J. Nonsmooth Anal. Optim. 6 (2026), 16354, DOI: 10.46298/jnsao-2026-16354
Submitted: 2025-08-15, accepted: 2026-08-17

© the authors, CC-BY-NC-ND 4.0

A TRUST-REGION METHOD FOR OPTIMAL CONTROL OF ODES WITH CONTINUOUS-OR-OFF CONTROLS AND TV REGULARIZATION

Markus Friedemann*, Gerd Wachsmuth†

Abstract A solution algorithm for a special class of optimal control problems subject to an ordinary differential equation is proposed. The controls possess a continuous-or-off structure and are priced by a convex function. Additionally, a total variation regularization is applied to penalize switches. Our solution method combines a trust-region method and a proximal gradient method. The subproblems are solved via Bellman's optimality principle. Convergence with respect to a criticality measure is proven. As a numerical example, we solve a simple optimal control problem involving an SIR model.

Keywords: integer optimal control problem, total variation regularization, switching-point optimization, proximal gradient method, trust-region method
MSC: 90C11, 49M37

1 INTRODUCTION

We consider the problem

$$\min_{u\in U_{\mathrm{ad}}} J(u) \coloneqq F(u) + \int_{t_0}^{T} g(u(t))\,\mathrm{d}t + \mathrm{TV}(\mathrm{sgn}(u)). \tag{1.1}$$

The objective function of (1.1) contains a differentiable part F that encodes the control-state system. The control $u\colon [t_0, T] \to \{0\} \cup [a, b]$ at each time instance t is either off ($u(t) = 0$) or takes values in the interval $[a, b]$ with $0 < a < b$. The total variation term $\mathrm{TV}(\mathrm{sgn}(u))$ penalizes each switch between these two modes. Finally, the function g acts as a pricing function and is assumed to be strongly convex on $[a, b]$. In Section 4, we study a variant of (1.1) with vector-valued controls. The precise assumptions on the data are given in Section 4.

Our investigation of (1.1) was motivated by the study of control problems for pandemic models. Tracing back to [16], the underlying processes can be approximated by nonlinear ordinary differential equations, see also [8]. The TV term is introduced as a penalization for the number of switches. Problem (1.1) has further applications. One of them is the optimal control of gas networks [12, 13, 3, 23]. Another application is the control of heatpumps, see [15], where effective cycle planning for multiple units can

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German research foundation) via grant number 458468407.

*Brandenburgische Technische Universität Cottbus–Senftenberg, Institute of Mathematics, 03046 Cottbus, Germany, markus.friedemann@b-tu.de, https://www.b-tu.de/fg-optimale-steuerung

†Brandenburgische Technische Universität Cottbus–Senftenberg, Institute of Mathematics, 03046 Cottbus, Germany, gerd.wachsmuth@b-tu.de, https://www.b-tu.de/fg-optimale-steuerung, ORCID: 0000-0002-3098-1503

reduce peak electricity demands while ensuring sufficient heating. It can also be applied to the cycle optimization of a single unit to minimize energy consumption while increasing the durability.

A problem very similar to (1.1) was studied in [7], although with completely different techniques. Up to this contribution, we are not aware of publications related to (1.1).

We propose to solve (1.1) by a combination of a proximal gradient method with a trust-region method. These algorithms are well established for solving nonlinear optimization problems, see [6, 11]. In [14, 17, 19, 20, 25] it was shown that the trust-region method can be applied to solve control problems with a switching structure which, in its simplest form, corresponds to problem (1.1) with $a = b = 1$, i.e., the control can only take the values 0 and 1, and each switch between these values is penalized. An abstract convergence analysis of the trust-region method was done by [18]. This abstract trust-region framework allows for streamlined proofs of classic convergence results.

This strongly hints at the fact, that a trust-region method will be well suited for (1.1), however for any trust-region method an efficient solution method for the arising subproblems is required. For a problem with switching structure, this can be done via combinatorial algorithms, see [21, 24]. In [21], it was shown that Bellman's optimality principle applies for the solution of subproblems in case of one-dimensional controls and it was extended in [22] to vector-valued controls.

Mathematically, this work can be viewed as an application of [18] by adapting the proximal gradient algorithm to this framework and combining it with the switching point optimization. We will show that the trust-region subproblems can be efficiently solved by an adaptation of the mentioned combinatorial algorithms. In this process, we will provide a minor generalization of [18]. In particular, we show that the algorithm already applies without a metric structure. The convergence proof follows the ideas of [18].

For a detailed overview of proximal gradient methods, we refer to the book [6] and the references therein. Recent contributions in particular concerning trust-region globalizations in the context of nonsmooth problems are given in, e.g., [4, 2].

In Section 2 the abstract trust-region method (Algorithm 2.1) is introduced, together with its necessary framework. As it will be shown, the analysis for the actual solution algorithm Algorithm 4.1 can be split into the part relating to a proximal gradient step and a switching point optimization. To this end, it is beneficial to prove the convergence of the proximal gradient method interpreted as a trust-region method using Algorithm 2.1 in Section 3. First, we prove convergence of a criticality measure. In the case that F has a compact Fréchet derivative, strong convergence of the control is proven. The main Section 4 is dedicated to the analysis of (1.1). We start by fixing the setting and showing the existence of solutions in Section 4.1. In Section 4.2, we derive optimality conditions, see Theorem 4.7. These stationarity conditions are used to define a criticality measure in Definition 4.11, which involves a proximal gradient operator on the set $\{\operatorname{sgn}(u) = 1\}$ and terms corresponding to the switching points. In Section 4.3, we apply the trust-region approach from Section 2 to (1.1). To this end, the oracle function T (which yields the suggestion for the next iterate) is given in Definition 4.15. Afterwards, we verify Assumption 2.1. Interestingly, all proofs basically split into a part which addresses the continuous contributions on $\{\operatorname{sgn}(u) = 1\}$ (where we can reuse results from Section 3) and a part which addresses the switching nature of the control (which can be handled by adapting the ideas from [18]). The final convergence result is Theorem 4.24. Section 4 is completed by the presentation of the inner solver in Section 4.4, combining a proximal gradient step with a switching-point solver based on Bellman's optimality principle. We present some numerical results in Section 5.

2 ABSTRACT TRUST-REGION METHOD

This section presents the result from [18, Section 4], such that it serves as an introduction to the abstract trust-region method with slight notational changes. While the problem is stated in the setting of metric

spaces in [18, Section 4], the convergence proof does not rely on this structure. Consequently, we state the method using an abstract set X.

Let X be a set and let $J : X \to \mathbb{R}$ denote a function that is bounded from below. We want to solve the abstract optimization problem

$$\min_{x \in X} J(x). \tag{OP}$$

The basis of our method is a function

$$T : X \times (0, \infty) \to X, \tag{oracle}$$

which is called "oracle", and a "prediction function" (or "predicted reduction")

$$\operatorname{pred} : X \times (0, \infty) \to [0, \infty). \tag{pred}$$

Given an iterate $x \in X$ and a trust-region radius $\Delta > 0$, $T(x, \Delta)$ is a suggestion for the next iterate and $\operatorname{pred}(x, \Delta)$ is a prediction for the decrease in the objective J. Typically, one has a model function $m_{x,\Delta} \colon X \to (-\infty, \infty]$ and $T(x, \Delta)$ is an approximate minimizer or Cauchy point of the model while $\operatorname{pred}(x, \Delta) = m_{x,\Delta}(x) - m_{x,\Delta}(T(x, \Delta))$, see Sections 3 and 4. Note that a model function is not used in the abstract convergence proof. Similarly, it is not enforced that $T(x, \Delta)$ stays within distance Δ to the point x since X is only an abstract set and might not carry a metric.

We define the "actual reduction" via

$$\operatorname{ared} : X \times (0, \infty) \to \mathbb{R}, \qquad \operatorname{ared}(x, \Delta) \coloneqq J(x) - J(T(x, \Delta)). \tag{ared}$$

Further, we are given a "criticality measure"

$$C : X \to [0, \infty). \tag{C}$$

The criticality measure C represents the encoding of an abstract stationarity property, where $C(x) = 0$ if and only if said property holds true.

The following set of assumptions connects the above functions, such that the iterates of a trust-region method satisfy the convergence $\lim_{n\to\infty} C(x_n) = 0$, see Theorem 2.2 below.

Assumption 2.1. Let the optimization problem (OP) be given. Assume that there exists a function $\underline{\Delta}_a : X \to [0, \infty]$, constants $\underline{\Delta}_b, \underline{\Delta}_c > 0$, a nondecreasing function $f : [0, \infty) \to [0, \infty)$, and constants $c \geq 0$, $L, \delta > 0$, $s > 1$, and $\eta \in (0, 1)$, such that for all $x \in X$ with $C(x) > 0$ the following properties hold:

monotonicity:

$$\operatorname{pred}(x, \Delta_1) \geq \operatorname{pred}(x, \Delta_2) \qquad \forall \Delta_1 \geq \Delta_2 > 0 \tag{I}$$

sufficient decrease:

$$\operatorname{pred}(x, \Delta) \geq f(C(x))\Delta - c\Delta^s \qquad \forall \Delta \in (0, \underline{\Delta}_a(x)) \tag{II}$$

accuracy:

$$\operatorname{ared}(x, \Delta) - \eta \operatorname{pred}(x, \Delta) \geq (1 - \eta)\left(f(C(x))\Delta - c\Delta^s\right) \qquad \forall \Delta \in (0, \underline{\Delta}_a(x)) \tag{III}$$

$$\begin{aligned} &\operatorname{ared}(x, \Delta) \geq \eta \operatorname{pred}(x, \Delta) \\ &\operatorname{pred}(x, \Delta) \geq \delta \end{aligned} \qquad \forall \Delta \in [\underline{\Delta}_a(x), \underline{\Delta}_b] \tag{III.E}$$

stability: $\forall \Delta \in (0, \underline{\Delta}_c]$:

$$\text{(IV)} \qquad |C(T(x,\Delta)) - C(x)| \le L\Delta$$
$$\text{(IV.E)} \qquad \text{or} \qquad \operatorname{pred}(x,\Delta) \ge \delta$$

With all necessary assumptions in place, Algorithm 2.1 can be stated. Under our abstract assumptions,

Algorithm 2.1: Abstract trust-region method

Data: $\eta \in (0,1)$, $0 < \gamma_1 < 1 \le \gamma_2$, $0 < \Delta_0 < \Delta_{\max} \in (0,\infty]$, $x_0 \in X$

$n \coloneqq 0$
while $C(x_n) > 0$ **do**
  if $\operatorname{ared}(x_n,\Delta_n) \ge \eta \ \operatorname{pred}(x_n,\Delta_n)$ **then**
    $x_{n+1} \coloneqq T(x_n,\Delta_n)$
    $\Delta_{n+1} \coloneqq \min\{\gamma_2\Delta_n, \Delta_{\max}\}$
  else
    $x_{n+1} \coloneqq x_n$
    $\Delta_{n+1} \coloneqq \gamma_1\Delta_n$
  $n \coloneqq n+1$

we obtain convergence of the criticality measure.

Theorem 2.2. *Let the optimization problem* (OP) *be given. If Assumption 2.1 is valid, then Algorithm 2.1 either terminates at an iterate x_n with $C(x_n) = 0$ or produces an infinite sequence of iterates (x_n), such that $\lim_{n\to\infty} C(x_n) = 0$.*

The proof of the theorem follows the ideas of [18, Section 4] and can be found in Appendix A.

Remark 2.3. In the context of Algorithm 2.1, an iteration n is called "accepted" or "successful" if $\operatorname{ared}(x_n,\Delta_n) \ge \eta \operatorname{pred}(x_n,\Delta_n)$ and subsequently $x_{n+1} = T(x_n,\Delta_n)$. Otherwise iteration n is called "unsuccessful".

Regarding Assumption 2.1(II) and (III), it is important to point out that a negative term on the right hand side is allowed, and that this term is of higher order than Δ. Whether there are finitely many of such terms with different exponents $s_j > 1$ ultimately is of no importance. In the worst case one can always limit the convergence proof to estimate every negative term using $\Delta^{s_i} \le \Delta^{\min_i s_i}$ for $\Delta \in [0,1]$.

As it becomes apparent from the proof, the trust-region radius is required to be scaled with $\gamma_1 < 1$ upon an unsuccessful iteration and is nondecreasing in case of a successful one, i.e., $\Delta_{n+1} \in [\Delta_n, \min\{\gamma_2\Delta_n, \Delta_{\max}\}]$ could be used. This corresponds to the classic trust-region method where the decision on increasing the trust-region radius Δ is made on the basis of how well prediction and actual reduction match.

3 THE PROXIMAL GRADIENT ALGORITHM AS A TRUST-REGION METHOD

The main work in the convergence proof for Algorithm 2.1 applied to the problem (OC) is the verification of Assumption 2.1 for a composite problem. Roughly speaking, the oracle function for (OC) includes a proximal gradient step on the set where the control is already active. Consequently, the aim of the current section is twofold. First, it provides a self-contained convergence proof for a proximal gradient algorithm by interpreting it as an abstract trust-region method. Second, the estimates of this section

will be used in Section 4 to analyze the trust-region method for (OC). For this section, we will be investigating problems of class

$$(OC_{\text{conv}}) \qquad \min_{v\in U} J(v) \coloneqq F(v) + G(v).$$

Here, U is a Hilbert space and $F\colon U \to \mathbb{R}$, $G\colon U \to (-\infty, \infty]$ are given functions. Throughout this section, we require the following standing assumption.

Assumption 3.1 (Standing assumption for Section 3). The function G is assumed to be convex, proper and lower semicontinuous and F is assumed to be Fréchet differentiable with Lipschitz continuous derivative with constant $L_{\nabla F}$. Further, we require that F and G are Lipschitz continuous on $\operatorname{dom}(G) \coloneqq \{v \in U \mid G(v) < \infty\}$ with constants L_F, L_G. Finally, $F + G$ is assumed to be bounded from below.

The final assumption on the Lipschitz continuity of F and G on the domain of G is quite strong and actually not needed for the convergence of a proximal gradient method. However, it is needed for the straightforward application of the abstract trust-region method from the previous section, see, in particular, Lemma 3.6 below.

In Section 3.1 it will be discussed briefly, why the function

$$(TRM_{\text{conv}}) \qquad m_{v,\Delta}(w) \coloneqq \langle \nabla F(v), w - v\rangle_U + G(w) - G(v) + \frac{1}{2\Delta}\|w - v\|_U^2,$$

whose minimizer is the proximal gradient step, is also a reasonable choice as a model function for Algorithm 2.1. Herein, v is the current iterate, the trust-region radius $\Delta > 0$ serves as a regularization parameter and the minimizer of the model function will act as the oracle, i.e., the proposal for the next iterate. Convergence with respect to the criticality measure will be proven in Section 3.2, see Theorem 3.13. In Theorem 3.15 it will be shown how the inclusion of a regularization term leads to strong convergence of the iterates under the additional assumption of ∇F being a compact operator.

3.1 PRELIMINARIES

Definition 3.2. Given $r > 0$, the proximal point $\hat{v} \coloneqq \operatorname{prox}_{rG}(v)$ of $v \in U$ w.r.t. rG is defined as the unique minimizer $\bar{v}$ of the strongly convex optimization problem

$$\min_{\hat{v}\in U}\left(G(\hat{v}) + \frac{1}{2r}\|\hat{v} - v\|_U^2\right).$$

It is well known that the proximal point $\hat{v} = \operatorname{prox}_{rG}(v)$ can be characterized via the optimality condition $v - \hat{v} \in r\partial G(\hat{v})$. From the literature we recall two crucial properties that will be used in the proofs.

Corollary 3.3. *For all $r > 0$ the proximal point mapping is nonexpansive, i.e.,*

$$\|\operatorname{prox}_{rG}(v) - \operatorname{prox}_{rG}(w)\|_U \le \|v - w\|_U \qquad \forall v, w \in U. \tag{3.1}$$

Further, given $v, d \in U$ and defining $\Phi\colon (0,\infty) \to [0,\infty)$ via

$$\Phi(r) \coloneqq \|\operatorname{prox}_{rG}(v + rd) - v\|_U, \tag{3.2}$$

the mapping $r \mapsto \Phi(r)$ is nondecreasing, while $r \mapsto \frac{\Phi(r)}{r}$ is nonincreasing.

Proof. (3.1) is proven in [5, Corollary 23.11(i)]. The monotonicity properties related to Φ are proven in the finite-dimensional setting in [6, Theorem 10.9], and the proof directly carries over to the Hilbert space setting. □

It is well known that the first-order stationarity conditions at a point $\bar{v} \in U$ for (OC_{conv}) are given by $-\nabla F(\bar{v}) \in \partial G(\bar{v})$, due to the Fréchet differentiability of F. The next, classical result shows that this can be expressed via a criticality measure involving the proximal map of G.

Corollary 3.4. *For fixed $r > 0$, we introduce the criticality measure $C\colon U \to [0,\infty)$ via*

$$C(v) \coloneqq \frac{1}{2r^2}\|v - \operatorname{prox}_{rG}(v - r\nabla F(v))\|_U^2. \tag{3.3}$$

Then, a point $\bar{v} \in U$ is stationary if and only if $C(\bar{v}) = 0$.

Proof. We write the first-order stationarity condition as $(\bar{v} - r\nabla F(\bar{v})) - \bar{v} \in \partial(rG)(\bar{v})$. Using the characterization of the proximal point in Definition 3.2, this is equivalent to $\bar{v} = \operatorname{prox}_{rG}(\bar{v} - r\nabla F(\bar{v}))$, proving the claim. □

Since the (possibly) nonconvex function F is linearized in (TRM_{conv}), the problem (TRM_{conv}) is strongly convex and computing the proximal point mapping directly yields the unique minimizer.

Definition 3.5. The oracle function $T : U \times (0,\infty) \to U$ for (OC_{conv}) is defined as

$$T(v,\Delta) \coloneqq \operatorname{prox}_{\Delta G}(v - \Delta\nabla F(v)).$$

Since Δ is an established symbol for the trust-region radius, the proximal gradient step used for the oracle function reads slightly unconventional, but will not lead to any ambiguities later. In a classical trust-region method, a local model is optimized over the trust region, i.e., the distance between the current iterate and the proposed next iterate is limited by the trust-region radius Δ. In the proximal gradient method, the proximity is only weakly enforced by the quadratic term in the model. Nonetheless, using the Lipschitz continuity from Assumption 3.1 we can prove that iterates remain in a scaled trust region.

Lemma 3.6. *There exists a $\rho > 0$, such that the oracle from Definition 3.5 satisfies*

$$\|T(v,\Delta) - v\|_U \le \rho\Delta \qquad \forall v \in \operatorname{dom}(G), \Delta \in (0,\infty).$$

In particular, we can choose $\rho = 2(L_F + L_G) > 0$, where L_F, L_G denote the Lipschitz constants of F and G, both on the domain of G.

Proof. Since $\hat{v} \coloneqq T(v,\Delta)$ is the minimizer of the model function $m_{v,\Delta}$ from (TRM_{conv}), we obtain

$$0 = m_{v,\Delta}(v) \ge m_{v,\Delta}(\hat{v}) = \langle \nabla F(v), \hat{v} - v\rangle_U + G(\hat{v}) - G(v) + \frac{1}{2\Delta}\|\hat{v} - v\|_U^2.$$

Rearranging the terms and using Assumption 3.1, we find

$$\frac{1}{2\Delta}\|\hat{v} - v\|_U^2 \le \langle \nabla F(v), v - \hat{v}\rangle_U + G(v) - G(\hat{v}) \le (L_F + L_G)\|\hat{v} - v\|_U. \tag{3.4}$$

Here, we used

$$\langle \nabla F(v), v - \hat{v}\rangle_U = -\lim_{t \searrow 0} \frac{F(v + t(\hat{v} - v)) - F(v)}{t} \le L_F\|\hat{v} - v\|_U.$$

This yields the claim. □

We briefly mention that the assumed Lipschitz continuity of G is necessary for the claim of Lemma 3.6. As a simple example, we could consider $U = \mathbb{R}$, $F \equiv 0$ and

$$G(x) \coloneqq \begin{cases} -\sqrt{x} & \text{if } x \geq 0, \\ \infty & \text{else.} \end{cases}$$

Then,

$$\begin{aligned} T(0,\Delta) &= \operatorname*{arg\,min}_{\hat{v}\in\mathbb{R}} \left(0 \cdot (\hat{v} - 0) + G(\hat{v}) - G(0) + \frac{1}{2\Delta}(\hat{v} - 0)^2 \right) \\ &= \operatorname*{arg\,min}_{\hat{v}\in[0,\infty)} \left(-\sqrt{\hat{v}} + \frac{1}{2\Delta}\hat{v}^2 \right) = \left(\frac{\Delta}{2} \right)^{\frac{2}{3}}. \end{aligned}$$

This shows that for any value of $\rho > 0$, the assertion of Lemma 3.6 is violated for $\Delta < \frac{1}{4\rho^3}$. As the issue occurs for all trust-region radii below a certain threshold, it can not be excluded in the algorithm.

