

# SINGLE-LOOP APPROACHES TO NONSMOOTH BILEVEL OPTIMISATION

Ensio Suonperä\*

Tuomo Valkonen†

**Abstract** We study bilevel optimisation problems in which the inner problem is represented as a set-valued, parametric constraint. We develop relevant optimistic and pessimistic calculus rules, derive corresponding optimality conditions, and formulate nonsmooth adjoint inclusions based on both the Fréchet and limiting coderivatives. Founded on these results, we propose a single-loop algorithm that accommodates a wide range of inner and adjoint steps, including those of primal-dual methods. We prove its convergence. Numerical experiments on total variation regularised inverse problems demonstrate the practicality of the approach.

## 1 INTRODUCTION

We consider bilevel optimisation problems of the form

$$(1.1) \quad \min_{u,x} J(u) + R(x) \quad \text{subject to} \quad 0 \in G(u, x)$$

in normed spaces  $U$  and  $X$ . The set-valued function  $G$  typically encodes the optimality conditions for an inner problem; for example, for  $g$  convex in the first argument, we can take  $G = \partial g(u; x)$  corresponding to the parametric inner problem  $\min_u g(u; x)$ . Very few works, so far, allow  $g$  to be nonsmooth;  $G$  to be set-valued. Most bilevel optimisation works impose strong differentiability, smoothness, and second-order growth assumptions on  $G$ .

Moreover, numerically, bilevel optimisation can be very expensive. At the outset, it requires solving the inner problem several times to find a solution to the problem (1.1). Therefore, in recent years, there has been a growing interest in “single-loop” methods [43, 44, 18, 30, 11, 28, 35, 48, 14, 32, 31, 23], as well as methods that solve the subproblems to a set accuracy [34, 42, 20]. Single-loop methods take only a single or a fixed number of steps of a conventional optimisation method towards a solution of the inner objective on each step of an algorithm to solve the outer problem. Most works take  $G(u, x) = \nabla_u g(u, x)$  for a smooth  $g$ , and use (stochastic) gradient descent as the inner optimisation method. In [44], we took  $G$  as primal-dual optimality conditions of  $\min_u f(u; x) + g(Ku; x)$ , which let us use the Primal-Dual Proximal Splitting (PDPS) of Chambolle and Pock [10] for the inner steps. We also allowed using the steps of standard algorithms for the approximate solution of an adjoint problem. The latter is used to compute  $[J \circ S]'(x)$ , where the solution mapping  $S$  is defined through the satisfaction of the inner problem,  $0 = G(S(x), x)$ . However, we still required  $G$  to be smooth.

*Our present results, based on coderivative calculus and nonsmooth adjoint inclusions, allow  $G$  and  $S$  to be set-valued; for  $g$  to be nonsmooth while also weakening second-order growth assumptions to metric*

---

This research has been supported by the Academy of Finland grants 314701, 320022, and 345486, and the Vilho, Yrjö and Kalle Väisälä Foundation grant for PhD studies.

\*Department of Mathematics and Statistics, University of Helsinki, Finland. [ensio.suonpera@helsinki.fi](mailto:ensio.suonpera@helsinki.fi), ORCID: [0000-0002-2111-0239](https://orcid.org/0000-0002-2111-0239)

†MODEMAT Research Center in Mathematical Modeling and Optimization, Quito, Ecuador and Department of Mathematics and Statistics, University of Helsinki, Finland, [tuomo.valkonen@iki.fi](mailto:tuomo.valkonen@iki.fi), ORCID: [0000-0001-6683-3572](https://orcid.org/0000-0001-6683-3572)

*subregularity*. The theory has connections to the recent [46], where, for inner problems of the form  $\min_u f(u) + g(u; x)$  with strongly convex and smooth  $f$  and convex  $g$ , the authors take  $G(u, x) = \gamma^{-1}(u - \text{prox}_{\gamma g(\cdot, x)}(u - \tau \nabla f(u)))$  as a proximal reformulation of the first-order necessary and sufficient optimality conditions. Then they employ limiting coderivatives of  $G$ . Such a proximal reformulation is commonly used with semismooth Newton methods, see, e.g., [13], as well as in the SparseHO of [4] for bilevel problems. Unlike [46], the latter, however, requires significant differentiability from the proximal reformulation. Our theory also applies to  $G$  as above while also allowing it to be more general and set-valued.

Algorithmically, both [46, 4] are based on exact or near-exact differentials or coderivatives of the inner solution mapping in an outer forward-backward method or a nonsmooth trust region method. We propose a highly-efficient single-loop method instead, based on the previous, still smooth, works [43, 44, 18]. Our approach can work both with the Fréchet and limiting coderivatives of  $G$ . Moreover, our numerical examples would not be practical to solve with forward-backward splitting as the inner algorithm, as the involved proximal map would not have closed-form solution. Our approach allows the use of (the steps of) more appropriate primal-dual methods.

There are a few relevant recent more analytical works – that do not represent algorithms – for nonsmooth bilevel optimisation based on the solution mapping. In particular, [6, 7] treat nonsmooth inner problems through proximal reformulations. They rely on Rademacher’s theorem to obtain differentiability almost everywhere in  $\mathbb{R}^n$ . However, those tools are not available in our infinite dimensional functional analytic setting. On the other hand, [25, 24] show that metric regularity of an inner problem with linear dependence on the outer parameters (i.e., tilt stability), is, often, equivalent to the Lipschitz continuity of the solution mapping.

