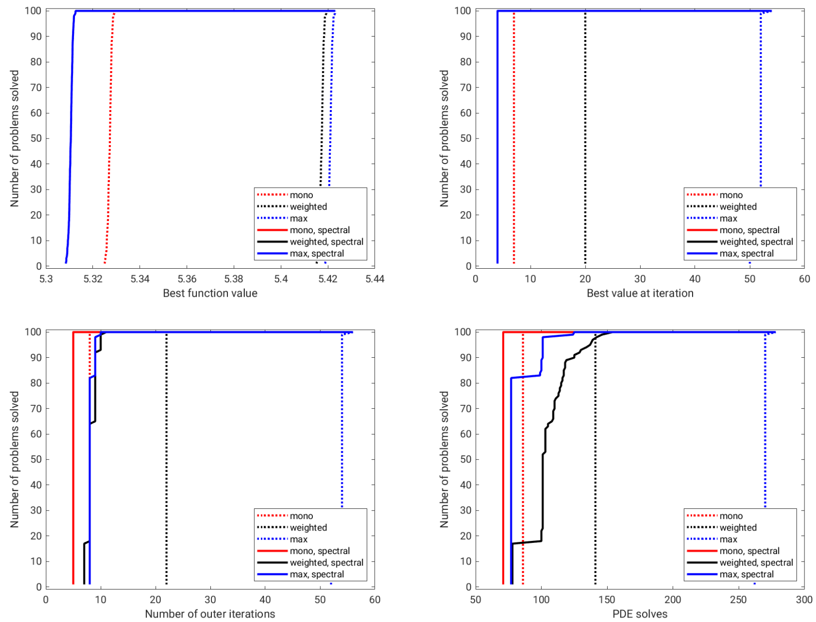

Fig. 3 Comparing variants of Algorithm 1 with monotone/non-monotone linesearch and plain/spectral stepsizes

References

1. Antil, H., Wachsmuth, D.: Sparse optimization problems in fractional order Sobolev spaces. Inverse Problems **39**(4), Paper No. 044001, 17 (2023)
2. Azmi, B., Kunisch, K.: Analysis of the Barzilai-Borwein step-sizes for problems in Hilbert spaces. J. Optim. Theory Appl. **185**(3), 819–844 (2020). DOI 10.1007/s10957-020-01677-y
3. Beck, A., Eldar, Y.C.: Sparsity constrained nonlinear optimization: optimality conditions and algorithms. SIAM J. Optim. **23**(3), 1480–1509 (2013). DOI 10.1137/120869778
4. Casas, E., Wachsmuth, D.: First and second order conditions for optimal control problems with an L^0 term in the cost functional. SIAM J. Control Optim. **58**(6), 3486–3507 (2020). DOI 10.1137/20M1318377
5. Chieu, N.H.: The Fréchet and limiting subdifferentials of integral functionals on the spaces $L_1(\Omega, E)$. J. Math. Anal. Appl. **360**(2), 704–710 (2009). DOI 10.1016/j.jmaa.2009.07.017
6. Clason, C., Ito, K., Kunisch, K.: A convex analysis approach to optimal controls with switching structure for partial differential equations. ESAIM Control Optim. Calc. Var. **22**(2), 581–609 (2016). DOI 10.1051/cocv/2015017
7. Clason, C., Kunisch, K.: Multi-bang control of elliptic systems. Ann. Inst. H. Poincaré C Anal. Non Linéaire **31**(6), 1109–1130 (2014). DOI 10.1016/j.anihpc.2013.08.005
8. De Marchi, A.: Proximal gradient methods beyond monotony. J. Nonsmooth Anal. Optim. **4**, Paper No. 10290, 18 (2023)
9. De Marchi, A., Jia, X., Kanzow, C., Mehlitz, P.: Constrained composite optimization and augmented Lagrangian methods. Math. Program. **201**(1-2, Ser. A), 863–896 (2023). DOI 10.1007/s10107-022-01922-4

10. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton's method. SIAM J. Numer. Anal. **23**(4), 707–716 (1986). DOI 10.1137/0723046
11. 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). DOI 10.1137/120896529
12. Jia, X., Kanzow, C., Mehlitz, P.: Convergence analysis of the proximal gradient method in the presence of the Kurdyka-Lojasiewicz property without global Lipschitz assumptions (2023). URL https://arxiv.org/abs/2301.05002
13. Jia, X., Kanzow, C., Mehlitz, P., Wachsmuth, G.: An augmented Lagrangian method for optimization problems with structured geometric constraints. Math. Program. **199**(1-2, Ser. A), 1365–1415 (2023). DOI 10.1007/s10107-022-01870-z
14. Kalise, D., Kunisch, K., Sturm, K.: Optimal actuator design based on shape calculus. Math. Models Methods Appl. Sci. **28**(13), 2667–2717 (2018). DOI 10.1142/S0218202518500586
15. Kanzow, C., Mehlitz, P.: Convergence properties of monotone and nonmonotone proximal gradient methods revisited. J. Optim. Theory Appl. **195**(2), 624–646 (2022). DOI 10.1007/s10957-022-02101-3
16. Kanzow, C., Mehlitz, P., Steck, D.: Relaxation schemes for mathematical programmes with switching constraints. Optim. Methods Softw. **36**(6), 1223–1258 (2021). DOI 10.1080/10556788.2019.1663425
17. Manns, P., Hahn, M., Kirches, C., Leyffer, S., Sager, S.: On convergence of binary trust-region steepest descent. J. Nonsmooth Anal. Optim. **4**, Paper No. 10164, 26 (2023). DOI 10.46298/jnsao-2023-10164
18. Mehlitz, P., Wachsmuth, G.: Subdifferentiation of nonconvex sparsity-promoting functionals on Lebesgue spaces. SIAM J. Control Optim. **60**(3), 1819–1839 (2022). DOI 10.1137/21M1435173
19. Natemeyer, C., Wachsmuth, D.: A proximal gradient method for control problems with non-smooth and non-convex control cost. Comput. Optim. Appl. **80**(2), 639–677 (2021). DOI 10.1007/s10589-021-00308-0
20. Natemeyer, C., Wachsmuth, D.: A penalty scheme to solve constrained non-convex optimization problems in $BV(\Omega)$. Pure Appl. Funct. Anal. **7**(5), 1857–1880 (2022)
21. Tröltzsch, F.: Optimal control of partial differential equations, *Graduate Studies in Mathematics*, vol. 112. American Mathematical Society, Providence, RI (2010). DOI 10.1090/gsm/112. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels
22. 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). DOI 10.1137/18M1194602
23. Wachsmuth, D.: Optimal control problems with $l^0(\omega)$constraints: maximum principle and proximal gradient method. Computational Optimization and Applications (2023). DOI 10.1007/s10589-023-00456-5
24. Wachsmuth, D.: Optimal control problems with $L^0(\Omega)$ constraints: maximum principle and proximal gradient method. Comp. Opt. Appl. (2023). DOI 10.1007/s10589-023-00456-5
25. Wachsmuth, D., Wachsmuth, G.: Second-order conditions for non-uniformly convex integrands: quadratic growth in L^1. J. Nonsmooth Anal. Optim. **3**, Paper No. 8733, 36 (2022). DOI 10.2140/pmp.2022.3.35
26. Zhang, H., Hager, W.W.: A nonmonotone line search technique and its application to unconstrained optimization. SIAM J. Optim. **14**(4), 1043–1056 (2004). DOI 10.1137/S1052623403428208