3.2 CONVERGENCE

In this section the individual parts of Assumption 2.1 are verified in separate lemmata. The convergence of the critical measure will be concluded from Theorem 2.2 in Theorem 3.13. Strong convergence for a subsequence of iterates is proven under additional assumptions in Theorem 3.15.

In view of Lemma 3.6, the space X appearing in the abstract setting of Section 2 is defined as

$$X \coloneqq \operatorname{dom}(G).$$

Note that the oracle T maps into $X = \operatorname{dom}(G)$ by definition of the proximal mapping.

Definition 3.7 (Prediction function). We define the prediction function $\operatorname{pred} : U \times (0,\infty) \to [0,\infty)$ using the model function (TRM_{conv}) via

$$\operatorname{pred}(v,\Delta) \coloneqq m_{v,\Delta}(v) - m_{v,\Delta}(T(v,\Delta)) = -m_{v,\Delta}(T(v,\Delta)).$$

Lemma 3.8 (Monotonicity). *The prediction function* $\operatorname{pred}(v,\Delta)$ *is nondecreasing with respect to* Δ.

Proof. As Δ increases, the model function $m_{v,\Delta}$ decreases. This yields the claim. □

Lemma 3.9 (Sufficient decrease). *For the value* $r > 0$ *that was used in the definition* (3.3) *of the criticality measure, the estimate*

$$\operatorname{pred}(v,\Delta) = -m_{v,\Delta}(T(v,\Delta)) \geq C(v)\Delta \qquad \forall v \in U, \Delta \in (0,r]$$

holds and therefore Assumption 2.1(II) *is valid with* $\underline{\Delta}_a \equiv r$ *and* $c = 0$.

Proof. For brevity, we set $\hat{v} \coloneqq T(v,\Delta)$. Due to the convexity of G, the model function $m_{v,\Delta}$ is strongly convex with constant Δ^{-1}. Consequently, the quadratic growth condition

$$m_{v,\Delta}(w) \geq m_{v,\Delta}(\hat{v}) + \frac{1}{2\Delta}\|w - \hat{v}\|_U^2 \qquad \forall w \in U$$

holds at its minimizer $\hat{v} \coloneqq T(v,\Delta) = \operatorname{prox}_{\Delta G}(v - \Delta\nabla F(v))$. In particular, inserting $w = v$ implies

$$\operatorname{pred}(v,\Delta) = m_{v,\Delta}(v) - m_{v,\Delta}(\hat{v}) \geq \frac{1}{2\Delta}\|v - \hat{v}\|_U^2.$$

It remains to relate the last term with the criticality measure (3.3). To this end, we employ the function $r \mapsto \Phi(r) = \|v - \operatorname{prox}_{rG}(v - r\nabla F(v))\|$ from Corollary 3.3 and the fact that $r \mapsto \frac{\Phi(r)}{r}$ is nonincreasing. We get

$$\begin{aligned}\frac{1}{2\Delta}\|v - \hat{v}\|_U^2 = \frac{1}{2\Delta}\|v - \operatorname{prox}_{\Delta G}(v - \Delta\nabla F(v))\|_U^2 = \frac{1}{2\Delta}\Phi(\Delta)^2 = \frac{\Delta}{2}\left(\frac{\Phi(\Delta)}{\Delta}\right)^2 \\ \geq \frac{\Delta}{2}\left(\frac{\Phi(r)}{r}\right)^2 = \frac{\Delta}{2r^2}\|\operatorname{prox}_{rG}(v - r\nabla F(v)) - v\|_U^2 = C(v)\Delta,\end{aligned}$$

closing the proof. □

Lemma 3.10 (Accuracy). *For an arbitrary $\eta \in (0, 1]$, we have*

$$\operatorname{ared}(v, \Delta) - \eta \operatorname{pred}(v, \Delta) \geq (1 - \eta)C(v)\Delta - \frac{1}{2}L_{\nabla F}\rho^2\Delta^2 \qquad \forall v \in \operatorname{dom}(G), \Delta \in (0, r],$$

where $r > 0$ is from (3.3) and $L_{\nabla F}$ denotes the Lipschitz constant of ∇F. Thus, Assumption 2.1(III) holds for $\underline{\Delta}_a \equiv r$, $s = 2$, and $c = \frac{1}{2}L_{\nabla F}\rho^2$.

Proof. We define $\hat{v} \coloneqq T(v, \Delta)$. Since ∇F is assumed to be Lipschitz continuous with constant $L_{\nabla F}$, the fundamental theorem of calculus implies

$$F(\hat{v}) \leq F(v) + \langle \nabla F(v), \hat{v} - v\rangle_U + \frac{L_{\nabla F}}{2}\|\hat{v} - v\|_U^2.$$

Consequently,

$$\begin{aligned}\operatorname{ared}(v, \Delta) &= J(v) - J(\hat{v}) = F(v) - F(\hat{v}) + G(v) - G(\hat{v}) \\ &\geq \langle \nabla F(v), v - \hat{v}\rangle_U + G(v) - G(\hat{v}) - \frac{L_{\nabla F}}{2}\|\hat{v} - v\|_U^2 \\ &= \operatorname{pred}(v, \Delta) + \frac{1}{2}\left(\frac{1}{\Delta} - L_{\nabla F}\right)\|\hat{v} - v\|_U^2 \\ &\geq \eta \operatorname{pred}(v, \Delta) + (1 - \eta)C(v)\Delta - \frac{1}{2}L_{\nabla F}\rho^2\Delta^2,\end{aligned}$$

where we used Lemmas 3.6 and 3.9 in the final estimate. □

Remark 3.11. The estimate in the proof of Lemma 3.10 even yields

$$\Delta \leq L_{\nabla F}^{-1} \quad \Longrightarrow \quad \frac{1}{\Delta} - L_{\nabla F} \geq 0 \quad \Longrightarrow \quad \operatorname{ared}(v, \Delta) - \eta \operatorname{pred}(v, \Delta) \geq 0.$$

This means that iterations with a sufficiently small trust-region radius are always successful. Such a result is expected under the given assumptions, cf. [6, Lemma 10.4].

Lemma 3.12 (Stability). *The criticality measure $C(v)$ from (3.3) satisfies*

$$|C(T(v, \Delta)) - C(v)| \leq L_C\Delta \qquad \forall v \in \operatorname{dom}(G), \Delta \in (0, \infty),$$

where $L_C \coloneqq \frac{\rho^2}{r}(2 + rL_{\nabla F})$ with $L_{\nabla F}$ being the Lipschitz constant of ∇F on $\operatorname{dom}(G)$ and ρ as defined in Lemma 3.6. Hence, Assumption 2.1(IV) holds for arbitrary $\underline{\Delta}_c > 0$.

Proof. Again, we set $\hat{v} \coloneqq T(v, \Delta)$. We first recall that Lemma 3.6 implies

$$\|v - \operatorname{prox}_{rG}(v - r\nabla F(v))\|_U = \|v - T(v, r)\|_U \leq \rho r,$$

and similarly for v replaced by $\hat{v}$. This lemma also yields $\|v - \hat{v}\|_U \leq \rho\Delta$. Further, we utilize the nonexpansivness of the proximal mapping as stated in Corollary 3.3 to estimate

$$\begin{aligned}
2r^2|C(\hat{v}) - C(v)| &= \left|\|\hat{v} - \operatorname{prox}_{rG}(\hat{v} - r\nabla F(\hat{v}))\|_U^2 - \|v - \operatorname{prox}_{rG}(v - r\nabla F(v))\|_U^2\right| \\
&= \left(\|\hat{v} - \operatorname{prox}_{rG}(\hat{v} - r\nabla F(\hat{v}))\|_U + \|v - \operatorname{prox}_{rG}(v - r\nabla F(v))\|_U\right) \\
&\qquad \cdot \left|\|\hat{v} - \operatorname{prox}_{rG}(\hat{v} - r\nabla F(\hat{v}))\|_U - \|v - \operatorname{prox}_{rG}(v - r\nabla F(v))\|_U\right| \\
&\leq 2\rho r\|v - \hat{v} + \left(\operatorname{prox}_{rG}(\hat{v} - r\nabla F(\hat{v})) - \operatorname{prox}_{rG}(v - r\nabla F(v))\right)\|_U \\
&\leq 2\rho r\left(\rho\Delta + \|\operatorname{prox}_{rG}(\hat{v} - r\nabla F(\hat{v})) - \operatorname{prox}_{rG}(v - r\nabla F(v))\|_U\right) \\
&\leq 2\rho r\left(\rho\Delta + \|\hat{v} - r\nabla F(\hat{v}) - (v - r\nabla F(v))\right)\|_U \\
&\leq 2\rho r\left(2\rho\Delta + r\|\nabla F(\hat{v}) - \nabla F(v)\|_U\right) \leq 2\rho^2 r(2 + rL_{\nabla F})\Delta.
\end{aligned}$$

□

Theorem 3.13. *The abstract trust-region method Algorithm 2.1 applied to* (OC) *with the proximal gradient step as oracle* $T(v, \Delta)$ *and the prediction function* $\operatorname{pred}(v, \Delta)$ *from Definition 3.7 either terminates at a stationary point* $\bar{v}$ *or produces an infinite sequence of iterates* (v_n)*, such that*

$$\left\|v_n - \operatorname{prox}_{\hat{r}G}\left(v_n - \hat{r}\nabla F(v_n)\right)\right\|_U \to 0$$

for all $\hat{r} > 0$.

Proof. We will apply Theorem 2.2 with $X = \operatorname{dom}(G)$. Monotonicity was proven in Lemma 3.8, predictability in Lemma 3.9, accuracy in Lemma 3.10 and stability in Lemma 3.12. This verifies Assumption 2.1 with $\underline{\Delta}_a(v) = \underline{\Delta}_c = r$, $\underline{\Delta}_b = r/2 < r$. Note that none of the exception clauses were needed, i.e., (III.E) is void due to $\underline{\Delta}_b < \underline{\Delta}_a$ and (IV.E) is not needed since (IV) always holds. Therefore, convergence of the criticality measure C for the iterates of the trust-region algorithm Algorithm 2.1 follows from Theorem 2.2. This shows the convergence for $\hat{r} = r$. In particular, the function $\Phi_n(\hat{r}) \coloneqq \|u_n - \operatorname{prox}_{\hat{r}G}(u_n - \hat{r}\nabla F(u_n))\|_U$ satisfies $\Phi_n(r) \to 0$. Using the monotonicity properties from Corollary 3.3, this implies

$$\Phi_n(\hat{r}) \leq \Phi_n(r) \to 0 \quad \forall \hat{r} < r \qquad \text{and} \qquad \frac{\Phi_n(\hat{r})}{\hat{r}} \leq \frac{\Phi_n(r)}{r} \to 0 \quad \forall \hat{r} > r.$$

This finishes the proof. □

Remark 3.14. In Remark A.3 it will be argued that convergence with respect to the criticality measure already follows from Assumption 2.1 (I)–(III) if a global lower bound of the trust-region radius is known. With Remark 3.11 convergence of Algorithm 2.1 for (OC) with the proximal gradient oracle already follows from Lemma 3.8, Lemma 3.9 and Lemma 3.10. Notably, global Lipschitz continuity of F and G is not needed. Regarding regularity of F, we still need Lipschitz continuity of ∇F. This recovers the standard convergence result for the proximal gradient method, cf. [6, Theorem 10.15(b)].

The result in this form does not yet yield convergence of the iterates. This can be shown under additional assumptions on the function F.

Theorem 3.15. *We assume that* $G\colon U \to (-\infty, \infty]$ *is convex, proper and lower semicontinuous. Let* $\hat{F}\colon U \to \mathbb{R}$ *be Fréchet differentiable such that* $\nabla\hat{F}\colon U \to U$ *is compact and Lipschitz continuous. We further assume that* $\hat{F}$ *and* G *are Lipschitz continuous on* $\operatorname{dom}(G)$*. We define* F *via*

$$F(v) \coloneqq \hat{F}(v) + \frac{\sigma}{2}\|v\|_U^2 \qquad \forall v \in U,$$

where $\sigma > 0$. Finally, we assume that $F + G$ is bounded from below. Then, the sequence of iterates contains a strongly convergent subsequence and every accumulation point is stationary.

Proof. We have to check that Assumption 3.1 is satisfied for F and G. The only problem is that $v \mapsto \|v\|_U^2$ is only Lipschitz continuous on bounded sets. For this, we use that $F(v_n) + G(v_n)$ is decreasing, in particular,

$$\|v_n\|_U \leq \sqrt{2 \cdot \frac{F(v_n) + G(v_n) - C}{\sigma}} \leq \sqrt{2 \cdot \frac{F(v_0) + G(v_0) - C}{\sigma}} =: R,$$

where C is a lower bound for $F + G$. This means that we can redefine G to take the value $+\infty$ outside the ball of radius R without changing the iterates of Algorithm 2.1. Consequently, F becomes Lipschitz continuous on $\operatorname{dom}(G)$.

From the boundedness of the iterates (v_n) we get weak convergence of a subsequence. However, it is not clear that the weak limit $\bar{v}$ satisfies $C(\bar{v}) = 0$. We will first prove that every weakly convergent subsequence is even strongly convergent and conclude that the limit point $\bar{v}$ is stationary.

Consider (without relabeling) a weakly convergent subsequence $(v_n) \rightharpoonup \bar{v}$. As shorthand we introduce $\hat{v}_n$ via a proximal gradient step with step length $\frac{1}{\sigma}$, i.e.,

$$\hat{v}_n := \operatorname{prox}_{\frac{G}{\sigma}}\left(v_n - \frac{\nabla F(v_n)}{\sigma}\right).$$

From Theorem 3.13, we get $\|\hat{v}_n - v_n\| \to 0$. This implies $\hat{v}_n \rightharpoonup \bar{v}$. The proximal point mapping is characterized by the optimality condition for the corresponding optimization problem, i.e.,

$$v_n - \hat{v}_n - \frac{1}{\sigma}\nabla F(v_n) \in \frac{1}{\sigma}\partial G(\hat{v}_n).$$

Using the definition of F, this yields

$$-\frac{1}{\sigma}\nabla \hat{F}(v_n) \in \partial\left(\frac{1}{\sigma}G + \frac{1}{2}\|\cdot\|^2\right)(\hat{v}_n).$$

Due to the compactness of $\nabla \hat{F}$, the left-hand side converges to $-\frac{1}{\sigma}\nabla \hat{F}(\bar{v})$. Since the function $\frac{1}{\sigma}G + \frac{1}{2}\|\cdot\|^2$ is strongly convex, its subdifferential has a Lipschitz continuous inverse, see [5, Theorem 18.15 (vii)⇒(i)]. This implies the strong convergence $\hat{v}_n \to \bar{v}$ and, consequently, $v_n \to \bar{v}$. By passing to the limit, we also get

$$-\frac{1}{\sigma}\nabla F(\bar{v}) = -\bar{v} - \frac{1}{\sigma}\nabla \hat{F}(\bar{v}) \in \frac{1}{\sigma}\partial G(\bar{v}).$$

This shows that $\bar{v}$ is a stationary point. □

4 CONTROL PROBLEMS WITH CONTINUOUS-OR-OFF CONTROLS AND TV REGULARIZATION

In this section, we will apply the trust-region algorithm to the problem

$$\min_{u \in U_{\text{ad}}} J(u) := F(u) + \sum_{i=1}^{N}\left(\int_{t_0}^{T} g_i(u_i(t))\,\mathrm{d}t + \operatorname{TV}(\operatorname{sgn}(u_i))\right). \tag{OC}$$

The notation and assumptions are introduced in Section 4.1 and we also prove existence of solutions therein. In Section 4.2, we derive stationarity conditions and define the criticality measure. The application of Section 2 is done in Section 4.3. Finally, we discuss the fast solution of the trust-region subproblems in Section 4.4.

4.1 PROBLEM SETTING, ASSUMPTIONS AND EXISTENCE OF SOLUTIONS

In problem (OC), the control variable is vector-valued and non-negative, i.e., $u\colon [t_0,T]\to[0,\infty)^N$. By $L^p(t_0,T;\mathbb{R}^N)$ we denote the usual Lebesgue space of measurable functions $u\colon [t_0,T]\to\mathbb{R}^N$ satisfying

$$\|u\|^p_{L^p(t_0,T;\mathbb{R}^N)} \coloneqq \int_{t_0}^T \|u(t)\|_p^p \,\mathrm{d}t = \int_{t_0}^T \sum_{i=1}^N |u_i(t)|^p \,\mathrm{d}t < \infty,$$

with the usual modification for $p=\infty$.

One of the main features of (OC) is the total variation term which acts on $\operatorname{sgn}(u_i)$, which is defined pointwise in time, i.e., $\operatorname{sgn}(u_i)(t)=1$ if $u_i(t)>0$ and $\operatorname{sgn}(u_i)(t)=0$ if $u_i(t)=0$. We extend a function $x\in \mathrm{BV}(t_0,T;\{0,1\})$ by 0 to the interval $[t_0-\varepsilon,T+\varepsilon]$ (for some arbitrary $\varepsilon>0$) and define

$$\mathrm{TV}(x) \coloneqq \sup\left\{ \int_{t_0-\varepsilon}^{T+\varepsilon} x(t)\varphi'(t)\,\mathrm{d}t \;\middle|\; \varphi\in C_c^1(t_0-\varepsilon,T+\varepsilon;[-1,1]) \right\}.$$

In contrast to the usual definition, our definition of $\mathrm{TV}(x)$ also includes possible jumps at t_0 and T. If we consider (OC), the term $\mathrm{TV}(\operatorname{sgn}(u_i))$ counts how often the control u_i is switched on or off, including the boundary points t_0 and T. Note that $\mathrm{TV}(\operatorname{sgn}(u_i))$ is always an even number. In the vector-valued case $x\in \mathrm{BV}(t_0,T;\{0,1\}^N)$, we set $\mathrm{TV}(x)\coloneqq\sum_{i=1}^N \mathrm{TV}(x_i)$.

The admissible set is given by

$$U_{\mathrm{ad}} \coloneqq \left\{ u\in L^\infty(t_0,T;\mathbb{R}^N) \;\middle|\; \begin{array}{l} u_i(t)\in\{0\}\cup[a_i,b_i] \text{ for a.a. } t\in[t_0,T] \text{ and} \\ \mathrm{TV}(\operatorname{sgn}(u_i))<\infty \text{ for all } i=1,\dots,N \end{array} \right\}, \tag{4.1}$$

where $0<a_i\le b_i$. For convenience, we also define the sets

$$\bar{M} \coloneqq \bigtimes_{i=1}^N [a_i,b_i], \qquad \bar{M}_0 \coloneqq \bigtimes_{i=1}^N (\{0\}\cup[a_i,b_i]),$$

which encode the possible range of the controls.

The assumptions on the objective terms F and g_i are collected next.

Assumption 4.1.

(i) We assume that $F\colon L^1(t_0,T;\mathbb{R}^N)\to\mathbb{R}$ is bounded from below and Fréchet differentiable such that the Fréchet derivative $\nabla F\colon L^1(t_0,T;\mathbb{R}^N)\to L^\infty(t_0,T;\mathbb{R}^N)$ is Lipschitz continuous with constant $L_{\nabla F}<\infty$. Further, we assume that for all $u\in L^1(t_0,T;\mathbb{R}^N)$, the function $\nabla F(u)\colon [t_0,T]\to\mathbb{R}^N$ is Lipschitz continuous with constant $L_{\nabla F(u)}<\infty$ and that these constants are uniformly bounded. For simplicity of the presentation, we denote the bound again by $L_{\nabla F}$, i.e., we assume $L_{\nabla F(u)}\le L_{\nabla F}$ for all $u\in L^1(t_0,T;\mathbb{R}^N)$.

(ii) The function $g_i:\mathbb{R}\to(-\infty,\infty]$ has the value $g_i(0)=0$, while $g|_{[a_i,b_i]}$ is assumed to be strongly convex with constant $m>0$ and Lipschitz continuous with constant L_g on $[a_i,b_i]$. Finally, $g_i|_{\mathbb{R}\setminus(\{0\}\cup[a_i,b_i])}=\infty$.