Methods of more second-order flavour have also been proposed: [12] study bundle-based trust region methods in special cases, while [33] propose a Levenberg–Marquardt method based on Newton-derivatives. The latter is based on finding roots of the optimality conditions of an alternative constrained single-level reformulation, known as the value function reformulation [39]. Although such a simple reformulation may sound attractive, not requiring the solution mapping, it still depends on knowing the value function  $\varphi(x) := \min_u g(u, x)$ , which is no easier to compute than the solution mapping  $S(u) = \arg \min_u g(u, x)$ . Here we recall that  $g$  is the inner objective when  $G$  arises as  $\partial g(u; x)$ . For smooth problems, similarly to our approach [43, 44] based on the solution mapping, [34] avoids knowing  $\varphi$  through solution quality control. Moreover, for nonsmooth problems, the value function reformulation poses several challenges with the satisfaction of constraint qualifications. This is also true of the conventional approach of reformulating (1.1) as an MPCC (Mathematical Program with Complementarity Constraints) [36, 21], and finding roots of the full-system KKT conditions. That reformulation is also no longer completely equivalent to the original problem, while for convex  $g$ , (1.1) remains equivalent. We do not, however, entirely avoid difficulties with coderivative calculus closely related to constraint qualifications. Alternative duality-based reformulations for nonsmooth problems have recently been studied in, e.g., [17].

We start our journey by developing optimistic and pessimistic calculus rules for the Fréchet sub-differential and coderivative in Section 2: it turns out that the directions of inclusions the calculus correspond to optimistic and pessimistic solutions of bilevel problems. We then use this calculus in Section 3 to derive optimality conditions for nonsmooth bilevel optimisation problems of form (1.1). Based on these optimality conditions, we devote Section 4 to discussing a general single-loop algorithms for (1.1). We base this on the general “tracking theory” of differential estimation from [18]. Section 5 focuses providing an actual algorithm when the inner problem is total variation regularised inverse problem. We describe our numerical experiments and their results in Section 6.

## NOTATION AND BASIC CONCEPTS

We write  $\langle x^* | x \rangle_{X^*, X} := x^*(x)$ , or, for short,  $\langle x^* | x \rangle$  for the dual pairing in a normed space  $X$  and its dual  $X^*$ . The notation  $\langle x, y \rangle$  is reserved for inner products in Hilbert spaces, where we also write  $\mathcal{R}x^* \in X$  for the Riesz representation of  $x^* \in X^*$ . The notation  $\mathbb{L}(X; Y)$  stands for the space of bounded linear operators between  $X$  and  $Y$ . The identity operator is written  $\text{Id}$ .

We now introduce several concepts of set-valued analysis; for further details we refer to [13]. Let  $F : X \rightarrow \overline{\mathbb{R}} := (-\infty, \infty]$  in a normed space  $X$ . We write  $\text{dom } F := \{x \in X \mid F(x) < \infty\}$  for the effective domain and say that  $F$  is proper if  $\text{dom } F \neq \emptyset$ . The Fenchel conjugate of  $F$  is written  $F^*$ , and the proximal map by  $\text{prox}_F$ .

The convex subdifferential of  $F$  at  $x$  reads

$$(1.2) \quad \partial F(x) := \{x^* \in X^* \mid F(\tilde{x}) - F(x) \geq \langle x^* | \tilde{x} - x \rangle \text{ for all } \tilde{x}\}.$$

The  $\varepsilon$ -subdifferential, for a  $\varepsilon \geq 0$ , has the definition

$$(1.3) \quad \partial_\varepsilon F(x) := \left\{ x^* \in X^* \mid \liminf_{\tilde{x} \rightarrow x} \frac{F(\tilde{x}) - F(x) - \langle x^* | \tilde{x} - x \rangle}{\|\tilde{x} - x\|} \geq -\varepsilon \right\}.$$

The Fréchet subdifferential is  $\partial_F F(x) = \partial_0 F(x)$ . The limiting subdifferential is then

$$\partial_M F(x) := \{x^* \in X^* \mid \text{there exist } \tilde{x} \rightarrow x, \varepsilon \searrow 0, \text{ and } \tilde{x}^* \in \partial_\varepsilon F(\tilde{x}) \text{ with } \tilde{x}^* \overset{*}{\rightharpoonup} x^* \text{ and } F(\tilde{x}) \rightarrow F(x)\}.$$

If  $X$  is reflexive, we can simplify

$$\partial_M F(x) = \{x^* \in X^* \mid \text{there exist } \tilde{x} \overset{*}{\rightharpoonup} x \text{ and } \tilde{x}^* \in \partial F(\tilde{x}) \text{ with } \tilde{x}^* \overset{*}{\rightharpoonup} x^*\}.$$

For convex  $F$ ,  $\partial_F F = \partial_M F = \partial F$ . For continuously differentiable  $F$ ,  $\partial_F F(x) = \partial_M F(x) = \{F'(x)\}$ .

For a set-valued  $S : X \rightrightarrows Y$ , where also  $Y$  is a normed space, the  $\varepsilon$ -Fréchet coderivative at  $x$  for  $y$ , denoted by  $\widehat{D}_\varepsilon^* F(x|y) : Y^* \rightrightarrows X^*$ , is defined through

$$(1.4) \quad \widehat{D}_\varepsilon^* F(x|y)(y^*) := \left\{ x^* \in X^* \mid \limsup_{\substack{\tilde{x} \rightarrow x, F(\tilde{x}) \ni \tilde{y} \rightarrow y \\ \tilde{y}^* \overset{*}{\rightharpoonup} y^*, \varepsilon \searrow 0}} \frac{\langle x^* | \tilde{x} - x \rangle - \langle y^* | \tilde{y} - y \rangle}{\|\tilde{x} - x\| + \|\tilde{y} - y\|} \leq \varepsilon \right\}.$$

That is,  $(x^*, -y^*) \in \widehat{N}_{\text{Graph } F}^\varepsilon(x, y)$ , where  $\widehat{N}_{\text{Graph } F}^\varepsilon$  is the  $(\varepsilon)$ -Fréchet normal cone to  $\text{Graph } F$ . We call  $\widehat{D}^* F(x|y) := \widehat{D}_0^* F(x|y)$  the Fréchet coderivative at  $x$  for  $y$ . We do not impose  $y \in F(x)$ : it is possible that  $\text{Graph } \widehat{D}^* F(x|y) \neq \emptyset$  for  $y \in \text{cl } F(x)$ .