(iii) The control bounds satisfy $0<a\le a_i\le b_i\le b<\infty$ for constants a and b

Remark 4.2. Clearly, $\|u\|_{L^1(t_0,T;\mathbb{R}^n)}\le b(T-t_0)N$ for all $u\in U_{\mathrm{ad}}$. The assumed Lipschitz continuity of ∇F with respect to the control directly implies

$$\begin{aligned} \|\nabla F(u)\|_{L^\infty(t_0,T;\mathbb{R}^N)} &\le \|\nabla F(u)-\nabla F(0)\|_{L^\infty(t_0,T;\mathbb{R}^N)} + \|\nabla F(0)\|_{L^\infty(t_0,T;\mathbb{R}^N)} \\ &\le C_{\nabla F} \coloneqq L_{\nabla F} b(T-t_0)N + \|\nabla F(0)\|_{L^\infty(t_0,T;\mathbb{R}^N)} < \infty \end{aligned} \tag{4.2}$$

for all $u\in U_{\mathrm{ad}}$. Subsequently the first-order Taylor expansion with integral remainder yields

$$|F(u)-F(\hat{u})| \le C_{\nabla F}\|u-\hat{u}\|_{L^1(t_0,T;\mathbb{R}^N)} \qquad \forall u,\hat{u}\in U_{\mathrm{ad}}. \tag{4.3}$$

Note that the assumption $g_i(0) = 0$ is not restrictive and can be achieved by a simple shift. For brevity, we define the functional $G\colon L^2(t_0, T; \mathbb{R}^N) \to (-\infty, \infty]$ via

$$G(u) \coloneqq \sum_{i=1}^{N} \int_{t_0}^{T} g_i(u_i(t)) \,\mathrm{d}t.$$

In particular, we have that $G(u) < \infty$ implies $u \in \bar{M}_0$ a.e. on $[t_0, T]$. Note that with our notation, the objective in (OC) can now be written as

$$F(u) + G(u) + \mathrm{TV}(\mathrm{sgn}(u)).$$

Finally, we need variants of g_i and G, which restrict their arguments to be non-zero throughout, i.e., we define $g_{i,+}\colon \mathbb{R} \to (-\infty, \infty]$ via

$$g_{i,+}(x) \coloneqq \begin{cases} g_i(x) & \text{if } x \in [a_i, b_i], \\ \infty & \text{else} \end{cases}$$

and $G_+\colon L^2(t_0, T; \mathbb{R}^N) \to (-\infty, \infty]$ via

$$G_+(u) \coloneqq \sum_{i=1}^{N} \int_{t_0}^{T} g_{i,+}(u_i(t)) \,\mathrm{d}t = \begin{cases} G(u) & \text{if } u(t) \in \bar{M} \text{ for a.a. } t \in (t_0, T), \\ \infty & \text{else.} \end{cases}$$

Note that g_i and G are nonconvex while $g_{i,+}$ and G_+ are convex functions, due to the convexity of g_i on $[a_i, b_i]$.

In order to prove the existence of solutions, we provide a lemma.

Lemma 4.3. *Let $(u_k) \subset U_{\mathrm{ad}}$ be a sequence of admissible controls such that $u_k \rightharpoonup \bar{u}$ in $L^2(t_0, T; \mathbb{R}^N)$ and $\mathrm{sgn}(u_k) \overset{\star}{\rightharpoonup} \bar{\alpha}$ in $BV(t_0, T; \mathbb{R}^N)$, where weak-$\star$ convergence in $BV(t_0, T; \mathbb{R}^N)$ is defined as in [1, Definition 3.11]. Then, $\mathrm{sgn}(\bar{u}) = \bar{\alpha}$ and $\bar{u} \in U_{\mathrm{ad}}$.*

Proof. For simplicity of notation, we only argue in the scalar case $N = 1$.

From [1, Corollary 3.49], we get $\mathrm{sgn}(u_k) \to \bar{\alpha}$ in $L^1(t_0, T)$. We choose a subsequence (without relabeling) such that $\|\mathrm{sgn}(u_k) - \bar{\alpha}\|_{L^1(t_0,T)} \le 2^{-k}$. Since $\mathrm{sgn}(u_k)$ and $\bar{\alpha}$ have values in $\{0, 1\}$ almost everywhere, the Lebesgue measure of the set on which they differ equals the $L^1(t_0, T)$ norm of their difference, i.e.,

$$|\{\mathrm{sgn}(u_k) \neq \bar{\alpha}\}| = \|\mathrm{sgn}(u_k) - \bar{\alpha}\|_{L^1(t_0,T)} \le 2^{-k}.$$

On the set $\bigcap_{l=k}^{\infty} \{\mathrm{sgn}(u_l) = 1\}$ we have $b \ge u_l \ge a > 0$ for all $l \ge k$ and, consequently, $b \ge \bar{u} \ge a > 0$ a.e. This shows $\mathrm{sgn}(\bar{u}) = 1 = \bar{\alpha}$ and $\bar{u} \in [a, b]$ on this set. Similarly, one can argue on the set $\bigcap_{l=k}^{\infty} \{\mathrm{sgn}(u_l) = 0\}$ and one obtains $\mathrm{sgn}(\bar{u}) = 0 = \bar{\alpha}$ and $\bar{u} = 0$ a.e. on this set. Thus, we get $\mathrm{sgn}(\bar{u}) = \bar{\alpha}$ and $\bar{u} \in \{0\} \cup [a, b]$ a.e. on the set $\bigcap_{l=k}^{\infty} \{\mathrm{sgn}(u_l) = \bar{\alpha}\}$ for all $k \in \mathbb{N}$. For the Lebesgue measure of the complement, we get

$$\left| (t_0, T) \setminus \bigcap_{l=k}^{\infty} \{\mathrm{sgn}(u_l) = \bar{\alpha}\} \right| \le \sum_{l=k}^{\infty} |\{\mathrm{sgn}(u_l) \neq \bar{\alpha}\}| \le 2^{-k+1}.$$

This shows $\mathrm{sgn}(\bar{u}) = \bar{\alpha}$ and $\bar{u} \in \{0\} \cup [a, b]$ a.e. on (t_0, T). In order to get $\bar{u} \in U_{\mathrm{ad}}$, we additionally need $\mathrm{TV}(\mathrm{sgn}(\bar{u})) < \infty$, but this follows from $\mathrm{TV}(\mathrm{sgn}(\bar{u})) = \mathrm{TV}(\bar{\alpha}) \le \liminf_{k\to\infty} \mathrm{TV}(\mathrm{sgn}(u_k))$, see [1, Remark 3.5]. □

Theorem 4.4. *We assume that F is sequentially weakly lower semicontinuous on $L^2(t_0,T;\mathbb{R}^N)$. Then, problem (OC) has a solution.*

Proof. Apply the direct method of calculus of variations. A minimizing sequence (u_k) is bounded in $L^2(t_0,T;\mathbb{R}^N)$, since U_{ad} is bounded in this space. Further, $(\operatorname{sgn}(u_k))$ is bounded in $\mathrm{BV}(t_0,T;\{0,1\}^N)$ due to the structure of the objective, since F and G are bounded from below. Consequently, we can extract a subsequence (without relabeling) such that $u_k \rightharpoonup \bar{u}$ in $L^2(t_0,T;\mathbb{R}^N)$ and $\operatorname{sgn}(u_k) \overset{\star}{\rightharpoonup} \bar{\alpha}$ in $\mathrm{BV}(t_0,T;\mathbb{R}^N)$, see [1, Theorem 3.23]. From Lemma 4.3, we get $\operatorname{sgn}(\bar{u}) = \bar{\alpha}$ and $\bar{u} \in U_{\mathrm{ad}}$. It remains to pass to the limit in the objective. By assumption, we have

$$\liminf_{k\to\infty} F(u_k) \geq F(\bar{u}),$$

and

$$\liminf_{k\to\infty} \mathrm{TV}(\operatorname{sgn}(u_k)) \geq \mathrm{TV}(\bar{\alpha}) = \mathrm{TV}(\operatorname{sgn}(\bar{u}))$$

follows from the properties of TV, see [1, Remark 3.5]. It remains to discuss G. To this end, we define the function $\tilde{g}_i\colon \mathbb{R} \to \mathbb{R}$ via

$$\tilde{g}_i(z) = \begin{cases} g_i(z) & \text{if } z \in [a_i,b_i], \\ g_i(a_i) - (z-a_i)L_g & \text{if } z \leq a_i, \\ g_i(b_i) + (z-b_i)L_g & \text{if } z \geq b_i. \end{cases}$$

Since L_g is a Lipschitz constant of g_i on $[a_i,b_i]$, one can check that $\tilde{g}_i$ is a convex function. For every $i \in \{1,\dots,N\}$, the sequence $\tilde{g}_i(u_{k,i})$ is bounded in $L^2(t_0,T)$, thus, w.l.o.g., we have $\tilde{g}_i(u_{k,i}) \rightharpoonup v_i$ in $L^2(t_0,T)$. We have $(u_{k,i},\tilde{g}_i(u_{k,i})) \in \operatorname{epi}(\tilde{g}_i)$ a.e. on $[t_0,T]$. Since the epigraph $\operatorname{epi}(\tilde{g}_i) \subset \mathbb{R}^2$ is closed and convex, the weak convergence implies $(\bar{u}_i,v_i) \in \operatorname{epi}(\tilde{g}_i)$ a.e. on $[t_0,T]$, i.e., $v_i \geq \tilde{g}_i(\bar{u}_i)$. Since $\operatorname{sgn}(u_{k,i}) \to \operatorname{sgn}(\bar{u}_i)$ in $L^2(t_0,T)$, see [1, Corollary 3.49], we get

$$g_i(u_{k,i}) = \tilde{g}_i(u_{k,i})\operatorname{sgn}(u_{k,i}) \rightharpoonup v_i \operatorname{sgn}(\bar{u}_i) \geq \tilde{g}_i(\bar{u}_i)\operatorname{sgn}(\bar{u}_i) = g_i(\bar{u}_i)$$

in $L^1(t_0,T)$. This shows

$$\lim_{k\to\infty} G_i(u_{k,i}) = \int_{t_0}^{T} v_i \operatorname{sgn}(\bar{u}_i)\,\mathrm{d}t \geq G_i(\bar{u}_i).$$

By combining the above estimates, we get

$$\liminf_{k\to\infty} J(u_k) \geq J(\bar{u}).$$

Since (u_k) is a (subsequence of a) minimizing sequence, this shows that $\bar{u}$ is a minimizer. □

4.2 STATIONARITY CONDITIONS AND CRITICALITY MEASURE

In this section, we will first investigate local optimality conditions for (OC) by combining theory from convex optimal control with results from [17, Proposition 4.4]. These stationarity conditions will be used to define the criticality measure associated with problem (OC).

Definition 4.5 (Minimal representation). Let $u \in L^\infty(t_0,T;\{0\}\cup[a,b])$ be given such that $\mathrm{TV}(\operatorname{sgn}(u)) < \infty$. We set $K := \mathrm{TV}(\operatorname{sgn}(u))/2$. A vector $\hat{t} \in \mathbb{R}^{2K}$, such that

$$t_0 \leq \hat{t}_1 < \hat{t}_2 < \cdots < \hat{t}_{2K} \leq T$$

and such that

$$\operatorname{sgn}(u) = \sum_{j=1}^{K} \chi_{[\hat{t}_{2j-1},\hat{t}_{2j}]} \tag{4.4}$$

holds a.e. on (t_0,T) is called minimal representation.

Proposition 4.6. *The minimal representation from Definition 4.5 is uniquely determined.*

Proof. This follows by adapting the arguments in [17, Proposition 4.4]. □

In what follows, we always redefine u on a set of measure zero such that (4.4) holds everywhere on $[t_0, T]$. This implies

$$\{\operatorname{sgn}(u_i) = 1\} \coloneqq \{t \in [t_0, T] \mid \operatorname{sgn}(u_i(t)) = 1\} = \bigcup_{j=1}^{K_j} [t_{2i-1}, t_{2i}].$$

We refer to $\hat{t}_u \coloneqq \hat{t}$ as the minimal representation of $\operatorname{sgn}(u)$.

In case of a vector-valued function $u \in L^\infty(t_0, T; \bar{M}_0)$ with $\operatorname{TV}(\operatorname{sgn}(u)) < \infty$, we say that $\hat{t}_u \coloneqq (\hat{t}_{u_1}, \dots, \hat{t}_{u_N})$ is the minimal representation of $\operatorname{sgn}(u)$.

We remind the reader, that the jump set of a function α of bounded variation is commonly referred to as J_α. Since we are interested in the switching points of functions $u \in L^\infty(t_0, T)$ with $\operatorname{TV}(\operatorname{sgn}(u)) < \infty$, which are the jump points of $\operatorname{sgn}(u)$, with some abuse of notation, we denote by J_u the jump set of $\operatorname{sgn}(u)$. Note that the jump set of $\operatorname{sgn}(u)$ satisfies

$$J_u = \{\hat{t}_1, \dots, \hat{t}_{2K}\},$$

where $\hat{t} = \hat{t}_u \in \mathbb{R}^{2K}$ is the minimal representation of $\operatorname{sgn}(u)$. This directly yields

$$J_u \coloneqq \{s \in [t_0, T] \mid \operatorname{sgn}(u)(s-) \neq \operatorname{sgn}(u)(s+)\},$$

with the conventions

$$\operatorname{sgn}(u)(t_0-) \coloneqq \operatorname{sgn}(u)(T+) \coloneqq 0.$$

Note that we used one-sided limits of $\operatorname{sgn}(u)$, since the function u itself might not possess one-sided limits. We mention that this characterization of J_u implies that $t_0 \in J_u$ if and only if $\operatorname{sgn}(u(t_0+)) = 1$, and similarly for T.

Since we are interested in defining an appropriate criticality measure, we investigate the necessary local optimality conditions of first order.

Theorem 4.7. *Let $\bar{u}$ be locally optimal for (OC) in the sense of $L^2(t_0, T; \mathbb{R}^N)$. Then for all $i = 1, \dots, N$, the function $\bar{u}_i$ (after modification on a set of measure zero) is continuous on the closed set $\{\operatorname{sgn}(\bar{u}_i) = 1\}$ and the following conditions hold:*

$$-(\nabla F(\bar{u})(t))_i \in \partial g_{i,+}(\bar{u}_i(t)) \qquad \forall t \in \{\operatorname{sgn}(\bar{u}_i) = 1\}, \tag{4.5a}$$

$$-(\nabla F(\bar{u})(t))_i = \frac{g_i(\bar{u}_i(t))}{\bar{u}_i(t)} \qquad \forall t \in J_{\bar{u}_i} \cap (t_0, T), \tag{4.5b}$$

$$-(\nabla F(\bar{u})(t_0))_i \geq \frac{g_i(\bar{u}_i(t_0))}{\bar{u}_i(t_0)} \qquad \text{if } t_0 \in J_{\bar{u}_i}, \tag{4.5c}$$

$$-(\nabla F(\bar{u})(T))_i \geq \frac{g_i(\bar{u}_i(T))}{\bar{u}_i(T)} \qquad \text{if } T \in J_{\bar{u}_i}. \tag{4.5d}$$

Proof. We prove (4.5a) and (4.5b)–(4.5d) separately.

We start by fixing $i \in \{1, \dots, N\}$ and vary only the component $\bar{u}_i$ of the solution. For this, we introduce the abbreviation

$$F_i(u_i) \coloneqq F(\bar{u}_1, \dots, \bar{u}_{i-1}, u_i, \bar{u}_{i+1}, \dots, \bar{u}_N).$$

First, we vary $\bar{u}_i$ only on the set $\{\operatorname{sgn}(\bar{u}_i) = 1\}$. That is, we consider

$$\min_{u_i \in L^2(\{\operatorname{sgn}(\bar{u}_i)=1\})} \quad F_i(\chi_{\{\operatorname{sgn}(\bar{u}_i)=1\}} u_i) + \int_{\{\operatorname{sgn}(\bar{u}_i)=1\}} g_{i,+}(u_i(t))\,\mathrm{d}t.$$

Clearly, $\bar{u}_i$ is a local minimizer of this auxiliary problem. The convexity of $g_{i,+}$ implies that the integral in the objective is a convex function of u_i. Therefore, the necessary optimality condition directly follows from the standard theory of convex optimization, i.e., we get

$$-[\nabla F(\bar{u})(t)]_i \in \partial g_{i,+}(\bar{u}_i(t)) \qquad \text{for a.a. } t \in \{\operatorname{sgn}(\bar{u}_i) = 1\}.$$

By assumption, $g_{i,+}$ is strongly convex and this implies that $(\partial g_{i,+})^{-1} = \partial g_{i,+}^\star$ is single-valued and Lipschitz continuous. Consequently,

$$\bar{u}_i(t) = \partial g_{i,+}^{-1}(-[\nabla F(\bar{u})(t)]_i) \qquad \text{for a.a. } t \in \{\operatorname{sgn}(\bar{u}_i) = 1\}.$$

Together with the assumed continuity of $\nabla F(\bar{u})\colon [t_0, T] \to \mathbb{R}^N$, this implies continuity of the right-hand side w.r.t. t and, therefore, $\bar{u}_i$ (after modification on a set of measure zero) is continuous on the closed set $\{\operatorname{sgn}(\bar{u}_i) = 1\}$. This verifies (4.5a).

For the other conditions we only vary $\operatorname{sgn}(\bar{u}_i)$. To this end, we denote by $\bar{v}_i \in C([t_0, T]; [a_i, b_i])$ an arbitrary continuous extension of $\bar{u}_i|_{\{\operatorname{sgn}(\bar{u}_i)=1\}}$. Further, for $\alpha \in L^1(t_0, T)$ we introduce

$$\tilde{F}(\alpha) \coloneqq F_i(\alpha \bar{v}_i) + \int_{t_0}^{T} \alpha(t) g(\bar{v}_i(t))\,\mathrm{d}t.$$

Now, it is clear that $\bar{\alpha} = \operatorname{sgn}(\bar{u}_i)$ is a local minimizer of

$$\min_{\alpha \in \mathrm{BV}(t_0, T; \{0,1\})} \quad \tilde{F}(\alpha) + \mathrm{TV}(\alpha).$$

Since $g(\bar{v}_i(t))$ is continuous with respect to t by Assumption 4.1(ii) and F already has the assumed regularity from Assumption 4.1(i), the function $\tilde{F}$ is Fréchet differentiable from $L^1(t_0, T)$ to $\mathbb{R}$ and its derivative is given by

$$\nabla \tilde{F}(\alpha) = \nabla F_i(\alpha \bar{v}_i)\bar{v}_i + g(\bar{v}_i) \in C([t_0, T]).$$

This regularity is enough to apply the optimality conditions from [17, Lemma 4.10] (see also the note after [21, Theorem 3.6]), to obtain

$$\nabla \tilde{F}(\operatorname{sgn}(\bar{u}_i))(t) = 0 \qquad \forall t \in J_{\bar{u}_i} \cap (t_0, T)$$

for each point from the jumping set, except at the boundary (recall, that in these references, the usual definition of the total variation is used). The continuity of $\bar{u}_i$ on $\{\operatorname{sgn}(\bar{u}_i) = 1\}$ implies that $\bar{v}_i(t) = \bar{u}_i(t) \geq a_i > 0$ at all switching points $t \in J_{\bar{u}_i}$. Consequently, (4.5b) follows.

If a jump occurs at a boundary point, then this switching point can only be moved in one direction and a simple argument yields

$$\nabla \tilde{F}(\operatorname{sgn}(\bar{u}_i))(t_0) \leq 0 \quad \text{if } t_0 \in J_{\bar{u}_i} \qquad \text{and} \qquad \nabla \tilde{F}(\operatorname{sgn}(\bar{u}_i))(T) \leq 0 \quad \text{if } T \in J_{\bar{u}_i}.$$

Together with the above formula for $\nabla \tilde{F}(\operatorname{sgn}(\bar{u}_i))$, the stationarity conditions (4.5) follow. □

In order to shed some light on the connection of (4.5a) and (4.5b), we provide a small lemma.

Lemma 4.8. *For all $i = 1, \dots, N$,*

- *the function $v \mapsto g_i(v)/v$ has a unique minimizer $u_i^\star$ on $[a_i, b_i]$,*

- *the inclusion* $g_i(v)/v \in \partial g_{i,+}(v)$ *has a unique solution* $v_i^\star \in [a_i, b_i]$,

and both coincide, i.e., $u_i^\star = v_i^\star$.

Proof. For simplicity, we drop the index i in the proof. We have the chain of equivalences

$$\begin{aligned} v^\star \text{ minimizes } v \mapsto g(v)/v \quad &\iff \quad \frac{g(v)}{v} \geq \frac{g(v^\star)}{v^\star} && \forall v \in [a,b] \\ &\iff \quad g(v) \geq g(v^\star) + \frac{g(v^\star)}{v^\star}(v - v^\star) && \forall v \in [a,b] \\ &\iff \quad \frac{g(v^\star)}{v^\star} \in \partial g_+(v^\star). \end{aligned}$$

Note that g and g_+ coincide on $[a,b]$. This shows that minimizers of $v \mapsto g(v)/v$ and solutions of $g(v)/v \in \partial g_+(v)$ coincide. The strong convexity of g implies

$$g(v) > g(v^\star) + \frac{g(v^\star)}{v^\star}(v - v^\star) \qquad \forall v \in [a,b] \setminus \{v^\star\}$$

and, consequently, $g(v)/v > g(v^\star)/v^\star$ for all $v \in [a,b] \setminus \{v^\star\}$. This provides the desired uniqueness. □

Remark 4.9. Let us assume that $\bar{u}$ satisfies the conditions (4.5a) and (4.5b). For a switching point $s \in J_{\bar{u}_i} \cap (t_0, T)$, this implies

$$\frac{g_i(\bar{u}_i(s))}{\bar{u}_i(s)} \in \partial g_{i,+}(\bar{u}_i(s)).$$

Due to Lemma 4.8, the value $\bar{u}_i(s)$ is uniquely determined and it equals

$$u_i^\star := \operatorname*{arg\,min}_{v \in [a_i, b_i]} \frac{g_i(v)}{v},$$

i.e., $\bar{u}_i(s) = u_i^\star$. This is somewhat surprising. It characterizes $\bar{u}_i(s)$ as the point at which a tangent of g_i goes through the origin. This means that independently of the choice for F, switches can only occur from and to the predetermined value $u_i^\star$, solely depending on g_i.

At the boundary switching points $s \in J_{\bar{u}_i} \cap \{t_0, T\}$, one can argue similarly. This shows that (4.5c) and (4.5d) imply $\bar{u}_i(s) \geq u_i^\star$.

We already know from Section 3 that a criticality measure for a convex problem can be defined via the proximal point mapping. The remaining part of the criticality measure is built from the additional stationarity conditions at the switching points. For the sake of clarity it is beneficial to introduce the criticality measure in two parts. For simplicity of notation we introduce the following shorthand.

Definition 4.10. Let $u \in U_{\mathrm{ad}}$ be given. For $p \in [1, \infty)$ and $f \in L^p(t_0, T; \mathbb{R}^N)$ we define

$$\|f\|^p_{L^p(\{\operatorname{sgn}(u)=1\})} := \sum_{i=1}^N \int_{\{\operatorname{sgn}(u_i)=1\}} |f_i(t)|^p \, \mathrm{d}t \in [0, \infty).$$

Similarly, for $u, w \in U_{\mathrm{ad}}$ and $f \in L^p(t_0, T; \mathbb{R}^N)$ we define

$$\|f\|^p_{L^p(\{\operatorname{sgn}(uw)=1\})} := \sum_{i=1}^N \int_{\{\operatorname{sgn}(u_i)=1\} \cap \{\operatorname{sgn}(w_i)=1\}} |f_i(t)|^p \, \mathrm{d}t \in [0, \infty).$$

Definition 4.11. The criticality measure for a component i of a feasible control $u \in U_{\mathrm{ad}}$ at a single switching point $s \in J_{u_i}$ is defined via:

$$\begin{aligned} v_{u,i}(s) &:= \operatorname*{arg\,min}_{z\in[a_i,b_i]} [\nabla F(u)(s)]_i z + g_i(z), \\ V_{u,i}(s) &:= \min_{z\in[a_i,b_i]} [\nabla F(u)(s)]_i z + g_i(z), \\ C_{\mathrm{switch}}(u,i,s) &:= \max\left\{ V_{u,i}(s), -V_{u,i}(s)\frac{\min\{s-t_0, T-s\}}{T-t_0} \right\}. \end{aligned}$$

For a fixed value $r > 0$, we define the criticality measure for the active part of the control and combine both to obtain the actual criticality measure:

$$C_{\mathrm{prox}}(u) := \frac{1}{2r^2} \|u - \mathrm{prox}_{rG_+}(u - r\nabla F(u))\|^2_{L^2(\{\mathrm{sgn}(u)=1\})},$$

$$C_{\mathrm{switch}}(u) := \max_{i=1,\dots,N} \max_{s\in J_{u_i}} C_{\mathrm{switch}}(u,i,s),$$

$$C(u) := \max\{C_{\mathrm{prox}}(u), C_{\mathrm{switch}}(u)\}.$$

In case $N = 1$ (which will be assumed in some proofs for simplicity of the presentation), we write $v_u(s)$, $V_u(s)$ and $C_{\mathrm{switch}}(u,s)$ instead of $v_{u,1}(s)$, $V_{u,1}(s)$ and $C_{\mathrm{switch}}(u,1,s)$, respectively.

Corollary 4.12. *For all $u \in U_{\mathrm{ad}}$, we have $C(u) \geq 0$. A feasible point $\bar{u} \in U_{\mathrm{ad}}$ is stationary in the sense of (4.5) if and only if $C(\bar{u}) = 0$.*

Proof. The term $\frac{\min\{s-t_0,T-s\}}{T-t_0}$ appearing in $C_{\mathrm{switch}}(u,i,s)$ is always non-negative, consequently, $C_{\mathrm{switch}}(u,i,s) \geq 0$. Together with $C_{\mathrm{prox}}(u) \geq 0$, we get $C(u) \geq 0$.

From the definition of G_+, it follows that prox_{G_+} can be evaluated componentwise and pointwise. This implies that $C_{\mathrm{prox}}(\bar{u}) = 0$ is equivalent to (4.5a).

For a switching point $s \in J_{\bar{u}_i} \cap (t_0, T)$ we get

$$\begin{aligned} C_{\mathrm{switch}}(u,i,s) = 0 &\iff V_{u,i}(s) = 0 \\ &\iff [\nabla F(\bar{u})(s)]_i v_{u,i}(s) + g_i(v_{u,i}(s)) = 0 \\ &\iff -[\nabla F(\bar{u})(s)]_i = \frac{g_i(v_{u,i}(s))}{v_{u,i}(s)}. \end{aligned}$$

Thus, the condition $C_{\mathrm{switch}}(u,i,s) = 0$ is equivalent to (4.5b) for all $s \in J_{\bar{u}_i} \cap (t_0, T)$. Similarly, for all $s \in J_{\bar{u}_i} \cap \{t_0, T\}$, $C_{\mathrm{switch}}(u,i,s) = 0$ if and only if $V_{u,i}(s) \leq 0$ which is equivalent to (4.5c), (4.5d). □

Remark 4.13. The asymmetric definition of $C_{\mathrm{switch}}(u,i,s)$ may appear surprising at first, but is a consequence of including switches at the boundary points t_0 and T for the computation of $\mathrm{TV}(\mathrm{sgn}(u))$. Scaling $-V_{u,i}(s)$ with the distance to the boundary prevents a discontinuity in the criticality measure, when moving the first switch to t_0 or the last switch to T, see the proof of Lemma 4.23 below.

Remark 4.14. We give a motivation for the definition of $V_{u,i}(s)$. For an interior switching point $s \in J_{u_i} \cap (t_0, T)$, the optimality condition $C(u) = 0$ from Corollary 4.12 implies $V_{u,i}(s) = 0$. If $V_{u,i}(s) \neq 0$, the sign of $V_{u,i}(s)$ indicates in which direction the switching point s should be moved in order to improve the objective value. For another control $w \in U_{\mathrm{ad}}$ in the neighborhood of u, a partial linearization of the objective leads to

$$\int_{t_0}^{T} \nabla F(u)(t)^\top (w(t) - u(t)) + g(w(t)) - g(u(t))\,\mathrm{d}t.$$

In case $V_{u,i}(s) > 0$, we have $[\nabla F(u)(s)]_i u_i(t) + g_i(u_i(t)) > 0$ for all t with $u_i(t) > 0$. By continuity, we get $[\nabla F(u)(t)]_i u_i(t) + g_i(u_i(t)) > 0$ for all t with $u_i(t) > 0$ close enough to s. Thus, we can improve the objective functional by switching off u_i in the neighborhood of the switching point s. In the other case $V_{u,i}(s) < 0$, we similarly get $[\nabla F(u)(t)]_i v_{u,i}(s) + g_i(v_{u,i}(s)) < 0 = [\nabla F(u)(t)]_i u_i(t) + g_i(u_i(t))$ for all t with $u_i(t) = 0$ close enough to s. Consequently, the objective can be improved by activating the control u_i in this neighborhood of the switch s and setting it to the value $v_{u,i}(s)$. This will be made precise in the proof of Lemma 4.19.

4.3 APPLICATION OF THE TRUST-REGION METHOD

The goal of this section is to apply the abstract trust-region method from Section 2 to solve (OC). We define the set X from Assumption 2.1 to be the admissible set, i.e.,

$$X := U_{\mathrm{ad}}.$$

[17, 18] showed convergence of a trust-region method for optimal control of ordinary differential equations with integer-valued controls of bounded variation. This would be applicable to (OC) if we would have $a_i = b_i$. In order to generalize to our situation $a_i < b_i$, we incorporate the proximal gradient method into the trust-region method for switching point optimization and verify Assumption 2.1 to prove convergence for the trust-region method applied to (OC) in Theorem 4.24. Now, we define the model function, the oracle and the prediction function.

Definition 4.15. Consider a point $u \in U_{\mathrm{ad}}$ and $\Delta > 0$. The model function $m_{u,\Delta} \colon U_{\mathrm{ad}} \to \mathbb{R}$ is defined as

$$\begin{aligned} m_{u,\Delta}(w) := {} & \langle \nabla F(u), w - u \rangle_{L^2(t_0,T;\mathbb{R}^N)} + G(w) - G(u) \\ & + \mathrm{TV}(\mathrm{sgn}(w)) - \mathrm{TV}(\mathrm{sgn}(u)) + \frac{1}{2\Delta} \|w - u\|^2_{L^2(\{\mathrm{sgn}(uw)=1\})}. \end{aligned} \tag{4.6}$$

The oracle function $T \colon U_{\mathrm{ad}} \times \mathbb{R}_+ \to U_{\mathrm{ad}}$ is defined by minimizing the model function over a trust region, i.e.,

$$T(u,\Delta) \in \arg\min \big\{ m_{u,\Delta}(w) \;\big|\; w \in U_{\mathrm{ad}}, \|\mathrm{sgn}(u) - \mathrm{sgn}(w)\|_{L^1(t_0,T;\mathbb{R}^N)} \le \Delta \big\}.$$

The prediction function is defined as

$$\mathrm{pred}(u,\Delta) := m_{u,\Delta}(u) - m_{u,\Delta}(T(u,\Delta)) = -m_{u,\Delta}(T(u,\Delta)). \tag{4.7}$$

Conceptually, the oracle function will select an optimal switching pattern $\mathrm{sgn}(w)$ for the model problem inside the trust region, while a proximal gradient step with step length Δ is performed on the set $\{\mathrm{sgn}(uw) = 1\}$. On the newly active set $\{\mathrm{sgn}(w) = 1\} \setminus \{\mathrm{sgn}(u) = 1\}$, an optimal continuation is chosen. Deriving an explicit formula for the oracle function will be done in Section 4.4.

Since the model function $m_{u,\Delta}$ is not convex, the minimizer might not be unique and for the oracle, we can use any of its minimizers. Note that the existence of a minimizer of the model function inside the trust region can be shown along the lines of the proof of Theorem 4.4 by using the following lemma.

Lemma 4.16. *Let $u \in U_{\mathrm{ad}}$ and a sequence $(w_k) \subset U_{\mathrm{ad}}$ be given such that $w_k \rightharpoonup \bar{w}$ in $L^2(t_0,T;\mathbb{R}^N)$ and $\mathrm{sgn}(w_k) \to \mathrm{sgn}(\bar{w})$ in $L^1(t_0,T;\mathbb{R}^N)$. Then,*

$$\|u - \bar{w}\|^2_{L^2(\{\mathrm{sgn}(u\bar{w})=1\})} \le \liminf_{k\to\infty} \|u - w_k\|^2_{L^2(\{\mathrm{sgn}(uw_k)=1\})}.$$

Proof. For simplicity, we argue in the case $N = 1$. From the weak lower semicontinuity of the norm in L^2, we get

$$\|u - \bar{w}\|^2_{L^2(\{\operatorname{sgn}(u\bar{w})=1\})} \le \liminf_{k\to\infty} \|u - w_k\|^2_{L^2(\{\operatorname{sgn}(u\bar{w})=1\})}.$$

Further,

$$\begin{aligned}
&\left| \|u - w_k\|^2_{L^2(\{\operatorname{sgn}(u\bar{w})=1\})} - \|u - w_k\|^2_{L^2(\{\operatorname{sgn}(uw_k)=1\})} \right| \\
&\quad = \|u - w_k\|^2_{L^2(\{\operatorname{sgn}(u\bar{w})=1\} \triangle \{\operatorname{sgn}(uw_k)=1\})} \\
&\quad \le \|u - w_k\|^2_{L^2(\{\operatorname{sgn}(\bar{w})=1\} \triangle \{\operatorname{sgn}(w_k)=1\})} \\
&\quad \le \|u - w_k\|^2_{L^\infty(t_0,T)} |\{\operatorname{sgn}(\bar{w}) = 1\} \triangle \{\operatorname{sgn}(w_k) = 1\}| \\
&\quad = \|u - w_k\|^2_{L^\infty(t_0,T)} \|\operatorname{sgn}(\bar{w}) - \operatorname{sgn}(w_k)\|_{L^1(t_0,T)} \to 0,
\end{aligned}$$

where $\triangle$ denotes the symmetric difference. In the last equality, we used that $\operatorname{sgn}(\bar{w})$ and $\operatorname{sgn}(w_k)$ only take values in $\{0, 1\}$. This shows the claim. □

Now, we have defined the ingredients for applying the abstract Algorithm 2.1 to (OC). It remains to verify Assumption 2.1.

Lemma 4.17 (Monotonicity). *The prediction function* (4.7) *is nondecreasing, i.e., Assumption 2.1*(I) *holds.*

Proof. As Δ increases, the model function $m_{u,\Delta}$ decreases and the trust region becomes larger. This yields the claim. □

For the verification of the remaining parts of Assumption 2.1, we combine the arguments from [18, Section 5] with the results from Section 3.2. This means that we have to define $\underline{\Delta}_a(u)$ accordingly.

Definition 4.18. Given $u \in U_{\mathrm{ad}}$ and the minimal representation $\hat{t}_u$ for $\operatorname{sgn}(u)$ from Proposition 4.6, we define $\underline{\Delta}_a(u) \in (0, r]$ via the smallest gap between two sign changes

$$\underline{\Delta}_a(u) := \min\left\{ r, \min_{i=1,\dots,N} \min_{j=1,\dots,2K_i-1} ((\hat{t}_{u_i})_{j+1} - (\hat{t}_{u_i})_j) \right\},$$

where $K_i := \operatorname{TV}(\operatorname{sgn}(u_i))/2$ and r is as in the definition of the criticality measure, see Definition 4.11.

Lemma 4.19 (Sufficient decrease). *The prediction function* (4.7) *satisfies Assumption 2.1*(II) *with* $\underline{\Delta}_a(u)$ *as defined in Definition 4.18,* $f \equiv \mathrm{Id}$, $c = \frac{1}{2} L_{\nabla F} b$ *and* $s = 2$.

Proof. Let $u \in U_{\mathrm{ad}}$ and $\Delta \in (0, \underline{\Delta}_a(u))$ be given. Since the predicted decrease $\operatorname{pred}(u, \Delta)$ is defined via the solution of the model function $m_{u,\Delta}$, we obtain the estimate

$$\operatorname{pred}(u, \Delta) \ge -m_{u,\Delta}(w)$$

for all $w \in U_{\mathrm{ad}}$ with $\|\operatorname{sgn}(u) - \operatorname{sgn}(w)\|_{L^1(t_0,T;\mathbb{R}^N)} \le \Delta$. We will construct several feasible points w for comparison. First, we consider

$$w_{\mathrm{prox}} := \operatorname{prox}_{\Delta G_+}(u - \Delta \nabla F(u))$$

and set $w_i := w_{\mathrm{prox},i} \chi_{\{\operatorname{sgn}(u_i)=1\}}$, $i = 1, \dots, N$. It is clear that $\operatorname{sgn}(w) = \operatorname{sgn}(u)$. Consequently, this case is completely analogous to Lemma 3.9. Since $\Delta < \underline{\Delta}_a(u) \le r$, where $r > 0$ was used in Definition 4.11 for the definition of the criticality measure, this lemma implies

$$\operatorname{pred}(u, \Delta) \ge -m_{u,\Delta}(w) \ge \frac{1}{2r^2} \|u - \operatorname{prox}_{rG}(u - r\nabla F(u))\|^2_{L^2(\{\operatorname{sgn}(u)=1\})} \Delta = C_{\mathrm{prox}}(u)\Delta. \tag{4.8}$$

Next, we consider a switching point $s \in J_{u_i} \cap [t_0, T)$. We construct w by shifting this switch by at most $\Delta < \underline{\Delta}_a(u)$. Since the distance between any two switches is at least $\underline{\Delta}_a(u)$, we have $\mathrm{TV}(\mathrm{sgn}(u)) = \mathrm{TV}(\mathrm{sgn}(w))$. W.l.o.g., we consider the case $N = 1$ and, further, $\mathrm{sgn}(u(s-)) = 0$ and $\mathrm{sgn}(u(s+)) = 1$, i.e., the control is switched on at s. The other cases can be handled similarly. From Assumption 4.1(i) we get that $\nabla F(u)$ is Lipschitz continuous with constant $L_{\nabla F(u)} \leq L_{\nabla F}$ and this implies

$$\begin{aligned}\left|\int_s^{s+h} (\nabla F(u)(s) - \nabla F(u)(t))w(t)\,\mathrm{d}t\right| &\leq \|w\|_{L^\infty(s,s+h)} \int_s^{s+h} |\nabla F(u)(s) - \nabla F(u)(t)|\,\mathrm{d}t \\ &\leq L_{\nabla F(u)} b \int_s^{s+h} |s-t|\,\mathrm{d}t \leq \frac{1}{2}h^2 L_{\nabla F} b \end{aligned} \tag{4.9}$$

for feasible $w \in U_{\mathrm{ad}}$ and $h > 0$ small enough such that $s + h \in (t_0, T)$.

We make a distinction by cases with respect to the sign of $V_u(s)$ from Definition 4.11. As mentioned in Remark 4.14, the sign of $V_u(s)$ encodes in which direction the switching point should be moved.

In case $V_u(s) > 0$, we first note that $s < T - \Delta$, since the control must be switched off at least once in $(s, T]$ and any two switching points are more than Δ apart. Thus, we can define $w := u - u\chi_{[s,s+\Delta]}$, i.e., the control is switched on a little bit later. We immediately get $\mathrm{TV}(\mathrm{sgn}(w)) = \mathrm{TV}(\mathrm{sgn}(u))$ and $\|w - u\|^2_{\{\mathrm{sgn}(uw)=1\}} = 0$. Together with $w = 0$ on $(s, s + \Delta)$, we have

$$\begin{aligned}\mathrm{pred}(u,\Delta) \geq -m_{u,\Delta}(w) &= \langle -\nabla F(u), w - u\rangle_{L^2(t_0,T)} + G(u) - G(w) \\ &= \int_s^{s+\Delta} -\nabla F(u)(t)(w(t) - u(t))\,\mathrm{d}t + \int_s^{s+\Delta} g(u(t)) - g(w(t))\,\mathrm{d}t \\ &\overset{(4.9)}{\geq} \int_s^{s+\Delta} \nabla F(u)(s)u(t) + g(u(t))\,\mathrm{d}t - \frac{1}{2}\Delta^2 L_{\nabla F} b \\ &\geq \Delta V_u(s) - \frac{1}{2}\Delta^2 L_{\nabla F} b = \Delta C_{\mathrm{switch}}(u,1,s) - \frac{1}{2}\Delta^2 L_{\nabla F} b,\end{aligned}$$

where the last estimate follows from the definition of $V_u(s)$ and the last equality is the definition of $C_{\mathrm{switch}}(u, 1, s)$.

In case $V_u(s) < 0$, we have

$$0 < C_{\mathrm{switch}}(u,1,s) = -V_u(s)\frac{\min\{s - t_0, T - s\}}{T - t_0}$$

and this implies $s > t_0$. Now, we want to move the switching point to the left. If s is near the initial time t_0, this might limit our possibility to move the switch. Therefore, we set $h := \min\{\Delta, s - t_0\}$. We use $v_u(s)$ from Definition 4.11 (i.e., the minimizer of $w \mapsto [\nabla F(u)(s)]w + g(w)$) to define $w := u + v_u(s)\chi_{[s-h,s]}$, i.e., the control is switched on slightly earlier with the fixed value $v_u(s)$. We proceed similar to the above estimate. Using $u = 0$ on $(s - h, s)$, we get

$$\begin{aligned}\mathrm{pred}(u,\Delta) &\geq \int_{s-h}^{s} -\nabla F(u)(t)(w(t) - u(t))\,\mathrm{d}t + \int_{s-h}^{s} -g(w(t)) + g(u(t))\,\mathrm{d}t \\ &\overset{(4.9)}{\geq} \int_{s-h}^{s} -\nabla F(u)(s)v_u(s) - g(v_u(s))\,\mathrm{d}t - \frac{1}{2}h^2 L_{\nabla F}\|v_u(s)\|_{L^\infty(s-h,s)} \\ &= h(-V_u(s)) - \frac{1}{2}h^2 L_{\nabla F} b.\end{aligned}$$

We continue with

$$\begin{aligned}h(-V_u(s)) = -V_u(s)\min\{\Delta, s - t_0\} &= -V_u(s)\frac{(s - t_0)\Delta}{\max\{s - t_0, \Delta\}} \\ &\geq -V_u(s)\frac{s - t_0}{T - t_0}\Delta \geq \Delta C_{\mathrm{switch}}(u,1,s),\end{aligned}$$

where the first estimate uses $\Delta \le T - t_0$, $s - t_0 \le T - t_0$, and the second estimate follows from the definition of $C_{\text{switch}}(u, 1, s)$. Combining the last two equations, we again get

$$\operatorname{pred}(u, \Delta) \ge \Delta C_{\text{switch}}(u, 1, s) - \frac{1}{2} h^2 L_{\nabla F} b.$$

Finally, if $V_u(s) = 0$, the same inequality holds trivially, since $\operatorname{pred}(u, \Delta) \ge 0$ and $C_{\text{switch}}(u, 1, s) = 0$.