The (normal) limiting/Mordukhovich coderivative is then defined as

$$D^* F(x|y)(y^*) := \text{w-}^*\text{-lim sup}_{\substack{(\tilde{x}, \tilde{y}) \rightarrow (x, y), \\ \tilde{y}^* \overset{*}{\rightharpoonup} y^*, \varepsilon \searrow 0}} \widehat{D}_\varepsilon^* F(\tilde{x}|\tilde{y})(\tilde{y}^*).$$

It also has the smaller “mixed” variant

$$D_M^* F(x|y)(y^*) := \text{w-}^*\text{-lim sup}_{\substack{(\tilde{x}, \tilde{y}) \rightarrow (x, y), \\ \tilde{y}^* \rightarrow y^*, \varepsilon \searrow 0}} \widehat{D}_\varepsilon^* F(\tilde{x}|\tilde{y})(\tilde{y}^*).$$

We call  $F$   $N$ -regular at  $(x, y)$  if  $D_M F(x|y) = \widehat{D}^* F(x|y)$ .

We recall that if  $F$  is Gateaux-differentiable, and the differential  $DF : X \rightarrow X^*$  is Lipschitz continuous with factor  $L$ , then the *descent inequality* holds (see, e.g., [13, Lemma 7.1])

$$F(y) \leq F(x) + \langle DF(x) | y - x \rangle + \frac{L}{2} \|x - y\|^2.$$

If  $F$  is convex and  $X$  is reflexive, the two properties are equivalent. If  $F$  is Fréchet differentiable, we write  $F'$  instead of  $DF$  for the differential. We write partial derivatives (generally Fréchet) of  $G : X \times Y \rightarrow Z$ ,  $G : (x, y) \rightarrow z$  as  $G^{(x)}(x, y) \in \mathbb{L}(X; Z)$  and  $G^{(y)}(x, y) \in \mathbb{L}(Y; Z)$ .

For a set  $B$  and a point  $a$  in a metric space  $(X, d)$ , we define

$$\text{dist}(B, a) := \text{dist}(a, B) := \inf_{b \in B} d(a, b).$$

When  $M \in \mathbb{L}(X; X^*)$  is self-adjoint and positive (semi-)definite, we write  $\|x\|_M := \sqrt{\langle Mx|x \rangle}$  as well as  $\text{dist}_M(A, u) = \inf_{\tilde{u} \in A} \|\tilde{u} - u\|_M$ . We also write  $B_M(u, \delta) = \{\tilde{u} \in U \mid \|\tilde{u}\|_M \leq \delta\}$ .

Finally, we write  $\text{proj}_A(x) := \{z \in A \mid \text{dist}(A, x) = d(z, x)\}$ .

## 2 CALCULUS FOR THE FRÉCHET SUBDIFFERENTIAL AND CODERIVATIVE

In this section, we provide chain and sum rules for the Fréchet subdifferential and coderivative, for which results are generally much less developed than for the limiting coderivative. We refer to [41, 13, 37] for such classical results. In case of the sum rule, we exploit convexity of the second component. In case of the chain rule, we obtain otherwise relaxed assumptions by considering optimistic and pessimistic variants of scalar functions. Towards this end, for a function  $F : X \rightrightarrows \overline{\mathbb{R}}$ , we define the optimistic and pessimistic selections

$$\text{Inf}_F(x) := \inf\{v \mid v \in F(x)\} \quad \text{and} \quad \text{Sup}_F(x) := \sup\{v \mid v \in F(x)\}.$$

For the composition  $J \circ S$  with  $J : U \rightarrow \overline{\mathbb{R}}$  and  $S : X \rightrightarrows U$ , we also define the optimistic and pessimistic inner maps

$$S^\sharp(\tilde{x}) := \{u \in S(\tilde{x}) \mid J(u) = \text{Inf}_{J \circ S}(\tilde{x})\}$$

and

$$S^\flat(\tilde{x}) := \{u \in S(\tilde{x}) \mid J(u) = \text{Sup}_{J \circ S}(\tilde{x})\}.$$

We start by introducing the regularity properties that we will work with.

### 2.1 REGULARITY PROPERTIES

We call  $F : X \rightrightarrows Y$  between normed spaces  $X$  and  $Y$  *inner Lipschitz at  $x$*  relative to a set  $A \subset Y$  if for every  $y \in A$  there exist  $\lambda_y \geq 0$  and  $\rho_y > 0$  (which may depend on the point  $y$ ) such that

$$\text{dist}(F(\tilde{x}), y) \leq \lambda_y \|x - \tilde{x}\| \quad \text{for all} \quad \tilde{x} \in B(x, \rho_y) \cap \text{dom} F.$$

We call  $F$  *inner semi-Lipschitz at  $x$*  relative to a set  $A \subset Y$  if, given a sequence  $\text{dom} F \ni x^k \rightarrow x$ , there exists a subsequence, unrelabelled, and *some*  $y \in A$  and  $\lambda_y \geq 0$ , such that

$$\text{dist}(F(x^k), y) \leq \lambda_y \|x - x^k\| \quad \text{for all} \quad k \in \mathbb{N}.$$

If  $A = F(x)$ , we drop the specification “relative to  $A$ ”.

**Remark 2.1.** Aside from Lemma 2.5 in the next subsection, where it is possible that  $A \cap \text{cl} F(x) \setminus F(x) \neq \emptyset$ , we will take  $A \subset F(x)$ .

Clearly an inner Lipschitz mapping is also inner semi-Lipschitz relative to any  $A \subset F(x)$ , and a single-valued mapping that is Lipschitz at the point  $x$ , is inner Lipschitz at  $x$ .

**Remark 2.2 (Outer Lipschitz, linear recession and inner calmness properties).** The inner Lipschitz properties bear superficial resemblance to outer Lipschitz and linear recession properties, however, are distinct; see [19, §3.4] and [29]. If  $F^{-1}$  is inner Lipschitz at  $y$  (relative to a set  $A \subset X$ ) then  $F$  satisfies inner calmness (w.r.t.  $A$ ) property introduced in [3, 2]. The latter is weaker property as it doesn't require Lipschitz inequality to hold in entire neighbourhood but only for some sequence in it.