Since the criticality measure is defined as the maximum of the proximal part and the term for the switching points, we obtain

$$\begin{aligned}\operatorname{pred}(u, \Delta) &\ge \max\left\{C_{\text{prox}}(u)\Delta, C_{\text{switch}}(u)\Delta - \frac{1}{2} L_{\nabla F} \Delta^2 b\right\} \\ &\ge C(u)\Delta - \frac{1}{2} L_{\nabla F} \Delta^2 b \qquad \forall \Delta \in (0, \underline{\Delta}_a(u)).\end{aligned}$$

□

Remark 4.20. The above proof reveals some interesting properties. First, we can guarantee the prediction estimate by analyzing the position of the switches and the improvement by the proximal gradient method on $\{\operatorname{sgn}(u) = 1\}$ independently. Further a switching point is either stationary in the sense of (4.5b)–(4.5c) or can always be moved to guarantee an improvement. While the proof is constructive, the constructed improvement always retains the switching behavior. However, a minimizer $T(u, \Delta)$ of the trust-region subproblem (Definition 4.15) is not forced to follow the same switching pattern as u. Consequently, the trust-region algorithm may produce solutions with better function values compared to a pure switching-point optimization approach.

In the next lemma, we check that the oracle stays in a scaled trust region, similar to Lemma 3.6.

Lemma 4.21. *For all $u \in U_{\text{ad}}$, we have*

$$\|T(u, \Delta) - u\|_{L^1(t_0, T; \mathbb{R}^N)} \le \rho \Delta \qquad \forall \Delta \in (0, \infty),$$

with $\rho \coloneqq 2(C_{\nabla F} + L_g)N(T - t_0) + b$.

Proof. We set $\hat{u} = T(u, \Delta)$. We use the fact that the norm can be split into the part where both controls are active and a remainder. For $i \in \{1, \dots, N\}$ and $t \in \{\operatorname{sgn}(u_i \hat{u}_i) = 1\}$, we have

$$\hat{u}_i(t) = \operatorname{prox}_{\Delta g_i^+}(u_i(t) - \Delta[\nabla F(u)(t)]_i).$$

Now, we argue as in (3.4) and obtain

$$|\hat{u}_i(t) - u_i(t)| \le 2\big(|[\nabla F(u)(t)]_i| + L_g\big)\Delta \le 2\big(C_{\nabla F} + L_g\big)\Delta \qquad \forall t \in \{\operatorname{sgn}(u_i \hat{u}_i) = 1\},$$

see (4.2). On the other hand, if $t \notin \{\operatorname{sgn}(u_i \hat{u}_i) = 1\}$, at least one of the controls is zero. Unless both are zero, we have $|\operatorname{sgn}(\hat{u}_i(t)) - \operatorname{sgn}(u_i(t))| = 1$. This implies

$$|\hat{u}_i(t) - u_i(t)| \le b|\operatorname{sgn}(\hat{u}_i(t)) - \operatorname{sgn}(u_i(t))| \qquad \forall t \notin \{\operatorname{sgn}(u_i \hat{u}_i) = 1\}.$$

By integrating over these inequalities, we get

$$\begin{aligned}\|\hat{u} - u\|_{L^1(t_0, T; \mathbb{R}^N)} &= \int_{t_0}^{T} \sum_{i=1}^{N} |\hat{u}_i(t) - u_i(t)| \, \mathrm{d}t \\ &\le 2\big(C_{\nabla F} + L_g\big)\Delta N(T - t_0) + b\|\operatorname{sgn}(\hat{u}) - \operatorname{sgn}(u)\|_{L^1(t_0, T; \mathbb{R}^N)} \\ &\le 2\big(C_{\nabla F} + L_g\big)\Delta N(T - t_0) + b\Delta,\end{aligned}$$

where we used the trust-region constraint from the definition of the oracle, see Definition 4.15. This shows the claim. □

Lemma 4.22 (Accuracy). *For an arbitrary $\eta \in (0,1]$, Assumption 2.1(III) and (III.E) are satisfied with C, r as in Definition 4.11, $\underline{\Delta}_a$ from Definition 4.18, ρ from Lemma 4.21, $\delta = 1$, $s = 2$,*

$$c \coloneqq \frac{1}{2} L_{\nabla F}\left(b + \frac{\rho^2}{1-\eta}\right), \qquad \text{and} \qquad \underline{\Delta}_b \coloneqq \min\left\{\frac{r}{2}, \frac{1}{(C_{\nabla F} + L_g)b}, \sqrt{\frac{2(1-\eta)}{L_{\nabla F}\rho^2}}\right\}.$$

Proof. The proof can be carried out similarly to Lemma 3.10, but we have to be careful with the case of switching points being too close, i.e., $\underline{\Delta}_a(u) \le \Delta$. This leads to the corresponding exception (III.E), which will be verified analogously to [18, Lemma 5.4].

We start with an arbitrary $\Delta \in (0, r)$. First, we define $\hat{u} \coloneqq T(u, \Delta)$ and $\operatorname{ared}(u, \Delta)$ will be estimated using the first-order approximation

$$F(u) - F(\hat{u}) \ge \langle \nabla F(u), u - \hat{u} \rangle_{L^2} - \frac{1}{2} L_{\nabla F} \|u - \hat{u}\|^2_{L^1(0,T;\mathbb{R}^N)}$$

of F, which follows from the Lipschitz continuity of $\nabla F \colon L^1(t_0, T; \mathbb{R}^N) \to L^\infty(t_0, T; \mathbb{R}^N)$. Using this estimate yields

$$\begin{aligned}
\operatorname{ared}(u,\Delta) &= J(u) - J(\hat{u}) \\
&= F(u) + G(u) + \operatorname{TV}(\operatorname{sgn}(u)) - F(\hat{u}) - G(\hat{u}) - \operatorname{TV}(\operatorname{sgn}(\hat{u})) \\
&\ge \Bigg(- \langle \nabla F(u), \hat{u} - u \rangle_{L^2} + G(u) - G(\hat{u}) \\
&\qquad + (\operatorname{TV}(\operatorname{sgn}(u)) - \operatorname{TV}(\operatorname{sgn}(\hat{u}))) - \frac{1}{2\Delta} \|\hat{u} - u\|^2_{\operatorname{sgn}(u\hat{u})=1} \Bigg) \\
&\quad - \frac{1}{2} L_{\nabla F} \|u - \hat{u}\|^2_{L^1(0,T;\mathbb{R}^N)} + \frac{1}{2\Delta} \|\hat{u} - u\|^2_{L^2(\{\operatorname{sgn}(u\hat{u})=1\})} \\
&\ge \operatorname{pred}(u,\Delta) - \frac{1}{2} L_{\nabla F} \|u - \hat{u}\|^2_{L^1(0,T;\mathbb{R}^N)}.
\end{aligned}$$

Combining this with Lemma 4.21, we get

$$\operatorname{ared}(u,\Delta) - \eta \operatorname{pred}(u,\Delta) \ge (1-\eta) \operatorname{pred}(u,\Delta) - \frac{1}{2} L_{\nabla F} \rho^2 \Delta^2 \qquad \forall \Delta \in (0, r). \tag{4.10}$$

In case $\Delta \in (0, \underline{\Delta}_a(u))$, we can insert the predicted reduction from Lemma 4.19 and obtain

$$\operatorname{ared}(u,\Delta) - \eta \operatorname{pred}(u,\Delta) \ge (1-\eta) \left(C(u)\Delta - c\Delta^2\right)$$

with the constant c as defined in the claim, proving Assumption 2.1(III) for $\Delta \in (0, \underline{\Delta}_a(u))$.

It remains to check Assumption 2.1(III.E), i.e., the case $\Delta \in [\underline{\Delta}_a(u), \underline{\Delta}_b]$. Since $\underline{\Delta}_b < r$, this implies $\underline{\Delta}_a(u) < r$ and, consequently, there exist indices $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, 2K_i - 1\}$ such that $(\hat{t}_{u_i})_{j+1} - (\hat{t}_{u_i})_j = \underline{\Delta}_a(u) \le \Delta$. This means that in the component u_i we have a short time distance between two consecutive switches (either on-off-on or off-on-off). We will see that this situation can be improved by simply removing both switches. For brevity, we set $N = 1$ and only consider the on-off-on situation. We further set $s \coloneqq (\hat{t}_u)_j$ and $h \coloneqq \underline{\Delta}_a(u) = (\hat{t}_u)_{j+1} - (\hat{t}_u)_j$. We define $\hat{u} \in U_{\mathrm{ad}}$ via $\hat{u} = u + a\chi_{(s,s+h)}$. Due to $\|\operatorname{sgn}(\hat{u}) - \operatorname{sgn}(u)\|_{L^1(t_0,T)} = h \le \Delta$, $\hat{u}$ belongs to the trust region. One has $\operatorname{TV}(\operatorname{sgn}(\hat{u})) = \operatorname{TV}(\operatorname{sgn}(u)) - 2$ and the model function (4.6) can be estimated via

$$m_{u,\Delta}(\hat{u}) \le \int_s^{s+h} \nabla F(u)(t)(\hat{u}(t) - u(t)) \, \mathrm{d}t + \int_s^{s+h} g(t, \hat{u}(t)) - g(t, u(t)) \, \mathrm{d}t - 2.$$

With the global Lipschitz constant L_g of g and the Lipschitz constant $C_{\nabla F}$ of F from Remark 4.2, one has

$$\operatorname{pred}(u,\Delta) \geq -m_{u,\Delta}(\hat{u}) \geq 2 - C_{\nabla F}\|\hat{u}-u\|_{L^1(t_0,T;\mathbb{R}^N)} - L_g\|\hat{u}-u\|_{L^1(t_0,T;\mathbb{R}^N)}.$$

The trust-region constraint for (4.6) yields $\|\hat{u}-u\|_{L^1(t_0,T)} \leq b\Delta$ and this implies

$$\operatorname{pred}(u,\Delta) \geq 2 - (C_{\nabla F}+L_g)b\Delta \geq 1 \qquad \forall \Delta \in [\underline{\Delta}_a(u), \underline{\Delta}_b].$$

If we use this estimate for the predicted reduction we can revisit (4.10) and obtain

$$\operatorname{ared}(u,\Delta) - \eta \operatorname{pred}(u,\Delta) \geq (1-\eta) - \frac{1}{2}L_{\nabla F}\rho^2\Delta^2 \geq 0 \qquad \forall \Delta \leq \underline{\Delta}_b.$$

This verifies Assumption 2.1(III.E). □

Lemma 4.23 (Stability). *For the criticality measure $C(u)$ defined in Definition 4.11 and the corresponding prediction function (4.7), there exist constants $\underline{\Delta}_c > 0$ and $L > 0$, such that*

$$|C(T(u,\Delta)) - C(u)| \leq L\Delta \quad \vee \quad \operatorname{pred}(u,\Delta) \geq \frac{1}{2} \qquad \forall \Delta \leq \underline{\Delta}_c,$$

i.e., Assumption 2.1(IV) or (IV.E) hold for all $\Delta \leq \underline{\Delta}_c$ with $\underline{\Delta}_c \coloneqq \min\{((C_{\nabla F}+L_g)\rho)^{-1}, \underline{\Delta}_b\}$ for ρ as defined in Lemma 4.21.

Proof. Define $\hat{u} = T(u,\Delta)$. The proof is done by estimating the individual parts of the criticality measure and we start with the criticality measure for the proximal gradient step C_{prox}. We partition it into the part where u and $\hat{u}$ are active and a remainder term, which is bounded using the trust-region radius:

$$\begin{aligned} C_{\text{prox},u}(u) &\coloneqq \frac{1}{2r^2}\|u - \operatorname{prox}_{rG_+}(u - r\nabla F(u))\|^2_{L^2(\{\operatorname{sgn}(u\hat{u})=1\})} \\ C_{\text{prox}}(u) &= C_{\text{prox},u}(u) + \frac{1}{2r^2}\|u - \operatorname{prox}_{rG_+}(u - r\nabla F(u))\|^2_{L^2(\{\operatorname{sgn}(u)=1\wedge\operatorname{sgn}(\hat{u})=0\})} \\ &\leq C_{\text{prox},u}(u) + \frac{(b-a)^2}{2r^2}\Delta. \end{aligned}$$

Note that $\hat{u} = \operatorname{prox}_{\Delta G_+}(u - \Delta\nabla F(u))$ on $\{\operatorname{sgn}(u\hat{u}) = 1\}$. Hence, we can directly use Lemma 3.12 for $|C_{\text{prox},u} - C_{\text{prox},\hat{u}}| \leq L_C\Delta$ to recover Lipschitz continuity for the proximal gradient part, i.e.,

$$|C_{\text{prox}}(u) - C_{\text{prox}}(\hat{u})| \leq \left(L_C + \frac{(b-a)^2}{2r^2}\right)\Delta.$$

Regarding C_{switch}, an apparent issue is that the position of switching points of $\hat{u}$ in relation to those of u is not known at first. W.l.o.g., we limit the discussion to the case $N = 1$ for brevity of the notation. We will argue similar as in [18, Lemma 5.5] to relate the switching points of u and $\hat{u}$. We define $M \coloneqq (C_{\nabla F} + L_g)\rho$. Then,

$$\begin{aligned} |\operatorname{pred}(u,\Delta) - (\operatorname{TV}(\operatorname{sgn}(u)) - \operatorname{TV}(\operatorname{sgn}(\hat{u})))| &= \langle \nabla F(u), u-\hat{u}\rangle_{L^2(t_0,T)} + G(u) - G(\hat{u}) \\ &\leq M\Delta. \end{aligned}$$

Now the choice $\underline{\Delta}_c \coloneqq \min\{M^{-1}, \underline{\Delta}_b\}$ guarantees, that for $\Delta \leq \underline{\Delta}_c$ and $\operatorname{TV}(\operatorname{sgn}(\hat{u})) < \operatorname{TV}(\operatorname{sgn}(u))$ the predicted reduction admits the lower bound:

$$\operatorname{pred}(u,\Delta) \geq 1,$$

where we used that $\mathrm{TV}(\mathrm{sgn}(u))$ can only attain even values. Note that $\mathrm{TV}(\mathrm{sgn}(\hat{u})) > \mathrm{TV}(\mathrm{sgn}(u))$ cannot occur for $\Delta \leq \underline{\Delta}_c$, since the prediction is always positive. By the second part of Lemma 4.22, this prediction is also guaranteed, if $\Delta \in [\underline{\Delta}_a(u), \underline{\Delta}_b]$.

This leaves the case $\mathrm{TV}(\mathrm{sgn}(\hat{u})) = \mathrm{TV}(\mathrm{sgn}(u))$ and $\Delta < \underline{\Delta}_a(u)$. We abbreviate $t \coloneqq \hat{t}_u$ and $s \coloneqq \hat{t}_{\hat{u}}$, which both belong to $\mathbb{R}^{2K}$ with $K \coloneqq \mathrm{TV}(\mathrm{sgn}(u))/2$. Since the trust-region radius directly limits the measure of the set $\{\mathrm{sgn}(u) \neq \mathrm{sgn}(\hat{u})\}$, for all $j = 1, \dots, 2K-1$ there exists a ξ_j, such that

$$\xi_j \in (t_j, t_{j+1}) \quad \wedge \quad \mathrm{sgn}(\hat{u}(\xi_j)) = \mathrm{sgn}(u(t_j+)) = \mathrm{sgn}(u(t_{j+1}-)),$$

since otherwise $\|\mathrm{sgn}(u) - \mathrm{sgn}(\hat{u})\|_{L^1(t_j,t_{j+1})} = t_{j+1} - t_j > \underline{\Delta}_a(u) > \Delta$. For completeness of the argument, it is beneficial to extend the control with zero past $[t_0, T]$ and consider additionally $\xi_0 < t_0, T < \xi_T$, such that we can use the same argument as above to conclude $s_j \in (\xi_{j-1}, \xi_j)$ for $j = 1, \dots, 2K$. With the above arguments it becomes clear, that the trust-region radius enforces $s_j \in [t_j - \Delta, t_j + \Delta]$, i.e., u and $\hat{u}$ possess the same switching patterns and the switching points of $\hat{u}$ are (small) perturbations of the switching points of u. Consequently, the signs of u and $\hat{u}$ only differ in the neighborhoods of the switching points and we get

$$\Delta \geq \|\mathrm{sgn}(u) - \mathrm{sgn}(\hat{u})\|_{L^1(t_0,T)} = \sum_{j=1}^{2K} |t_j - s_j|.$$

Before we can move to the final estimate, we need to investigate the parts of $C_{\mathrm{switch}}(u)$ with respect to Lipschitz continuity. Therefore, we recall the assumed Lipschitz continuities of ∇F:

$$\begin{aligned} \|\nabla F(u) - \nabla F(\hat{u})\|_{L^\infty(t_0,T)} &\leq L_{\nabla F} \|u - \hat{u}\|_{L^1(t_0,T)}, \\ |\nabla F(u)(s) - \nabla F(u)(t)| &\leq L_{\nabla F(u)} |s - t| \leq L_{\nabla F} |s - t|. \end{aligned}$$

Together with Lemma 4.21, this implies

$$\begin{aligned} |\nabla F(u)(t_j) - \nabla F(\hat{u})(s_j)| &\leq |\nabla F(u)(t_j) - \nabla F(\hat{u})(t_j)| + |\nabla F(\hat{u})(t_j) - \nabla F(\hat{u})(s_j)| \\ &\leq L_{\nabla F}(\|u - \hat{u}\|_{L^1(t_0,T)} + |t_j - s_j|) \leq L_{\nabla F}(\rho\Delta + \Delta). \end{aligned}$$

Using the notation of Definition 4.11, we get

$$\begin{aligned} V_u(t_j) &= \nabla F(u)(t_j) v_u(t_j) + g(v_u(t_j)) \\ &\leq \nabla F(u)(t_j) v_{\hat{u}}(s_j) + g(v_{\hat{u}}(s_j)) \\ &= (\nabla F(u)(t_j) - \nabla F(\hat{u})(s_j)) v_{\hat{u}}(s_j) + V_{\hat{u}}(s_j) \\ &\leq L_{\nabla F}(\rho + 1)\Delta b + V_{\hat{u}}(s_j) \end{aligned}$$

and vice cersa. Consequently,

$$|V_u(t_j) - V_{\hat{u}}(s_j)| \leq L_{\nabla F}(\rho + 1)\Delta b.$$

Using the definition of C_{switch}, this allows for the estimate

$$\begin{aligned} C_{\mathrm{switch}}(\hat{u}, s_j) &= \max\left\{ V_{\hat{u}}(s_j), -V_{\hat{u}}(s_j) \frac{\min\{s_j - t_0, T - s_j\}}{T - t_0} \right\} \\ &\leq C_{\mathrm{switch}}(u, t_j) + L_{\nabla F}(\rho + 1)\Delta b \end{aligned}$$

and vice versa switching the roles of u and $\hat{u}$. From this point we can make the final estimate

$$\begin{aligned} |C_{\mathrm{switch}}(u) - C_{\mathrm{switch}}(\hat{u})| &= \left| \max_{j=1,\dots,2K} C_{\mathrm{switch}}(u, t_j) - \max_{j=1,\dots,2K} C_{\mathrm{switch}}(\hat{u}, s_j) \right| \\ &\leq \max_{j=1,\dots,2K} \left| C_{\mathrm{switch}}(u, t_j) - C_{\mathrm{switch}}(\hat{u}, s_j) \right| \\ &\leq L_{\nabla F}(\rho + 1)\Delta b. \end{aligned}$$

This shows that Assumption 2.1(IV) is satisfied. □

Now, we are in position to prove the convergence theorem.

Theorem 4.24. *Algorithm 2.1 applied to* (OC) *either terminates at a stationary point or produces a sequence of iterates for which* $C(u_n) \to 0$.

If in addition $\nabla F\colon L^2(t_0,T;\mathbb{R}^N) \to L^2(t_0,T;\mathbb{R}^N)$ *is a compact operator,* (u_n) *contains a subsequence which converges strongly in* $L^2(t_0,T;\mathbb{R}^N)$ *towards a stationary point* $\bar u$ *and* $\operatorname{sgn}(u_n) \to \operatorname{sgn}(\bar u)$ *in* $L^1(t_0,T;\mathbb{R}^N)$.

Proof. Convergence of the criticality measure immediately follows from Theorem 2.2, since Lemmas 4.17, 4.19, 4.22 and 4.23 verified the four parts of Assumption 2.1.

Similar to the these lemmas, the idea of the remaining part of the proof is to separate the arguments relating to the proximal gradient method and the part relating to the switching points. We will carry out the proof for $N = 1$, i.e. $(u_k) \in L^\infty(t_0,T)$ for clarity of notation. It will become apparent from the individual arguments that they generalize to $N > 1$ without any changes.

To this end we consider the sequence $(\operatorname{sgn}(u_k))$, which is bounded in $\mathrm{BV}(t_0,T;\{0,1\})$, since the objective includes the total variation regularization. By [1, Theorem 3.23], there exists a subsequence (without relabeling) that is weak-$\star$ convergent towards a limit point $\bar\alpha$. Further by [1, Proposition 3.12], this subsequence $(\operatorname{sgn}(u_k))$ already convergences strongly in $L^1(t_0,T;\{0,1\})$. Thus, we can choose another subsequence (without relabeling) such that $\|\operatorname{sgn}(u_k) - \operatorname{sgn}(u_{k+1})\|_{L^1(t_0,T)} \le 2^{-k}$. Here, u_{k+1} is the element of the subsequence following u_k.

For the same the subsequence, $(u_k) \subset U_{\mathrm{ad}} \subset L^2(t_0,T)$ is bounded by the Assumption 4.1(ii), see also (4.1). Thus, there exists (without relabeling) a weakly convergent subsequence $u_k \rightharpoonup \bar u$. From Lemma 4.3, we get $\operatorname{sgn}(\bar u) = \bar\alpha$ almost everywhere.

Let $n \in \mathbb{N}$ be arbitrary. As a next step we want to split $\|u_k - \bar u\|_{L^2(t_0,T)}$ into the part where the sign no longer changes, i.e., which is supported on

$$\mathcal{I}_n \coloneqq \bigcap_{l=n}^{\infty} \{\operatorname{sgn}(u_l) = 1\},$$

and a vanishing remainder. From a straightforward argument, we get for all indices $k \ge n$ the inclusion

$$(\{\bar\alpha = 1\} \cup \{\operatorname{sgn}(u_k) = 1\}) \setminus \mathcal{I}_n \subset \left(\bigcup_{l=n}^{\infty} \{\operatorname{sgn}(u_l) = 1\} \triangle \{\operatorname{sgn}(u_{l+1}) = 1\} \right).$$

Since the above functions all have values in $\{0,1\}$ we get

$$\begin{aligned} |(\{\bar\alpha = 1\} \cup \{\operatorname{sgn}(u_k) = 1\}) \setminus \mathcal{I}_n| &\le \sum_{l=n}^{\infty} \|\operatorname{sgn}(u_l) - \operatorname{sgn}(u_{l+1})\|_{L^1(t_0,T)} \\ &\le \sum_{l=n}^{\infty} 2^{-l} \le 2^{-n+1}. \end{aligned} \tag{4.11}$$

We split the domain into $\mathcal{I}_n$ and its complement, i.e., for $k \ge n$ we get

$$\begin{aligned} \|u_k - \bar u\|^2_{L^2(t_0,T;\mathbb{R}^N)} &= \|u_k - \bar u\|^2_{L^2(\mathcal{I}_n)} + \|u_k - \bar u\|^2_{L^2(\{\operatorname{sgn}(u_k)=1\}\cup\{\bar\alpha=1\})\setminus\mathcal{I}_n)} \\ &\le \|u_k - \bar u\|^2_{L^2(\mathcal{I}_n)} \\ &\quad + \|u_k - \bar u\|^2_{L^\infty(t_0,T)} |\{\operatorname{sgn}(u_k) = 1\} \cup \{\bar\alpha = 1\}) \setminus \mathcal{I}_n)|. \end{aligned}$$

Since the sequence (u_k) only contains admissible points, we get $\|u_k - \bar u\|^2_{L^\infty(t_0,T)} \le b^2$. With (4.11) for any $\varepsilon \ge 0$ there exists a $n \in \mathbb{N}$, such that for all $k \ge n$ we have

$$\|u_k - \bar u\|^2_{L^2(t_0,T)} \le \|u_k - \bar u\|^2_{L^2(\mathcal{I}_n)} + \frac{\varepsilon}{2}. \tag{4.12}$$

In order to bound the first addend on the right-hand side, we utilize that u_k is always positive on $\mathcal{I}_n$ for $k \geq n$ and this means that the criticality measure $C(u_k)$ contains information on a proximal gradient step on $\mathcal{I}_n$, see Definition 4.11. We define $\hat{u}_k \coloneqq u_k - \operatorname{prox}_{rG_+}(u_k - r\nabla F(u_k))$. From the convergence of the criticality measure we get

$$\|\hat{u}_k\|^2_{L^2(\{\operatorname{sgn}(u_k)=1\})} = \|u_k - \operatorname{prox}_{rG}(u_k - r\nabla F(u_k))\|^2_{L^2(\{\operatorname{sgn}(u_k)=1\})} = 2r^2 C_{\text{prox}}(u_k) \to 0.$$

Consequently,

$$\|\hat{u}_k\|^2_{L^2(\mathcal{I}_n)} \to 0. \tag{4.13}$$

Since G_+ is an integral functional, the proximal map can be evaluated pointwise. In particular, this yields

$$-r\nabla F(u_k) + \hat{u}_k \in r\partial g_+(u_k - \hat{u}_k) \qquad \text{a.e. on } \mathcal{I}_n,$$

where g_+ is defined via

$$g_+(v) \coloneqq \begin{cases} g(v) & \text{if } a \leq v \leq b \text{ a.e. on } (t_0, T), \\ \infty & \text{else.} \end{cases}$$

Since g_+ is assumed to be strongly convex, see Assumption 4.1, its convex conjugate $g_+^\star$ has a Lipschitz continuous subdifferential $\partial g_+^\star$. Further, from $u_k \rightharpoonup \bar{u}$ in $L^2(t_0, T)$ and the assumed compactness of ∇F, we get $\nabla F(u_k) \to \nabla F(\bar{u})$ in $L^2(t_0, T)$. Combining these two facts with $\partial g_+^\star = (\partial g_+)^{-1}$ and the strong convergence of $\hat{u}_k$ on $\mathcal{I}_n$ due to (4.13), we get

$$\bar{u} \leftharpoonup u_k = \hat{u}_k + \partial g_+^\star\left(\frac{\hat{u}_k}{r} - \nabla F(u_k)\right) \to \partial g_+^\star(-\nabla F(\bar{u})) \qquad \text{in } L^2(\mathcal{I}_n). \tag{4.14}$$

Hence, there exists $K \in \mathbb{N}$, $K \geq n$, such that

$$\|u_k - \bar{u}\|_{L^2(\mathcal{I}_n)} \leq \frac{\varepsilon}{2} \qquad \forall k \geq K$$

and with (4.12) we get

$$\|u_k - \bar{u}\|_{L^2(t_0,T)} \leq \varepsilon \qquad \forall k \geq K.$$

This shows $u_k \to \bar{u}$ in $L^2(t_0, T)$. It remains to show that $\bar{u}$ is stationary, i.e., $C(\bar{u}) = 0$.

Regarding C_{prox}, we note that (4.14) implies $\bar{u} = \operatorname{prox}_{rg_+}(\bar{u} - r\nabla F(\bar{u}))$ on $\mathcal{I}_n$ for all $n \in \mathbb{N}$. From (4.11), we get $|\{\bar{\alpha} = 1\} \setminus \mathcal{I}_n| \leq 2^{-n+1}$ and, thus, $\{\bar{\alpha} = 1\} = \{\operatorname{sgn}(\bar{u}) = 1\}$ coincides with $\bigcup_{n=1}^\infty \mathcal{I}_n$ (up to a set of measure zero). Consequently, $\bar{u} = \operatorname{prox}_{rg_+}(\bar{u} - r\nabla F(\bar{u}))$ a.e. on $\{\operatorname{sgn}(\bar{u}) = 1\}$. Hence, $C_{\text{prox}}(\bar{u}) = 0$.

To show the stationarity of the switching points, we fix a switching point $s \in J_{\bar{u}}$ of $\bar{u}$. From the convergence $u_k \to \bar{u}$ in $L^1(t_0, T)$, we infer the existence of a switching point $s_k \in J_{u_k}$ of u_k with $|s - s_k| \to 0$. From $C(u_k) \to 0$, we get $C_{\text{switch}}(u_k, 1, s_k) \to 0$. The continuity properties of the gradient ∇F from Assumption 4.1 ensure $\nabla F(u_k)(s_k) \to \nabla F(\bar{u})(s)$ and, consequently, $V_{u_k,1}(s_k) \to V_{\bar{u},1}(s)$. Altogether, this implies $C_{\text{switch}}(\bar{u}, 1, s) = 0$. Since s was an arbitrary switching point, we get $C_{\text{switch}}(\bar{u}) = 0$. Thus, $C(\bar{u}) = 0$ and $\bar{u}$ is stationary. □

4.4 FAST SOLUTION OF SUBPROBLEMS

It remains to give a possibility for the computation of the oracle function. We use the approach from [21, Section 5] to solve the corresponding trust-region subproblems efficiently using the Bellman principle.

We discretize the problem by choosing an equidistant discretization with step length $\tau = (T - t_0)/N_T$ and the controls are discretized by piecewise constant functions. Thus, the controls are identified with matrices $u \in \mathbb{R}^{N \times N_T}$, i.e., the entry $u_{i,j}$ corresponds the value of component i on the time interval

$[t_{j-1}, t_j)$. Consequently, the gradient $\nabla F(u)$ can be represented by a matrix in $\mathbb{R}^{N\times N_T}$, i.e., $\nabla F(u)_{i,j}$ is the partial derivative of F at u w.r.t. $u_{i,j}$ divided by τ. In each iteration of the trust-region method, we have to minimize a discretization of the model problem (4.6) (dropping the constant terms which do not depend on w), i.e.,

$$m_{u,\Delta}(w) \coloneqq \tau \sum_{j=1}^{N_T}\sum_{i=1}^{N}\left(\nabla F(u)_{i,j} w_{i,j} + g_i(w_{i,j}) + \frac{\operatorname{sgn}(u_{i,j})\operatorname{sgn}(w_{i,j})}{2\Delta}|u_{i,j}-w_{i,j}|^2\right) + \sum_{j=0}^{N_T}\sum_{i=1}^{N}|\operatorname{sgn}(w_{i,j}) - \operatorname{sgn}(w_{i,j+1})|.$$

Note that the third term in parenthesis is the discretization of $(2\Delta)^{-1}\|u-w\|^2_{L^2(\{\operatorname{sgn}(uw)=1\})}$.

In order to account for the boundary terms in the TV regularization, we use the convention $\operatorname{sgn}(w_{i,0}) = \operatorname{sgn}(w_{i,N_T+1}) = 0$. The components of w are coupled due to the trust-region constraint $\sum_{i=1}^{N}\sum_{j=1}^{N_T}|\operatorname{sgn}(u_{i,j}) - \operatorname{sgn}(w_{i,j})| \le \mathcal{B}$. The unconventional symbol choice for this trust-region radius comes from the fact that this value can be understood as a "budget" for the amount of change regarding the switching pattern in the application of Bellman's principle below. Originally, we would have $\mathcal{B} = \Delta/\tau$, but we will see that it is beneficial to treat the proximal parameter Δ and the trust-region radius $\mathcal{B}$ appearing in the constraint differently. The oracle $\hat{u} = T(u,\Delta,\mathcal{B})$ is a minimizer of the model over the trust region.

The main idea will be to split the model problem and apply the celebrated Bellman principle. In what follows, we use the abbreviation $w_j \coloneqq (w_{1,j},\dots,w_{N,j})$ for the variable at time step j. To this end we define for $1 \le l \le r \le N_T$ the family of partial model problems via

$$m^{(l,r)}_{u,\Delta}(w) \coloneqq \tau \sum_{j=l}^{r}\sum_{i=1}^{N}\left(\nabla F(u)_{i,j} w_{i,j} + g_i(w_{i,j}) + \frac{\operatorname{sgn}(u_{i,j})\operatorname{sgn}(w_{i,j})}{2\Delta}|u_{i,j}-w_{i,j}|^2\right) + \sum_{j=l}^{r-1}\|\operatorname{sgn}(w_j) - \operatorname{sgn}(w_{j+1})\|_1.$$

Note that $m^{(1,N_T)}_{u,\Delta}(w)$ does not contain the jump terms at the boundary, i.e.,

$$m_{u,\Delta}(w) = m^{(1,N_T)}_{u,\Delta}(w) + \|\operatorname{sgn}(w_1)\|_1 + \|\operatorname{sgn}(w_{N_T})\|_1.$$

By optimizing a part of the model function only w.r.t. $w_1,\dots,w_l$ gives rise to the definition of the value function

$$\Phi(l,\alpha,B) \coloneqq \min\left\{ m^{(1,l)}_{u,\Delta}(w) + \|\operatorname{sgn}(w_1)\|_1 \;\middle|\; \begin{array}{l} w_1,\dots,w_l \in \bar{M}_0,\\ \operatorname{sgn}(w_l) = \alpha,\\ \displaystyle\sum_{j=1}^{l}\|\operatorname{sgn}(u_j) - \operatorname{sgn}(w_j))\|_1 = B \end{array} \right\}$$

for $l \in \{1,\dots,N_T\}$, $\alpha \in \{0,1\}^N$, $B \in \{0,\dots,\mathcal{B}\}$. Note that α encodes the terminal value of $\operatorname{sgn}(w_l)$ and this will be important for the recursion formula below. Now, we can rewrite the value of the model problem at the minimizer $\hat{u} = T(u,\Delta,\mathcal{B})$ via

$$m_{u,\Delta}(\hat{u}) = \min\left\{\Phi(N_T,\alpha,B) + \|\alpha\|_1 \;\middle|\; \alpha \in \{0,1\}^N,\ B \in \{0,\dots,\mathcal{B}\}\right\},$$

where $\|\alpha\|_1$ is the cost for a jump at the terminal time T. Now, we derive a recursion formula for the value function. For $1 \le l < N_T$ we have

$$\Phi(l+1,\alpha,B)$$

$$= \min\left\{ \begin{array}{l} m_{u,\Delta}^{(1,l)}(w) + m_{u,\Delta}^{(l+1,l+1)}(w) \\ + \|\operatorname{sgn}(w_1)\|_1 \\ + \|\operatorname{sgn}(w_l) - \alpha\|_1 \end{array} \middle| \begin{array}{l} w_1, \dots, w_l, w_{l+1} \in \bar{M}_0, \\ \operatorname{sgn}(w_{l+1}) = \alpha, \\ \sum_{j=1}^{l+1} \|\operatorname{sgn}(u_j) - \operatorname{sgn}(w_j)\|_1 = B \end{array} \right\}$$

$$= \min\left\{ \begin{array}{l} m_{u,\Delta}^{(1,l)}(w) + m_{u,\Delta}^{(l+1,l+1)}(w) \\ + \|\operatorname{sgn}(w_1)\|_1 \\ + \|\beta - \alpha\|_1 \end{array} \middle| \begin{array}{l} w_1, \dots, w_l, w_{l+1} \in \bar{M}_0, \\ \beta \in \{0,1\}^N,\ \tilde{B} \in \{0, \dots, B\}, \\ \operatorname{sgn}(w_{l+1}) = \alpha,\ \operatorname{sgn}(w_l) = \beta \\ \|\operatorname{sgn}(u_{l+1}) - \alpha\|_1 = B - \tilde{B} \\ \sum_{j=1}^{l} \|\operatorname{sgn}(u_j) - \operatorname{sgn}(w_j)\|_1 = \tilde{B} \end{array} \right\}$$

$$= \min\left\{ \begin{array}{l} \Phi(l, \beta, \tilde{B}) + m_{u,\Delta}^{(l+1,l+1)}(w) \\ + \|\beta - \alpha\|_1 \end{array} \middle| \begin{array}{l} w_{l+1} \in \bar{M}_0,\ \beta \in \{0,1\}^N,\ \tilde{B} \in \{0, \dots, B\}, \\ \operatorname{sgn}(w_{l+1}) = \alpha,\ \|\operatorname{sgn}(u_{l+1}) - \alpha\|_1 = B - \tilde{B} \end{array} \right\}.$$

In the first equality, we have rewritten the objective, in the second equality we introduced the additional optimization variables β and $\tilde{B}$, which allows us to carry out the optimization w.r.t. $w_1, \dots, w_l$ in the third equality explicitly. The remaining optimization w.r.t. w_{l+1} can be carried out explicitly and the minimizer (which also depends on α_i) is given by

$$w_{i,l+1}^{\alpha_i} := \begin{cases} 0 & \text{if } \alpha_i = 0, \\ \operatorname{prox}_{\Delta g_{i,+}}\big(u_{i,j} - \Delta \nabla F(u)_{i,j}\big) & \text{if } \alpha_i = 1 \text{ and } \operatorname{sgn}(u_{i,j}) = 1, \\ \arg\min_{v \in [a_i, b_i]} \big(\nabla F(u)_{i,j} v + g_{i,+}(v)\big) & \text{if } \alpha_i = 1 \text{ and } \operatorname{sgn}(u_{i,j}) = 0. \end{cases} \tag{4.15}$$

The initial value of the value function can be computed explicitly as

$$\Phi(1, \alpha, B) = \min\left\{ m_{u,\Delta}^{(1,1)}(w) + \|\alpha\|_1 \quad \middle| \begin{array}{l} w_1 \in \bar{M}_0, \\ \operatorname{sgn}(w_1) = \alpha,\ \|\operatorname{sgn}(u_1) - \alpha)\|_1 = B \end{array} \right\}.$$

Again, this minimization w.r.t. w_1 can be evaluated as in (4.15).

Finally, we mention how the minimizer $\hat{u} = T(u, \Delta)$ of the model problem can be evaluated. For this, we store in $\Psi(l+1, \alpha, B)$ one of the minimizing values of β appearing in the recursion formula for $\Phi(l+1, \alpha, B)$. Then, the optimal switching vector $\hat{\alpha}$ can be evaluated via

$$(\hat{\alpha}_{N_T}, \hat{B}_{N_T}) = \arg\min\big\{\Phi(N_T, \alpha, B) + \|\alpha\|_1 \,\big|\, \alpha \in \{0,1\}^N,\ B \in \{0, \dots, \mathcal{B}\}\big\} \tag{4.16a}$$

and

$$\hat{\alpha}_l = \Psi(l+1, \hat{\alpha}_{l+1}, \hat{B}_{l+1}), \qquad \hat{B}_l = \hat{B}_{l+1} - \|\operatorname{sgn}(u_{l+1}) - \hat{\alpha}_{l+1}\|_1 \tag{4.16b}$$

for $l = N_T - 1, \dots, 1$. Finally, the minimizer $\hat{u}$ is given by

$$\hat{u}_{i,j} = w_{i,j}^{\hat{\alpha}_{i,j}} \tag{4.16c}$$

by using (4.15).

In the definition of the oracle $T(u, \Delta)$, see Definition 4.15, the value Δ is used both as a trust-region radius and as a parameter scaling the proximal term. In the discretized setting, this corresponds to the fixed choice $\mathcal{B} = \Delta/\tau$. If we use this fixed scaling in Algorithm 2.1, the algorithm barely makes any

progress, since the trust-region radius and the proximal-point parameter should scale very differently. Consequently, we separated both parameters.

The utilization of the value function and the separation of Δ and $\mathcal{B}$ have a further advantage. Once we have computed Φ and Ψ, we can optimize the model $m_{u,\Delta}$ over a smaller trust region with radius $\hat{\mathcal{B}} \leq \mathcal{B}$. That is, the minimizer $\hat{u}$ of

$$\min\left\{ m_{u,\Delta}(w) \;\middle|\; \sum_{j=1}^{N_T} \|\operatorname{sgn}(u_j) - \operatorname{sgn}(w_j)\|_1 \leq \hat{\mathcal{B}} \right\}$$

can be obtained by replacing $\mathcal{B}$ by $\hat{\mathcal{B}}$ in (4.16). Note that this is computationally cheap and does not require the expensive recomputation of Φ and Ψ. The separation of Δ and $\mathcal{B}$ and the possible fast reevaluation for smaller trust-region radii $\mathcal{B}$ give rise to Algorithm 4.1 which is a variation of Algorithm 2.1. Note that the convergence of Algorithm 4.1 is not covered directly by the convergence analysis of Theorem 4.24. The investigation of convergence guarantees of Algorithm 4.1 is subject to future research.