**Remark 2.3 (The Aubin property).** If  $F$  has the Aubin property at  $\bar{x}$  for  $\bar{y}$ , i.e.,

$$\text{dist}(F(\tilde{x}), \tilde{y}) \leq \kappa \text{dist}(F^{-1}(\tilde{y}), \tilde{x}) \quad \text{for all } \tilde{x} \in B(\bar{x}, \delta_x) \text{ and } \tilde{y} \in B(\bar{y}, \delta_y),$$

for some factor (called the *graphical modulus*)  $\kappa > 0$  and radii  $\delta_x, \delta_y > 0$ , then it also has the inner Lipschitz property at  $\bar{x}$  relative to  $A = B(\bar{y}, \delta_y)$ . However, the inner Lipschitz property does not imply the Aubin property, especially as it only considers  $\tilde{x} \in \text{dom } F$ , and allows the factors  $\lambda_y$  to depend on  $y$ .

The next lemma considers inverses of translated radial functions, in particular, inverses of  $J(x) = \frac{1}{2}\|x - b\|^2$ .

**Lemma 2.4.** *In a normed space  $Y$ , let  $J : Y \rightarrow \mathbb{R}$ ,  $J(y) = \varphi(\|y - b\|)$  for some  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  with a  $\kappa$ -Lipschitz inverse. Then  $J^{-1}$  is inner Lipschitz.*

*Proof.* Let  $v = J(y)$  for some  $y$ . Then

$$J^{-1}(\tilde{v}) = \{\tilde{y} \mid \|\tilde{y} - b\| = \varphi^{-1}(\tilde{v})\} \ni \frac{\varphi^{-1}(\tilde{v})}{\varphi^{-1}(v)}(y - b) + b.$$

It follows for any  $\tilde{v} \in \text{dom } J^{-1} = \text{ran } J$  that

$$(2.1) \quad \text{dist}(J^{-1}(\tilde{v}), y) \leq \left| \frac{\varphi^{-1}(\tilde{v})}{\varphi^{-1}(v)} - 1 \right| \|y - b\| = |\varphi^{-1}(\tilde{v}) - \varphi^{-1}(v)| \leq \kappa |\tilde{v} - v|.$$

That is,  $J^{-1}$  is inner Lipschitz. □

## 2.2 CHAIN RULES

We now start our calculus results with chain rules. Our main result follows from the next lemma when we take  $g = \text{Inf}_{J \circ S}$  or  $g = \text{Sup}_{J \circ S}$ , but it can also be applied to derive calculus rules for arbitrary selections of  $J \circ S$ . In (ii) and (iii), note that it is possible that  $u \in A \subset S_g^b(x)$  with  $u \in \text{cl } S_g^\#(x) \setminus S_g^\#(x)$ , and  $S_g^\#(x|u)(u^*) \neq \emptyset$ .

**Lemma 2.5.** *Let  $S : X \rightrightarrows U$ ,  $J : U \rightarrow \mathbb{R}$ , and  $g : X \rightarrow \overline{\mathbb{R}}$  on normed spaces  $U$  and  $X$ . Define*

$$S_g^\#(x) := \{u \in S(x) \mid g(x) \geq J(u)\} \quad \text{and} \quad S_g^b(x) := \{u \in S(x) \mid g(x) \leq J(u)\}.$$

*Then the following hold at any  $x \in X$ :*

(i) *If  $J$  is continuously differentiable near  $x$  and  $S_g^\#(x)$  is non-empty, then*

$$\partial_F g(x) \subset \bigcap_{u \in S_g^\#(x)} \widehat{D}^* S_g^b(x|u)(J'(u)).$$

(ii) *If  $S_g^\#$  is inner Lipschitz at  $x$  relative to an  $A \subset S_g^b(x)$ , then*

$$\bigcup_{u \in A, u^* \in \partial_F J(u)} \widehat{D}^* S_g^\#(x|u)(u^*) \subset \partial_F g(x).$$

(iii) If  $S_g^\sharp$  is inner semi-Lipschitz at  $x$  relative to an  $A \subset S_g^b(x)$ , then

$$\bigcap_{u \in A} \bigcup_{u^* \in \partial_F J(u)} \widehat{D}^* S_g^\sharp(x|u)(u^*) \subset \partial_F g(x).$$

*Proof.* (i): Let  $x^* \in \partial_F g(x)$ . Then, by definition

$$(2.2) \quad \limsup_{x^k \rightarrow x} \frac{\langle x^* | x^k - x \rangle - (g(x^k) - g(x))}{\|x^k - x\|} \leq 0.$$

By assumption there exists some  $u \in S_g^\sharp(x)$ . Take any. Then  $J(u) \leq g(x)$ . Since  $J(u^k) \geq g(x^k)$  for any  $u^k \in S_g^b(x^k)$ , the characterisation (2.2) implies

$$\limsup_{x^k \rightarrow x, S_g^b(x^k) \ni u^k \rightarrow u} \frac{\langle x^* | x^k - x \rangle - (J(u^k) - J(u))}{\|x^k - x\| + \|u^k - u\|} \leq 0.$$

By the mean value theorem, for some  $\zeta^k \in [u, u^k]$ , this implies

$$\limsup_{x^k \rightarrow x, S_g^b(x^k) \ni u^k \rightarrow u} \frac{\langle x^* | x^k - x \rangle - \langle J'(u) | u^k - u \rangle + \langle J'(u) - J'(\zeta^k) | u^k - u \rangle}{\|x^k - x\| + \|u^k - u\|} \leq 0.$$

By the continuity of  $J'$ , we have

$$\lim_{u^k \rightarrow u} \frac{\langle J'(u) - J'(\zeta^k) | u^k - u \rangle}{\|u^k - u\|} = 0.$$

Thus,

$$(2.3) \quad \limsup_{x^k \rightarrow x, S_g^b(x^k) \ni u^k \rightarrow u} \frac{\langle x^* | x^k - x \rangle - \langle J'(u) | u^k - u \rangle}{\|x^k - x\| + \|u^k - u\|} \leq 0,$$

which is to say  $x^* \in \widehat{D}^* S_g^b(x|u)(J'(u))$ . Since  $u \in S_g^\sharp(x)$ , we obtain (i).