Algorithm 4.1: TRM for (OC)

Data: $\eta \in (0,1)$, $0 < \gamma_1 < 1 \leq \gamma_2$, $0 < \Delta_0 < \Delta_{\max} \in (0,\infty]$, $\mathcal{B}_0 \coloneqq \mathcal{B}_{\max} \in \mathbb{N}$, $u_0 \in U_{\mathrm{ad}}$

$n \coloneqq 0$
while $C(u_n) > 0$ **do**
  if $\operatorname{ared}(u_n, \Delta_n, \mathcal{B}_n) \geq \eta \ \operatorname{pred}(u_n, \Delta_n, \mathcal{B}_n)$ **then**
    $u_{n+1} \coloneqq T(u_n, \Delta_n, \mathcal{B}_n)$
    $\Delta_{n+1} \coloneqq \min\{\gamma_2 \Delta_n, \Delta_{\max}\}$
    $\mathcal{B}_{n+1} \coloneqq \min\{\lceil \gamma_2 \mathcal{B}_n \rceil + 1, \mathcal{B}_{\max}\}$
  else
    $u_{n+1} \coloneqq u_n$
    if $\mathcal{B}_n \coloneqq 0$ **then**
      $\Delta_{n+1} \coloneqq \gamma_1 \Delta_n$
      $\mathcal{B}_{n+1} \coloneqq \mathcal{B}_{\max}$
    else
      $\mathcal{B}_{n+1} \coloneqq \lfloor \gamma_1 \mathcal{B}_n \rfloor$
      $\Delta_{n+1} \coloneqq \Delta_n$
  $n \coloneqq n + 1$

5 NUMERICAL EXPERIMENTS

In this section, we solve (discretizations of) (OC) with Algorithm 4.1 by using the approach of Section 4.4 to solve the trust-region subproblems. We intend to highlight characteristic properties, as well as shortcomings of the algorithm. We will consider examples that come with a qualitative understanding of the resulting control, especially its switching behavior. In both examples, the parameters used in Algorithm 4.1 are

$$\gamma_1 = \frac{1}{2}, \qquad \gamma_2 = 2, \qquad \Delta_0 = 10^{-7}, \qquad \Delta_{\max} = 10, \qquad \eta = 10^{-3}.$$

The termination criterion "$C(u_n) = 0$" in Algorithm 4.1 is replaced by

$$C_{\mathrm{prox}}(u_n) \leq 10^{-10} \qquad \text{and} \qquad \mathcal{B}_n = 0.$$

The first condition ensures that the continuous parts of the control are almost stationary. The second condition ensures that the algorithm did not found reasonable steps and the budget $\mathcal{B}_n$ for moving the switching points has been reduced to 0. Note that $C_{\text{switch}}(u_n)$ will not converge to zero after discretization and this will be discussed below.

5.1 DECAY PROBLEM

As a first example, we consider an identification problem in which the terms of the objective are defined via

$$\begin{aligned} F(u) &:= \frac{\sigma_y}{2} \int_{t_0}^{T} \left(\frac{y_u(t) - y_d(t)}{y_d(t)} \right)^2 \mathrm{d}t + \frac{\sigma_T}{2} (y_u(T) - y_d(T))^2, \\ G(u) &:= \int_{t_0}^{T} g(u(t)) \, \mathrm{d}t, \\ J(u) &:= F(u) + G(u) + \sigma_{\text{sgn}} \mathrm{TV}(\operatorname{sgn}(u)) \end{aligned} \tag{5.1}$$

with parameters $t_0 = 0$, $T = 140$, $\sigma_y = 10$, $\sigma_T = 0.3$, $\sigma_{\text{sgn}} = 1$, and

$$g(x) = \begin{cases} 0 & \text{if } x = 0, \\ 0.7x^2 - 0.5x + 0.4 & \text{if } x \in [0.3, 1], \\ \infty & \text{else,} \end{cases}$$

i.e., $a = 0.3$ and $b = 1$. Further, the state y_u is the solution of the ordinary differential equation

$$\dot{y}_u(t) = -(0.025 + 0.05u(t)) y_u(t), \qquad y_u(t_0) = 1000.$$

Note that this differential equation describes an exponential decay of the state y_u and the control can be used to increase the decay rate from 0.025 to a value from $[0.04, 0.075]$. The choice of parameters comes from balancing the order of magnitude of the different terms in the objective, but have no further meaning. We use Algorithm 4.1 to solve discretized versions of this problem.

The target state y_d is computed using the control shown in Figure 5.1(a). Due to the structure of the state equation, y_d is strictly positive. The running cost of the active control is priced with the convex function g, see Figure 5.1(b). The tangent through the origin (shown in orange) gives the prognosis for the control value $u^\star$ at which switches are allowed, see Remark 4.9. The problem is solved on a discretization with 8192 time steps. Figure 5.1(c) depicts the computed solution. The dotted vertical lines show the switches of the control. The dotted horizontal line visualizes the precomputed control value $u^\star$ (see Remark 4.9) at which a jump may occur. Note the exception at t_0, where the optimality condition only yields that the control value is above the switching value. Figure 5.1(d) shows the pointwise relative error between target state and computed state. Note that the plot does not show jumps, but steep slopes between the points, where the computed control and the control for the computation of y_d have slightly shifted switching points. Figure 5.1(e) visualizes the objective over the iterations of the algorithm compared to the objective value obtained in the final iteration and Figure 5.1(f) shows both parts of the criticality measure. In Table 5.1 we present some numbers concerning the solution for different time discretization levels N_T and the choice $\mathcal{B}_{\max} = \max\{8, \lfloor N_T/16 \rfloor\}$.

From this example it should be noted, that the predicted switching value is attained in the solution and that the switching behavior mimics that of the control which was used to compute the target state y_d. The target state is identified well. The criticality measure C_{prox} is of magnitude 10^{-10} (which means that the algorithm terminated successfully), while C_{switch} stagnates at a rather high level (depending on the time discretization), which at first glance may not be the expected convergence. Note that we derived the optimality conditions Theorem 4.7 only for the non-discretized problem. The proof of

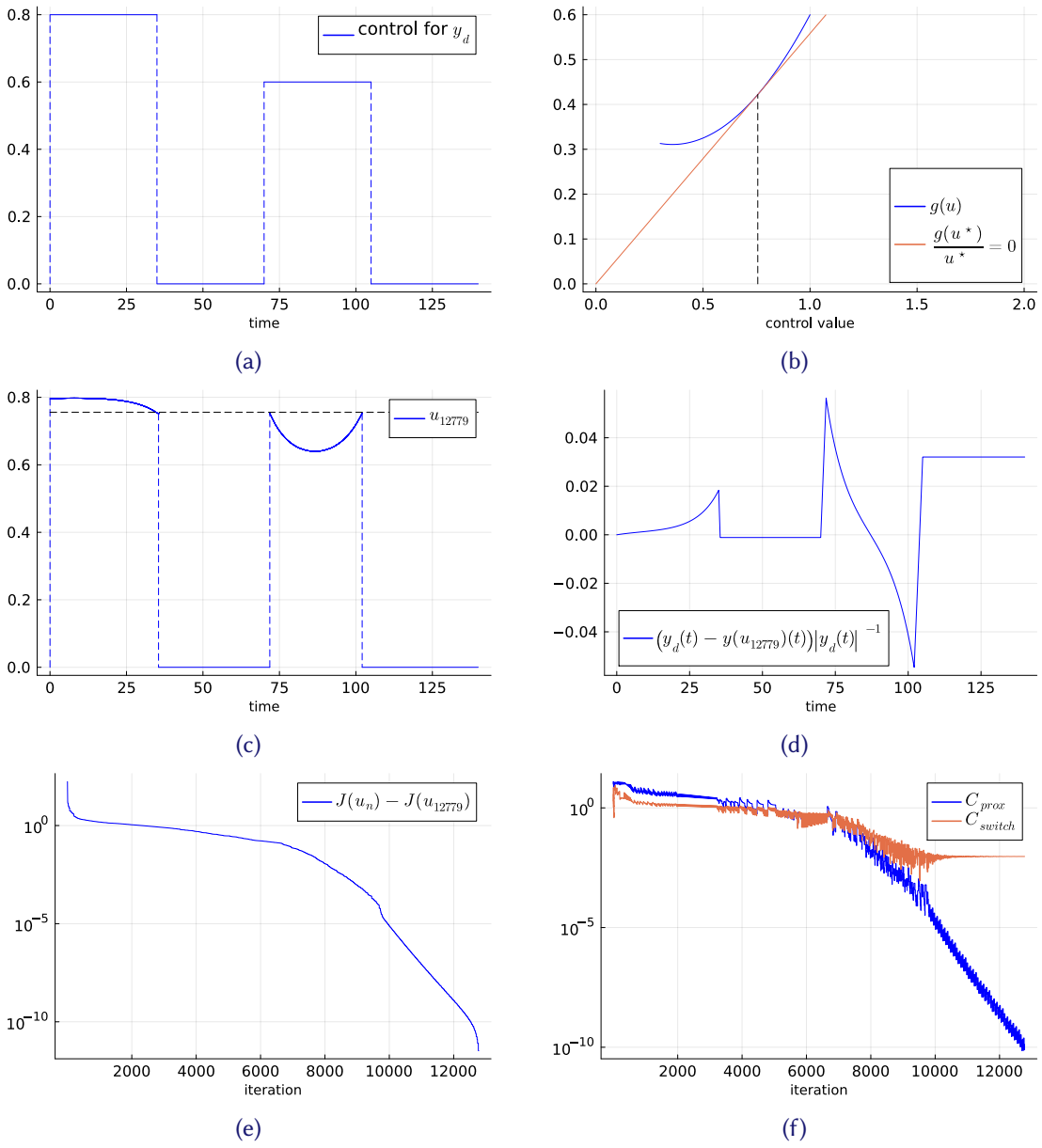


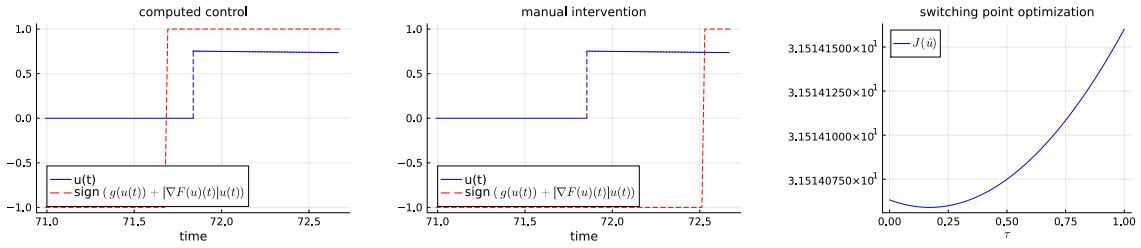
Figure 5.1: Identification of the decay problem.

condition (4.5a) (corresponding to C_{prox}) can be easily carried over to the discretized setting, but the proofs of (4.5b)–(4.5d) cannot be transferred since they rely on small shifts of the switching points, which is not possible after time discretization. In Table 5.1, we can see that the final value of C_{switch} decreases with N_T.

Further visualization of this issue is given in Figure 5.2, where we zoomed in around the switching point at $t \approx 71.75$. The computed switching point can be seen in Figure 5.2(a) together with the signum of $g(u(t)) + [\nabla F(u)(t)]u(t)$. The optimality condition suggests, that the sign changes of both functions occur at the same time point. The value of $g(u(t)) + [\nabla F(u)(t)]u(t)$ suggests moving the switch to a later point in time. Figure 5.2(b) shows the result of manually moving the switch one time step τ to the right. We observe that the manual correction shifts the sign change, such that the algorithm in its next iteration would reverse our intervention. In this sense the switching point is indeed locally optimal even if it does not perfectly align with the exact position of the sign change of the control and the

N_T	$\mathcal{B}_{\max}$	iterations	time in s	J	C_{prox}	C_{switch}
32	8	4499	0.44	31.4678	9.80040e-11	0.785092
64	8	3496	0.57	32.1746	9.81943e-11	0.728040
128	8	3115	0.88	31.8243	3.91882e-11	0.549244
256	16	4179	2.67	31.5575	9.83474e-11	0.320774
512	32	4671	6.04	31.4768	9.03884e-11	0.122810
1024	64	5168	20.04	31.4857	8.43537e-11	0.062130
2048	128	6881	68.46	31.4955	8.73450e-11	0.025920
4096	256	9519	273.32	31.5029	6.94831e-11	0.014348
8192	512	12779	1157.26	31.5069	7.76904e-11	0.009325
16384	1024	15133	4568.48	31.5089	9.46679e-11	0.003030

Table 5.1: Computations with different discretizations for the decay problem



(a) (b) (c)

Figure 5.2: Identification of the decay problem.

criticality measure retains a rather high value. Further, Figure 5.2(c) shows a gradual movement of the switch between the original grid points, where 0 represents no change and 1 pushing it a full time step. This suggests a discretization refinement around the switching points would benefit the algorithm, but a global refinement of the discretization is not feasible looking at the scaling behavior in Table 5.1.

5.2 SIR PROBLEM

For the second example, we want to consider a more practical example with multiple controls u controlling a simple SIR model. The underlying system of ordinary differential equations is given for a population with a fixed number of N individuals divided into "susceptible" (S), "infected" (I) and "recovered" (R) individuals, where the infection is driven by a contact term with time-dependent infection rate $\beta(t)$ (see below) and recovery occurs with rate ρ. This results in the system

$$\text{(SIR)} \qquad \begin{aligned} \dot{S}(t) &= -\beta(t)S(t)I(t)/N, \\ \dot{I}(t) &= \beta(t)S(t)I(t)/N - \rho I(t), \\ \dot{R}(t) &= \rho I(t). \end{aligned}$$

When planning to contain the outbreak of a pandemic, one possible strategy is to achieve sufficient suppression in order to wait for a fundamental change of circumstances, like a new treatment or vaccination. A second approach is to obtain a controlled infection while "flattening the curve". We try to model the latter via the objective

$$F(I,S) := \frac{\sigma_I}{2}\|I\|_{L^2}^2 + \frac{\sigma_S}{2}S(T)^2,$$

where the first term penalizes the severity of the outbreak and the second term penalizes the remaining number of susceptibles at the end of the considered time horizon. The control modifies the infection parameter, such that $\beta(t) = \beta_0(1-u_1(t)-u_2(t))$ and the controls are priced with functions representing a cheap measure

$$g_1(x) = \begin{cases} 0 & \text{if } x = 0, \\ 30x^2 + x + 1 & \text{if } x \in [0.1, 0.6], \\ \infty & \text{else,} \end{cases}$$

and an expensive measure

$$g_2(x) = \begin{cases} 0 & \text{if } x = 0, \\ 30000x^2 + x + 3000 & \text{if } x \in [0.1, 0.4], \\ \infty & \text{else.} \end{cases}$$

The stark contrast is intentionally chosen to highlight shortcomings of the algorithm and at the same time allows for some intuition on the structure of the solution. The full objective is given by

$$J(u) \coloneqq F(I(u), S(u)) + \sum_{i=1}^{2} \left(\int_{t_0}^{T} g_i(u_i(t)) \, \mathrm{d}t \right) + \sigma_{\text{sgn}} \mathrm{TV}(\mathrm{sgn}(u)).$$

The chosen parameters are

$$\sigma_I = 2, \ \sigma_S = 800, \ \sigma_{\text{sgn}} = 10000, \ \rho = 0.1, \ \beta_0 = 0.6, \ t_0 = 0, \ T = 140.$$

The solution for a discretization with 8192 time steps is shown in Figure 5.3. From a qualitative standpoint, the solution is reasonable. The suggested control manages a controlled infection with a flattened curve, while the cheap measure is used to its fullest capacity and the expensive measure is balanced with its cost, while the amount of switches are minimized, due to the high switching cost. On closer inspection, we see an amplified version of the detailed problem regarding the convergence of C_{switch}. For the cheap control the switching value (visualized by the dotted line) seems to be ignored, while the expensive measure does not exactly reach its switching value either. This suggests that the current time discretization is too coarse to resolve these features of the optimal control. In Table 5.2,

N_T	$\mathcal{B}_{\max}$	iterations	time in s	J	C_{prox}	C_{switch}
32	8	3711	1.01	1.43464e6	8.1725e0-11	21168.2
64	8	3103	1.67	1.34994e6	9.96217e-11	24282.7
128	8	1087	0.96	1.36792e6	6.42651e-11	38060.6
256	16	4110	6.03	1.12967e6	9.84629e-11	8838.4
512	32	8016	31.40	1.02804e6	9.57865e-11	3777.2
1024	64	4819	43.77	9.48856e5	7.17141e-11	2010.9
2048	128	6689	152.28	8.89106e5	9.08372e-11	1086.8
4096	256	7008	483.05	8.39067e5	8.54805e-11	1034.9
8192	512	11402	2488.06	7.92406e5	5.96074e-11	716.4

Table 5.2: Computations with different discretizations for the SIR problem

we give further results on different time discretization levels. Similarly to the decay problem, we see that C_{switch} decreases as the time discretization parameter N_T increases.

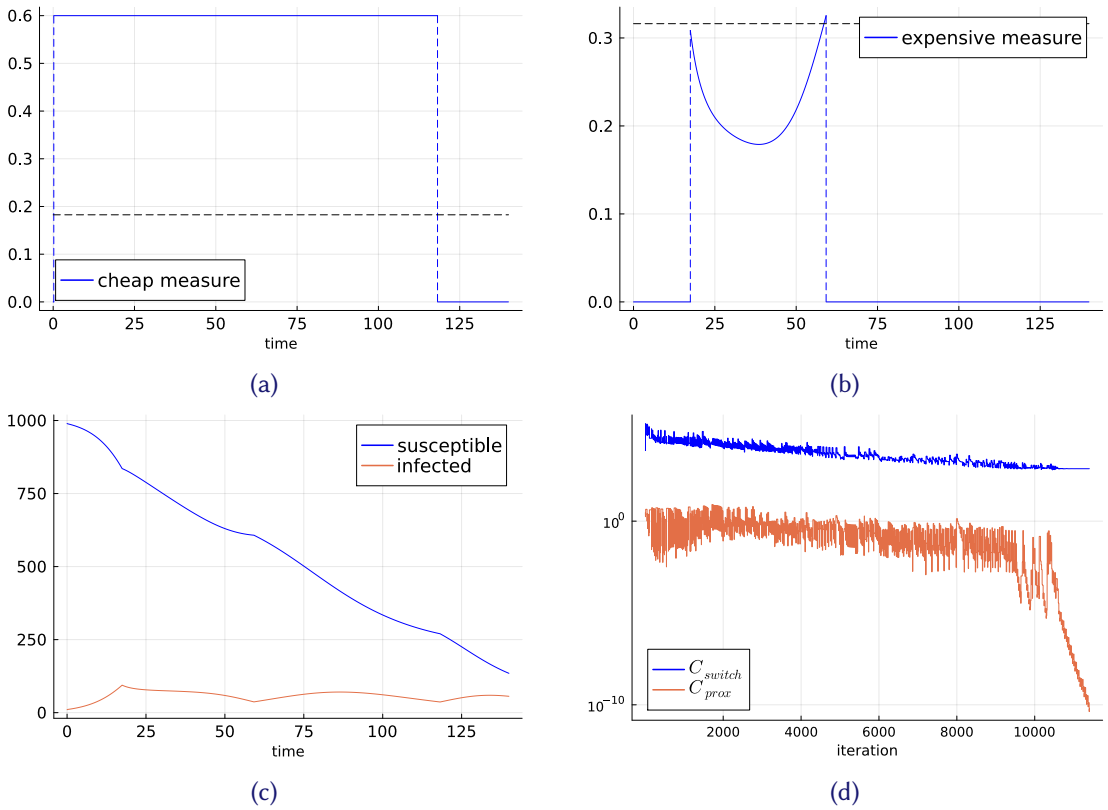


Figure 5.3: Control of an SIR dynamic

6 OUTLOOK

In this section we present some interesting questions and ideas that arise from our work.

SWITCHING POINT OPTIMIZATION BETWEEN GRID POINTS

We have seen in Section 5, that we do not have the desired convergence with respect to the part of the criticality measure C_{switch} that relates to the optimality of the switching points. At the same time, we showed for the example problem that the method indeed numerically converges given the discretization. This suggests to refine the discretization, but the algorithm presented in Section 4.4 that is used to solve the model problems is expensive. This issue occurs even if $a_i = b_i$.

We suggest to investigate a post processing step in form of a switching point optimization that operates between grid points. This could be realized by changing the Euler step from

$$y_{j+1} = y_j + (t_{j+1} - t_j) f(t_j, y_j, u_{j+1})$$

to

$$y_{j+1} = y_j + h(t_{j+1} - t_j) f(t_j, y_j, u_{j+1})$$

with $h \in [0, 1]$. The new optimization variable h describes a switch strictly between the time points t_j and t_{j+1}. This essentially means, that intermediate values can be expressed as a simple proportional scaling when applying the forward Euler method to solve the state and adjoint equations required for the computation of $F(u)$ and $\nabla F(u)$. Utilizing this principle, a gradient descent method can be deployed, that moves switches between time points by introducing the appropriate scaling. The main benefit is that the computed gradient remains numerically exact. In first experiments, we could confirm

convergence with respect to C_{switch} for (SIR) and even the features of the cheap pricing function were correctly resolved.