(ii): Suppose  $x^* \in \widehat{D}^* S_g^\sharp(x|u)(u^*)$  for a  $u \in A \subset S_g^b(x)$  and  $u^* \in \partial_F J(u)$ . The former is characterised by the inequality analogous to (2.3),

$$(2.4) \quad \limsup_{x^k \rightarrow x, S_g^b(x^k) \ni u^k \rightarrow u} \frac{\langle x^* | x^k - x \rangle - \langle u^* | u^k - u \rangle}{\|x^k - x\| + \|u^k - u\|} \leq 0.$$

We need to show that  $x^* \in \partial_F g(x)$ . By definition, this holds if and only if (2.2) holds. Therefore, take any  $x^k \rightarrow x$ . Since  $S_g^\sharp$  is inner Lipschitz at  $x$  relative to  $A \ni u$  there exists  $\kappa_u > 0$  such that, for any  $\kappa' > \kappa_u$ , for large enough  $k$ , we can find  $u^k \in S_g^b(x^k)$  with  $\|u^k - u\| \leq \kappa' \|x^k - x\|$ . Thus, in particular,  $S_g^b(x^k) \ni u^k \rightarrow u$ . But then (2.4) implies

$$(2.5) \quad \limsup_{k \rightarrow \infty} \frac{\langle x^* | x^k - x \rangle - \langle u^* | u^k - u \rangle}{\|x^k - x\| + \|u^k - u\|} \leq 0.$$

By the definition (1.2) of the Fréchet subdifferential, we have

$$\limsup_{k \rightarrow \infty} \frac{\langle u^* | u^k - u \rangle - (J(u^k) - J(u))}{\|x^k - x\|} \leq 0.$$

Using this and  $\|u^k - u\| \leq \kappa' \|x^k - x\|$  in (2.5), we obtain

$$(2.6) \quad \limsup_{k \rightarrow \infty} \frac{\langle x^* | x^k - x \rangle - (J(u^k) - J(u))}{\|x^k - x\|} \leq 0.$$

Keeping in mind that  $g(x^k) \geq J(u^k)$  by the definition of  $S^\sharp(x^k) \ni u^k$ , and  $g(x) \leq J(u)$  by the definition of  $S^b(x) \ni u$ , since the sequence  $x^k \rightarrow x$  was arbitrary, this proves (2.2).

(iii): We proceed analogously to (ii), except in (2.5), after the use of the inner semi-Lipschitz property relative to  $S_g^\sharp(x)$ , we have passed to a subsequence, unrelabelled, and  $u \in A \subset S_g^b(x)$  is no longer arbitrary, but determined by the property. To justify passing from (2.6) to (2.2), we argue that we can repeat the argument for an arbitrary subsequence of the original. Note that the choice of  $u$  does not depend on  $u^* \in \partial_F J(u)$ , therefore, the union over  $u^*$  can be inside the intersection over  $u$  in the claim.  $\square$

**Theorem 2.6.** *Let  $S : X \rightrightarrows U$ , and  $J : U \rightarrow \mathbb{R}$  on normed spaces  $U$  and  $X$ . Then the following hold:*

(i) *If  $J$  is continuously differentiable near  $x$  and  $S^\sharp(x)$  is non-empty, then*

$$\partial_F \text{Inf}_{J \circ S}(x) \subset \bigcap_{u \in S^\sharp(x)} \widehat{D}^* S(x|u)(J'(u)).$$

(ii) *If  $S$  is inner Lipschitz at  $x$  relative to an  $A \subset S^b(x) \subset S(x)$ , then*

$$\bigcup_{u \in A, u^* \in \partial_F J(u)} \widehat{D}^* S(x|u)(u^*) \subset \partial_F \text{Sup}_{J \circ S}(x).$$

(iii) *If  $S$  is inner semi-Lipschitz at  $x$  relative to an  $A \subset S^b(x) \subset S(x)$ , then*

$$\bigcap_{u \in A} \bigcup_{u^* \in \partial_F J(u)} \widehat{D}^* S(x|u)(u^*) \subset \partial_F \text{Sup}_{J \circ S}(x).$$

*Proof.* (i): Apply Lemma 2.5 (i) with  $g = \text{Inf}_{J \circ S}$ . Then  $S_g^\sharp = S^\sharp$ , and  $S_g^b = S$ .

(ii) and (iii): Apply Lemma 2.5 (ii) and (iii) with  $g = \text{Sup}_{J \circ S}$ . Then  $S_g^\sharp = S$ , and  $S_g^b = S^b$ .  $\square$

### 2.3 SUM RULES

We now provide a sum rule that relaxes assumptions in known sum rules; see, e.g. [13, 41]. For  $A \subset X$ , we set,

$$(2.7) \quad F_A(x) := \begin{cases} \emptyset, & x \notin A, \\ F(x), & x \in A. \end{cases}$$

Then we define the  $A$ -lower subdifferential

$$\partial_{|A} R(x) := \left\{ x^* \in \partial_F R(x) \left| \limsup_{A \ni x^k \rightarrow x} \frac{R(x^k) - R(x) - \langle x^* | x^k - x \rangle}{\|x^k - x\|} \leq 0 \right. \right\}.$$

Despite relaxing standard results, we still have the relatively strict assumption  $\partial_{|A} R(x) \neq \emptyset$ , which, in particular, excludes the one-norm, but does allow several forms of constraints.

**Example 2.7.**  $\partial_{|\text{dom} R} R(x) = \partial_F R(x)$  if  $R$  is Lipschitz differentiable.

**Example 2.8.** Since  $0 \in \partial \delta_C(x)$  for all  $x \in C$  for a non-empty closed convex set  $C$ , we always have  $0 \in \partial_{|\text{dom} \delta_C} \delta_C(x)$ .

**Example 2.9.** For  $R(x) = |x|$  on  $\mathbb{R}$ , we have  $\partial_{|[0,\infty)} R(0) = \{1\}$  and  $\partial_{|(-\infty,0]} R(0) = \{-1\}$ .