The second benefit is that this approach yields a proper termination criterion which can significantly shorten the computational time, since one does not have to use the expensive inner solver until no further step is taken.

GENERALIZATIONS OF THE CONTROL STRUCTURE

The structure of the objective was mainly motivated by control problems arising for pandemic models, where the decision between activating or deactivating a measure with its associated switching cost is highly important. The proposed structure has two natural extensions.

First would be the inclusion of a time dependence of g_i, a_i and b_i, where we suspect that one then needs at least Lipschitz continuity with respect to the time variable. We have to emphasize that the proofs do not extend trivially to this case.

The second question would be regarding multiple activity sets $u_i(t) \in [a_i^1, b_i^1] \cup [a_i^2, b_i^2] \cup \cdots \cup [a_i^K, b_i^K]$ with $a_i^1 \leq b_i^1 < a_i^2 \leq b_i^2 < \cdots < a_i^K \leq b_i^K$ and a cost function, where we only assume strong convexity for each $g_i|_{[a_i^k, b_i^k]}$. Note that this is related to the concept of multibang control, see, e.g., [9, 10]. We expect that the algorithm extends naturally to this case, while it is less obvious what to expect for the jumping value $u^\star$. We conjecture that a jump between two activity sets can only occur from and to the control value where $g_i|_{[a_i^k, b_i^k]}$ and $g_i|_{[a_i^l, b_i^l]}$ share the same tangent and only if this tangent lies below $g_i|_{[a_i^j, b_i^j]}$ for all $i < j < k$.

COMBINATION OF MULTIPLE ORACLES

Lastly, Section 4 poses a question regarding the abstract trust-region method. Large parts of the proofs involved splitting the argument into the proximal gradient part and the part relating to the switching point optimization. However, the proofs could not be carried out in full separation, especially the verification of the stability assumption Assumption 2.1(IV).

It would be interesting to investigate under which conditions for the underlying problem two separate oracles, which individually satisfy Assumption 2.1, can be combined. We suggest to focus on a combined criticality measure which is simply the maximum of the two individual components.

We suspect, that a strong assumption of interdependence is required with respect to the Lipschitz continuity, if both oracles are evaluated in each step and share the same trust-region radius. In contrast, we expect significantly less strict assumptions if the choice of the deployed oracle is made dependent on the currently dominating part of the criticality measure.

It is highly likely, that the convergence of Algorithm 4.1 can be proven by combining our results for the proximal gradient part in Section 3 directly with the results from [18, Section 5] utilizing such an abstract combination result for separate oracle functions.

APPENDIX A PROOF OF THEOREM 2.2

We prepare the proof of Theorem 2.2 with a simple lemma.

Lemma A.1. *Let Assumption 2.1 be valid. There exists a nondecreasing function $\underline{\Delta}_{\text{pred}} \colon (0, \infty) \to (0, \infty)$ such that for all $x \in X$ with $C(x) > 0$ and trust-region radii $\Delta \in (0, \underline{\Delta}_a(x))$ with $\Delta \leq \underline{\Delta}_{\text{pred}}(C(x))$ we have*

1. *(sufficient decrease):*

$$\operatorname{pred}(x, \Delta) \geq \frac{f\left(C(x)\right)}{2} \Delta,$$

i.e., for every nonstationary x there exists a trust-region radius that guarantees a positive prediction,

2. *(accuracy):*

$$\operatorname{ared}(x,\Delta) \geq \eta \operatorname{pred}(x,\Delta).$$

In particular, if $\hat{x} = T(x,\Delta)$ is not accepted in Algorithm 2.1, we get a lower bound for the trust-region radius $\Delta \geq \min\{\underline{\Delta}_{\text{pred}}(C(x)), \underline{\Delta}_a(x)\}$.

Proof. From straightforward calculations, we obtain

$$f(C(x))\Delta - c\Delta^s \geq \frac{1}{2} f(C(x))\Delta \qquad \forall \Delta \in \left[0, \left(\frac{f(C(x))}{2c}\right)^{1/(s-1)}\right].$$

Thus, the assertion follows from Assumption 2.1(II) and (III) with

$$\underline{\Delta}_{\text{pred}}(\zeta) \coloneqq \left(\frac{f(\zeta)}{2c}\right)^{1/(s-1)} \qquad \forall \zeta > 0.$$

□

Now we are in position to prove Theorem 2.2.

Proof of Theorem 2.2. Algorithm 2.1 terminates at x_n if and only if $C(x_n) = 0$. For the remainder of the proof, we assume that the sequence of iterates (x_n) is infinite.

First, we argue that for each x_n any trust-region radius below a certain threshold leads to an accepted step (for clarification of the term compare with Remark 2.3). If $\Delta_n < \underline{\Delta}_a(x_n)$ then Lemma A.1 yields the threshold $\underline{\Delta}_{\text{pred}}(C(x_n))$. If $\Delta_n \geq \underline{\Delta}_a(x_n)$ the threshold is $\underline{\Delta}_b$ by Assumption 2.1(III.E), especially if $\underline{\Delta}_a(x_n) = 0$. Hence, at most finitely many scalings with $\gamma_1 < 1$ lead to an accepted step. Thus, Algorithm 2.1 performs infinitely many successful steps if it does not terminate. We continue with proof by contradiction and assume that there exists a subsequence (x_{n_k}) of successful iterations with $C(x_{n_k}) > 2\varepsilon > 0$. Using this subsequence, we show that we get a certain decrease of the objective J and this yields a contradiction, since J is assumed to be bounded from below.

The radius $\underline{\Delta}_{\text{pred}}(\varepsilon)$ from Lemma A.1 and the assumed monotonicity of f yield

$$\operatorname{ared}(x,\Delta) \geq \eta \operatorname{pred}(x,\Delta) \qquad \wedge \qquad \operatorname{pred}(x,\Delta) \geq \frac{f(\varepsilon)}{2}\Delta \tag{A.1}$$

whenever $\Delta \in (0, \underline{\Delta}_a(x))$, $\Delta \leq \underline{\Delta}_{\text{pred}}(\varepsilon)$, and $C(x) > \varepsilon$. We define

$$\underline{\Delta} \coloneqq \min\left\{\underline{\Delta}_b, \underline{\Delta}_c, \underline{\Delta}_{\text{pred}}(\varepsilon), \frac{\varepsilon}{L}\right\}. \tag{A.2}$$

Next we set $i \coloneqq n_1$. First, we investigate the case $\Delta_i \geq \underline{\Delta}$. We define $\mathcal{I}_1 \coloneqq \{i\}$ and use a further distinction by cases.

- If $\Delta_i \geq \underline{\Delta}_a(x_i)$ and $\underline{\Delta}_a(x_i) \leq \underline{\Delta}_b$, we use Assumption 2.1(III.E) together with the monotonicity of the prediction to estimate

$$\operatorname{pred}(x_i,\Delta_i) \overset{\text{(I)}}{\geq} \operatorname{pred}(x_i, \underline{\Delta}_a(x_i)) \overset{\text{(III.E)}}{\geq} \delta. \tag{A.3}$$

- If $\Delta_i \geq \underline{\Delta}_a(x_i)$ and $\underline{\Delta}_a(x_i) > \underline{\Delta}_b$ then (A.2) implies $\underline{\Delta} \leq \underline{\Delta}_b < \underline{\Delta}_a(x_i)$. Thus,

$$\operatorname{pred}(x_i,\Delta_i) \overset{\text{(I)}}{\geq} \operatorname{pred}(x_i,\underline{\Delta}) \overset{\text{(A.1)}}{\geq} \frac{f(\varepsilon)}{2}\underline{\Delta} =: \delta_1 > 0. \tag{A.4}$$

- If $\underline{\Delta}_a(x_i) > \Delta_i$, we get $\underline{\Delta} \leq \Delta_i < \underline{\Delta}_a(x_i)$ and we can argue as in (A.4).

In the other case $\Delta_i < \underline{\Delta}$, we define $\mathcal{I}_1 \coloneqq \{i, \dots, K\}$, where $K \geq i$ is the smallest index such that at least one of the following conditions is satisfied:

(i) $\operatorname{pred}(x_K, \Delta_K) \geq \delta$,

(ii) $\sum_{l=i}^{K+1} \Delta_l \geq \underline{\Delta}$,

(iii) $\operatorname{ared}(x_{K+1}, \Delta_{K+1}) < \eta \operatorname{pred}(x_{K+1}, \Delta_{K+1})$, i.e. iteration $K+1$ is not successful.

We briefly argue that K is finite. By construction, $\mathcal{I}_1$ encodes only successful iterations for which the trust-region radius is nondecreasing. Condition (ii) leads to the bound $K \leq i + \frac{\underline{\Delta}}{\Delta_i} - 1 < \infty$. By definition, one of the conditions (i)–(iii) is always valid at K. Note that the construction directly yields $\sum_{l=i}^{K} \Delta_l < \underline{\Delta}$, since $\Delta_i \leq \underline{\Delta}$ was assumed. In particular, $\Delta_l < \underline{\Delta}$ for $l = i, \dots, K$. Next, we perform a distinction by cases on which of these conditions hold at K.

Case (i): By construction, the iteration K is successful and $\operatorname{pred}(x_K, \Delta_K) \geq \delta > 0$.

Case ¬(i): Assumption 2.1(III.E) together with $\neg(i)$ implies $\Delta_l \notin [\underline{\Delta}_a(x_l), \underline{\Delta}_b]$ for all $l \in \mathcal{I}_1$. Since we already know $\Delta_l \leq \underline{\Delta} \leq \underline{\Delta}_b$, we get

$$\Delta_l < \underline{\Delta}_a(x_l) \qquad \forall l = i, \dots, K. \tag{A.5}$$

Similarly, Assumption 2.1(IV.E) combined with $\neg(i)$ and $\Delta_l \leq \underline{\Delta} \leq \underline{\Delta}_c$ implies

$$|C(x_{l+1}) - C(x_l)| \leq L\Delta_l \qquad \forall l = i, \dots, K.$$

As a consequence, for any $l = i, \dots, K$, we can estimate

$$C(x_{l+1}) \geq C(x_l) - L\Delta_l \geq C(x_i) - L\sum_{k=i}^{l} \Delta_k \geq C(x_i) - L\underline{\Delta} \geq 2\varepsilon - L\frac{\varepsilon}{L} = \varepsilon. \tag{A.6}$$

Case (ii)∧¬(i): Since the iterations $l = i, \dots, K$ are successful, $\Delta_{K+1} \leq \gamma_2 \Delta_K$ implies

$$(1+\gamma_2)\sum_{l=i}^{K} \Delta_l \geq \sum_{l=i}^{K+1} \Delta_l + \underbrace{\gamma_2 \sum_{l=i}^{K-1} \Delta_l}_{\geq 0} \geq \underline{\Delta},$$

thus $\sum_{l=i}^{K} \Delta_l \geq \frac{\underline{\Delta}}{1+\gamma_2}$. From (A.6) we get $C(x_l) \geq \varepsilon$. Due to (A.5) we can apply (A.1) and get

$$\sum_{l=i}^{K} \operatorname{pred}(x_j, \Delta_j) \geq \frac{f(\varepsilon)}{2} \sum_{l=i}^{K} \Delta_l \geq \frac{f(\varepsilon)}{2} \frac{\underline{\Delta}}{1+\gamma_2} \eqqcolon \delta_2 > 0.$$

Case ¬(ii)∧¬(i): From ¬(ii), we get $\Delta_{K+1} < \underline{\Delta}$. In case that, additionally, $\Delta_{K+1} < \underline{\Delta}_a(x_{K+1})$, (A.1) implies acceptance of iteration $K+1$. Otherwise, $\Delta_{K+1} \geq \underline{\Delta}_a(x_{K+1})$ and (III.E) yields acceptance as well. This implies

$$(iii) \implies (i) \vee (ii),$$

i.e., the distinction by cases is complete. In all cases, we have shown that

$$\sum_{l \in I_1} \operatorname{pred}(x_l, \Delta_l) \geq \min\{\delta, \delta_1, \delta_2\} > 0.$$

We proceed inductively and define I_m, $m > 1$, by repeating the above construction with $i \coloneqq n_k$, where k is the smallest index such that $n_k \notin I_{m-1}$. This allows for the final estimate

$$\begin{aligned} J(x_1) - \lim_{n\to\infty} J(x_n) &= \sum_{n=1}^{\infty} J(x_n) - J(x_{n+1}) \\ &\geq \sum_{m=1}^{\infty} \sum_{l\in I_m} J(x_l) - J(x_{l+1}) \\ &= \sum_{m=1}^{\infty} \sum_{l\in I_m} \operatorname{ared}(x_l, \Delta_l) \\ &\geq \sum_{m=1}^{\infty} \sum_{l\in I_m} \eta \operatorname{pred}(x_l, \Delta_l) \\ &\geq \sum_{m=1}^{\infty} \eta \min\{\delta, \delta_1, \delta_2\} = \infty. \end{aligned}$$

For the first inequality, we use that the trust-region method only accepts steps that decrease the objective and in the third line we get the equality exactly from the restriction to successful iterates. In conclusion, this violates the assumption of J being bounded from below leading to a contradiction. Thus, no such subsequence x_{n_k} with $C(x_{n_k}) > 2\varepsilon$ can exist, finishing the proof. □

Remark A.2. The proof of Theorem 2.2 is constructive. The main ingredient is that for each criticality measure there exists a corresponding trust-region radius that guarantees sufficient improvement of the objective. This is described in detail for the first major case $\Delta_i > \underline{\Delta}$. It may however happen, that after an accepted step the next criticality measure is large, while the trust region radius is too small in relation. The construction of the second part of the proof uses the assumed Lipschitz continuity to ensure that the criticality measure remains large enough, while the algorithm expands the trust region. In this process it is proven that an unsuccessful iterate can only occur (condition (iii)) after a successful step with sufficient decrease was performed.

Remark A.3. As an interesting case, we discuss the situation in which $\underline{\Delta}_a(x) \equiv \underline{\Delta}_a > 0$, i.e., it does not depend on the point $x \in X$. Consequently, one could choose $\underline{\Delta}_b \in (0, \underline{\Delta}_a)$ arbitrarily and (III.E) would be automatically satisfied. If, additionally, there exists a bound $\underline{\Delta}_{\text{acc}} > 0$ such that

$$\operatorname{ared}(x, \Delta) \geq \eta \operatorname{pred}(x, \Delta) \qquad \forall x \in X, \Delta \in (0, \underline{\Delta}_{\text{acc}}),$$

then all trust-region radii are bounded from below by $\gamma_1 \underline{\Delta}_{\text{acc}}$. Now, it can be checked that Assumption 2.1(IV) and (IV.E) are no longer needed in the proof of Theorem 2.2. Thus, we can define

$$\underline{\Delta} \coloneqq \min\left\{\underline{\Delta}_{\text{pred}}(\varepsilon), \frac{\underline{\Delta}_{\text{acc}}}{\gamma_2}, \underline{\Delta}_a\right\} \tag{A.7}$$

and (A.1) holds for all $\Delta \in (0, \underline{\Delta}]$. By monotonicity, we also get $\operatorname{pred}(x, \Delta) \geq \frac{f(\varepsilon)}{2}\underline{\Delta}$ for all $\Delta > \underline{\Delta}$. This leads to the estimate

$$\operatorname{ared}(x_{n_k}, \Delta_{n_k}) \geq \eta \operatorname{pred}(x_{n_k}, \Delta_{n_k}) \geq \eta \frac{f(\varepsilon)}{2}\underline{\Delta}.$$

This is enough to argue as in the final estimate in the proof of Theorem 2.2. In particular, all the distinctions by cases can be skipped.

Note that this situation appears precisely in Section 3 for the proximal gradient algorithm, see Lemma 3.9 and Remark 3.11.

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, 2000, doi:10.1093/oso/9780198502456.001.0001.

[2] A. Y. Aravkin, R. J. Baraldi, and D. Orban, A Proximal Quasi-Newton Trust-Region Method for Nonsmooth Regularized Optimization, *SIAM Journal on Optimization* 32 (2022), 900–929, doi:10.1137/21m1409536.

[3] F. Bachmann, D. Beermann, J. Lu, and S. Volkwein, POD-Based Mixed-Integer Optimal Control of the Heat Equation, *Journal of Scientific Computing* 81 (2019), 48–75, doi:10.1007/s10915-019-00924-3.

[4] R. J. Baraldi and D. P. Kouri, A proximal trust-region method for nonsmooth optimization with inexact function and gradient evaluations, *Mathematical Programming* 201 (2022), 559–598, doi:10.1007/s10107-022-01915-3.

[5] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer International Publishing, 2017, doi:10.1007/978-3-319-48311-5.

[6] A. Beck, *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics, 2017, doi:10.1137/1.9781611974997.

[7] P. Bock de Barillas, F. M. Hante, and M. Hintermüller, Exact relaxation in optimal switching control for the heat equation, *ESAIM: Control, Optimisation and Calculus of Variations* 32 (2026), 39, doi:10.1051/cocv/2026019.

[8] F. Brauer, C. Castillo-Chavez, and Z. Feng, *Mathematical Models in Epidemiology*, Springer New York, 2019, doi:10.1007/978-1-4939-9828-9.

[9] C. Clason and K. Kunisch, A convex analysis approach to multi-material topology optimization, *ESAIM: Mathematical Modelling and Numerical Analysis* 50 (2016), 1917–1936, doi:10.1051/m2an/2016012.

[10] C. Clason, C. Tameling, and B. Wirth, Vector-Valued Multibang Control of Differential Equations, *SIAM Journal on Control and Optimization* 56 (2018), 2295–2326, doi:10.1137/16m1104998.

[11] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust Region Methods*, Society for Industrial and Applied Mathematics, 2000, doi:10.1137/1.9780898719857.

[12] S. Göttlich, F. M. Hante, A. Potschka, and L. Schewe, Penalty alternating direction methods for mixed-integer optimal control with combinatorial constraints, *Mathematical Programming* 188 (2021), 599–619, doi:10.1007/s10107-021-01656-9.

[13] S. Göttlich, A. Potschka, and C. Teuber, A partial outer convexification approach to control transmission lines, *Computational Optimization and Applications* 72 (2018), 431–456, doi:10.1007/s10589-018-0047-6.

[14] M. Hahn, S. Leyffer, and S. Sager, Binary optimal control by trust-region steepest descent, *Mathematical Programming* 197 (2022), 147–190, doi:10.1007/s10107-021-01733-z.

[15] Z. Huang, G. He, H. Yan, and H. Yu, Switch sequence optimization of heat pumps for micro-grid peak clipping, *Energy Procedia* 152 (2018), 64–70, doi:10.1016/j.egypro.2018.09.060.

[16] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115 (1927), 700–721, doi:10.1098/rspa.1927.0118.

[17] S. Leyffer and P. Manns, Sequential linear integer programming for integer optimal control with total variation regularization, *ESAIM: Control, Optimisation and Calculus of Variations* 28 (2022), 66, doi:10.1051/cocv/2022059.

[18] P. Manns, Convergence of Trust-Region Algorithms in Metric Spaces, *SIAM Journal on Optimization* 36 (2026), 841–865, doi:10.1137/24m1718627.

[19] P. Manns, M. Hahn, C. Kirches, S. Leyffer, and S. Sager, On Convergence of Binary Trust-Region Steepest Descent, *Journal of Nonsmooth Analysis and Optimization* Volume 4 (2023), doi:10.46298/jnsao-2023-10164.

[20] P. Manns and A. Schiemann, On Integer Optimal Control with Total Variation Regularization on Multidimensional Domains, *SIAM Journal on Control and Optimization* 61 (2023), 3415–3441, doi:10.1137/22m152116x.

[21] J. Marko and G. Wachsmuth, Integer optimal control problems with total variation regularization: Optimality conditions and fast solution of subproblems, *ESAIM: Control, Optimisation and Calculus of Variations* 29 (2023), 81, doi:10.1051/cocv/2023065.

[22] J. Marko and G. Wachsmuth, Vector-valued integer optimal control with TV regularization: Optimality conditions and algorithmic treatment, *ESAIM: Control, Optimisation and Calculus of Variations* 31 (2025), 51, doi:10.1051/cocv/2025037.

[23] M. Schmidt and F. M. Hante, Gas Transport Network Optimization: PDE-Constrained Models, in *Encyclopedia of Optimization*, Springer International Publishing, 2023, 1–7, doi:10.1007/978-3-030-54621-2_872-1.

[24] M. Severitt and P. Manns, Efficient Solution of Discrete Subproblems Arising in Integer Optimal Control with Total Variation Regularization, *INFORMS Journal on Computing* 35 (2023), 869–885, doi:10.1287/ijoc.2023.1294.

[25] M. Sharma, M. Hahn, S. Leyffer, L. Ruthotto, and B. van Bloemen Waanders, Inversion of convection–diffusion equation with discrete sources, *Optimization and Engineering* 22 (2020), 1419–1457, doi:10.1007/s11081-020-09536-5.