**Theorem 2.10.** Suppose  $R : X \rightarrow \overline{\mathbb{R}}$  is convex, proper, and locally Lipschitz on its effective domain (not only the interior of the domain), and  $F : X \rightarrow \mathbb{R}$  on a normed space  $X$ . Let  $x \in \text{dom} R$ . Then the following hold:

(i) The inclusion<sup>1</sup>

$$\partial_F F(x) + \partial_F R(x) \subset \partial_F [F + R](x)$$

(ii) If,  $\partial_{|A} R(x) \neq \emptyset$  for a set  $A \subset X$ , the inclusion

$$\partial_F [F + R](x) \subset \partial_F F_A(x) + \tilde{x}^* \quad \text{for any } \tilde{x}^* \in \partial_{|A} R(x).$$

*Proof.* By definition,  $x^* \in \partial_F [F + R](x)$  if and only if

$$(2.8) \quad \limsup_{x^k \rightarrow x} \frac{\langle x^* | x^k - x \rangle - ([F + R](x^k) - [F + R](x))}{\|x^k - x\|} \leq 0.$$

(i): Let  $\tilde{x}^* \in \partial_F R(x)$ . Using the definition (1.3) of the latter, we see (2.8) to hold if

$$(2.9) \quad \limsup_{x^k \rightarrow x} \frac{\langle x^* - \tilde{x}^* | x^k - x \rangle - (F(x^k) - F(x))}{\|x^k - x\|} \leq 0.$$

This says exactly that  $x^* - \tilde{x}^* \in \partial_F F(x)$ , proving (i).

(ii): In the other direction, restricting  $x^k \in A$ , (2.8) implies for any  $\tilde{x}^* \in X^*$  that

$$(2.10) \quad \limsup_{A \ni x^k \rightarrow x} \frac{\langle x^* - \tilde{x}^* | x^k - x \rangle - (F(x^k) - F(x))}{\|(x^k - x, F(x^k) - F(x))\|} \leq \limsup_{A \ni x^k \rightarrow x} \frac{R(x^k) - R(x) - \langle \tilde{x}^* | x^k - x \rangle}{\|x^k - x\|}.$$

The right-hand-side is non-positive for any  $\tilde{x}^* \in \partial_{|A} R(x) \subset \partial_F R(x)$ . We have assumed  $\partial_{|A} R(x)$  non-empty. Therefore, such a choice of  $\tilde{x}^*$  exists. From (2.10) we now get that  $x^* \in \partial_F F_A(x) + \tilde{x}^*$ . This proves (ii).  $\square$

**Remark 2.11.** Similar sum rule to (ii) is obtained for  $\tilde{x}^* \in -\partial_F(-R(x))$  in [38]. The function  $x \mapsto -\partial_F(-R(x))$  is called Fréchet upper subdifferential therein, and it behaves like  $A$ -lower subdifferential in Examples 2.7 to 2.9.

<sup>1</sup>This inclusion holds in general.

### 3 OPTIMALITY CONDITIONS FOR NONSMOOTH BILEVEL PROBLEMS

Throughout this section, on normed spaces  $U$ ,  $W_*$ ,  $W = (W_*)^*$ , and  $X$ , where  $W_*$  is a predual space of  $W$ , we consider the bilevel optimisation problem (1.1), i.e.,

$$(3.1) \quad \min_{u,x} J(u) + R(x) \quad \text{subject to} \quad 0 \in G(u, x).$$

Here  $G : U \times X \rightrightarrows W_*$ ,  $J : U \rightarrow \mathbb{R}$ , and  $R : X \rightarrow \overline{\mathbb{R}}$ .

We start in Section 3.1 by deriving optimality conditions for (3.1) or, more precisely, the optimistic and pessimistic variants

$$\min_x [\text{Inf}_{J \circ S} + R](x) \quad \text{and} \quad \min_x [\text{Sup}_{J \circ S} + R](x),$$

where the optimistic and pessimistic selections  $\text{Inf}_{J \circ S}$  and  $\text{Sup}_{J \circ S}$ , are defined in Section 2 along with the corresponding slices  $S^\#$  and  $S^b$  of the inner solution mapping

$$(3.2) \quad S : X \rightrightarrows U, \quad S(x) := \{u \in U \mid 0 \in G(u, x)\}.$$

In Section 3.2 we further characterise these optimality conditions with the help of nonsmooth adjoint inclusions. We finish the section in Section 3.3 by writing out the full inner-adjoint-outer optimality system with the help of additional variables.

#### 3.1 CRITICALITY CONDITIONS

By the Fermat principles for the Fréchet and limiting subdifferentials [13, Theorem 16.2 and Corollary 16.6], for  $\bar{x}$  to be an optimistic minimiser of (3.1), the condition

$$(3.3) \quad 0 \in \check{\partial}[\text{Inf}_{J \circ S} + R](\bar{x})$$

is necessary for both  $\check{\partial} = \partial_M, \partial_F$ . Likewise for  $\bar{x}$  to be a pessimistic minimiser of (3.1), it is necessary that

$$0 \in \check{\partial}[\text{Sup}_{J \circ S} + R](\bar{x}).$$

The next results provide further expanded conditions under additional assumptions. For the first result, recall the definition of  $S_A$  from (2.7).

**Theorem 3.1.** *Let  $J$ ,  $S$ , and  $R$  be as in (3.1) and (3.2) with  $J$  continuously differentiable. Suppose  $\text{Inf}_{J \circ S} + R$  is minimised at  $\bar{x}$ . If  $\partial_{|A}R(\bar{x}) \neq \emptyset$  for a set  $A \subset X$ , then*

$$(3.4) \quad 0 \in \widehat{D}^*S_A(\bar{x}|\bar{u})(J'(\bar{u})) + \partial_{|A}R(\bar{x}) \quad \text{for all} \quad \bar{u} \in S^\#(\bar{x}).$$

*If, in particular,  $\text{dom} R = X$  with  $\partial_{|X}R(\bar{x}) \neq \emptyset$  (e.g., if  $R$  is Lipschitz continuously differentiable), then*

$$(3.5) \quad 0 \in \widehat{D}^*S(\bar{x}|\bar{u})(J'(\bar{u})) + \partial_F R(\bar{x}) \quad \text{for all} \quad \bar{u} \in S^\#(\bar{x}).$$

*Proof.* Let  $\bar{v} := \text{Inf}_{J \circ S}(\bar{x})$ . By (3.3), Theorem 2.10 (ii) and Theorem 2.6 (i)

$$\begin{aligned} 0 \in \partial_F[\text{Inf}_{J \circ S} + R](\bar{x}) &\subset \widehat{D}^*[J \circ S]_A(\bar{x}|\bar{v})(1) + \partial_{|A}R(\bar{x}) \\ &= \widehat{D}^*[J \circ S_A](\bar{x}|\bar{v})(1) + \partial_{|A}R(\bar{x}) \\ &\subset \bigcap_{u \in S^\#(\bar{x})} \widehat{D}^*S_A(\bar{x}|u)(J'(u)) + \partial_{|A}R(\bar{x}). \end{aligned}$$

This proves (3.4). For (3.5) we take  $A = X$ . □

Since the limiting coderivative and subdifferential contain the Fréchet coderivative and subdifferential, the following is immediate.

**Corollary 3.2.** *Theorem 3.1 holds with the limiting coderivative in place of the Fréchet coderivative.*

We will, however, wish to avoid the  $A = \text{dom } R$  restriction on  $S$ , and the restrictions on  $R$ , so base our work with the Fréchet coderivative mainly on the condition of the next lemma, which is neither necessary nor sufficient, but is a *sufficient condition for the pessimistic necessary condition* (3.3). It will be necessary in each case to verify that (3.6) or (3.7) can be satisfied (e.g., through becoming necessary); it is no way necessary that they can be satisfied.

**Theorem 3.3.** *Consider  $F := J \circ S + R$ , where  $J$ ,  $S$ , and  $R$  are as in (3.1) and (3.2). Then the pessimistic necessary optimality condition  $0 \in \partial_F[\text{Sup}_{J \circ S} + R](\bar{x})$  holds in the following cases:*

(i)  $S$  is inner Lipschitz at  $\bar{x}$  relative to  $\{\bar{u}\}$  for a  $\bar{u} \in S^b(\bar{x})$ , and

$$(3.6) \quad 0 \in \bigcup_{u^* \in \partial_F J(\bar{u})} \widehat{D}^* S(\bar{x}|\bar{u})(u^*) + \partial_F R(\bar{x})$$

(ii)  $S$  is inner semi-Lipschitz at  $\bar{x}$  relative to an  $A \subset S^b(\bar{x})$ , and

$$(3.7) \quad 0 \in \bigcap_{\bar{u} \in A} \bigcup_{u^* \in \partial_F J(\bar{u})} \widehat{D}^* S(\bar{x}|\bar{u})(u^*) + \partial_F R(\bar{x}).$$

*Proof.* (i): Let  $\bar{v} := J(\bar{u})$ . By Theorem 2.10 (i) and Theorem 2.6 (ii), we have

$$(3.8) \quad \partial_F[\text{Sup}_{J \circ S} + R](\bar{x}) \supset \widehat{D}^*[J \circ S](\bar{x}|\bar{v})(1) + \partial_F R(\bar{x}) \supset \bigcup_{u^* \in \partial_F J(\bar{u})} \widehat{D}^* S(\bar{x}|\bar{u})(u^*) + \partial_F R(\bar{x}).$$

(ii): We use Theorem 2.6 (iii) instead of (ii) in the final equality of (3.8).  $\square$

**Remark 3.4 (Limiting coderivative).** If  $\partial_M \text{Sup}_{J \circ S}(\bar{x}) \subset D^*[J \circ S](\bar{x}|\bar{v})(1)$  for  $\bar{v} = \text{Sup}_{J \circ S}(\bar{x})$ , such as when  $S$  is single-valued, and various qualification conditions hold, the sum and chain rules of [37, Theorems 3.10 and 3.13] can be used to derive expanded necessary optimality conditions. We do not pursue this route here, concentrating on optimalistic optimality conditions for the limiting coderivative.

**Remark 3.5 (Optimistic/pessimistic, necessary/sufficient, and limiting/Fréchet).** We note that with the limiting coderivative, we easily get necessary optimality conditions for the optimistic bilevel problem. The smaller Fréchet coderivative, on the other hand, seems to better correspond to the pessimistic bilevel problem, but – according to what we have been able to derive here – does give neither a necessary nor sufficient optimality condition, but a sufficient-for-necessary optimality condition. Under additional assumptions, it can be made either sufficient or necessary.

### 3.2 NONSMOOTH ADJOINT INCLUSIONS

The next lemma establishes for both  $\check{D}^* = \widehat{D}^*$  and  $\check{D}^* = D^*$  the *nonsmooth adjoint inclusion*

$$(-u^*, x^*) \in \text{ran } \check{D}^* G(u, x|0).$$

For the limiting coderivative this is necessary for  $x^* \in D^* S(x|u)(u^*)$ , subject to a qualification condition, while for the Fréchet coderivative it is only sufficient. In the latter case it provides a way to show that  $\widehat{D}^* S(x|u)(u^*)$  is non-empty, and thus to a suggestion that the sufficient-for-necessary condition (3.6) could be satisfied.

**Lemma 3.6.** *Suppose  $u \in S(x)$ . Then, for all  $u^* \in U^*$ ,*

$$\{x^* \in X^* \mid (-u^*, x^*) \in \text{ran } \widehat{D}^*G(u, x|0, 0)\} \subset \widehat{D}^*S(x|u)(u^*) \subset D^*S(x|u)(u^*).$$

If

$$(3.9) \quad D_M^*G^{-1}(0|u, x)(0) = \{0\},$$

then

$$D^*S(x|u)(u^*) \subset \{x^* \in X^* \mid (-u^*, x^*) \in \text{ran } D^*G(u, x|0)\}$$

*Proof.* By definition, we have  $x^* \in \widehat{D}^*S(x|u)(u^*)$  if and only if

$$(3.10) \quad \limsup_{x^k \rightarrow x, u^k \rightarrow u, G(u^k, x^k) \ni 0} \frac{\langle x^* | x^k - x \rangle - \langle u^* | u^k - u \rangle}{\|(x^k - x, u^k - u)\|} \leq 0.$$

Likewise  $(-u^*, x^*) \in \widehat{D}^*G(u, x|0)(w)$  for some  $w \in W$  if and only if

$$(3.11) \quad \limsup_{x^k \rightarrow x, u^k \rightarrow u, G(u^k, x^k) \ni w_*^k \rightarrow 0} \frac{-\langle w | w_*^k - 0 \rangle + \langle x^* | x^k - x \rangle - \langle u^* | u^k - u \rangle}{\|(x^k - x, u^k - u, w_*^k - 0)\|} \leq 0.$$

Restricting attention to  $w_*^k = 0$  in (3.11), it implies (3.10). This proves the first inclusion.

For the second inclusion, we use [37, Theorem 3.8] with  $\Theta = \{0\}$  and  $F = G$ . Since  $N_\Theta(0) = X^*$ , the qualification condition in the result requires that  $\ker \tilde{D}_M^*G(u, x|0) = \{0\}$  for the “reverse mixed coderivative”, defined for  $F : X \rightrightarrows Y$  as [37, (1.40)]  $\tilde{D}_M^*F(x|y)(y^*) := \{x^* \in X^* \mid y^* \in -D_M^*F^{-1}(y|x)(-x^*)\}$ . That is, the qualification condition reads  $w_* \in -D_M^*G^{-1}(0|u, x)(0) \implies w_* = 0$  with  $0 \in -D_M^*G^{-1}(0|u, x)(0)$ . This can be equivalently be written (3.9). Noting that  $\text{Graph } S^{-1} = G^{-1}(\Theta)$ , that  $\Theta$  is partially sequentially normally compact, and  $(u, x) \mapsto G(u, x) \cap \Theta \subset \{0\}$  is inner semicompact,<sup>2</sup> it follows that

$$\begin{aligned} D^*S(x|u)(u^*) &= \{x^* \mid (x^*, -u^*) \in N_{\text{Graph } S}(x, u)\} = \{x^* \mid (-u^*, x^*) \in N_{\text{Graph } S^{-1}}(u, x)\} \\ &= \{x^* \mid (-u^*, x^*) \in N_{G^{-1}(\Theta)}(u, x)\} \\ &\subset \{x^* \mid (-u^*, x^*) \in D^*G(u, x|w)(w^*), w^* \in N_\Theta(w), w \in G(u, x) \cap \Theta\} \\ &= \{x^* \in X^* \mid (-u^*, x^*) \in \text{ran } D^*G(u, x|0)\}. \end{aligned} \quad \square$$

**Remark 3.7.** The “mixed” limiting coderivative is used in (3.9). When  $Y^*$  is finite-dimensional, and the “mixed” and “normal” limiting coderivatives agree, the condition can be written as

$$0 \in D^*G(u, x|0)(w) \implies w = 0 \quad \text{i.e.} \quad \ker D^*G(u, x|0) = \{0\}.$$

**Corollary 3.8.** *Suppose  $u \in S(x)$ . If the qualification condition (3.9) holds and  $G$  is  $N$ -regular at  $(u, x)$  for 0, then  $S$  is  $N$ -regular at  $x$  for  $u$ , and we have*

$$\widehat{D}^*S(x|u)(u^*) = \{x^* \in X^* \mid (-u^*, x^*) \in \text{ran } \widehat{D}^*G(u, x|0)\}.$$

The following example demonstrates that the non-smooth (optimality) condition  $0 \in \check{D}^*S(x|u)(J'(u))$  generalises the smooth one, e.g. [44].

<sup>2</sup>There are inconsistencies in the definition of inner semicompactness in [37]. According to Definition 1.63 therein,  $S : X \rightrightarrows Y$  is inner semicompact at  $\bar{x}$  if for every  $x^k \rightarrow \bar{x}$ , there exist  $y^k \in S(x^k)$  and a convergent subsequence  $\{y^{k_\ell}\}_{\ell \in \mathbb{N}}$ . Our mapping does not satisfy this definition. However, in practise in the proofs,  $x^k \in \text{dom } S$ . Our mapping satisfies this domain-restricted definition. Indeed, in the proof of [37, Theorem 3.8], one takes  $x^k \in F^{-1}(\Theta)$ , which guarantees the existence of  $y^k \in \Theta \cap F(x^k)$ . In our case,  $y^k \equiv 0$ .

**Example 3.9.** Suppose  $G$  is single-valued and Fréchet differentiable. Then by, e.g., [13, Theorem 20.12], when  $0 = G(u, x)$ , i.e.  $u \in S(x)$ , we have, for  $\check{D}^* = \widehat{D}^*, D^*$ ,

$$\check{D}^*G(u, x|0)(w) = \{G'(u, x)^*w\} = \{(G^{(u)}(u, x)^*w, G^{(x)}(u, x)^*w)\}.$$

Consequently, by the lemma, for any  $w \in W$  and

$$u^* = -G^{(u)}(u, x)^*w \quad \text{and} \quad x^* = G^{(x)}(u, x)^*w,$$

we have  $x^* \in \check{D}^*S(x|u)(u^*)$ . In particular, if  $G^{(u)}(u, x)^*$  is invertible, for arbitrary  $u^*$ , solving  $w = G^{(u)}(u, x)^{-1}u^*$ , we have  $-G^{(x)}(u, x)^*G^{(u)}(u, x)^{-1}u^* \in \check{D}^*S(x|u)(u^*)$ .

### 3.3 FULL SYSTEM OPTIMALITY CONDITIONS

We can now write our full-system optimality conditions. We start with a sufficient-for-pessimistic necessary condition based on the Fréchet coderivative. For simplicity, we now always take  $J$  continuously differentiable.

**Theorem 3.10.** Consider  $F := J \circ S + R$ , where  $J, S$ , and  $R$  are as in (3.1) and (3.2) with  $J$  continuously differentiable. Then the pessimistic necessary optimality condition  $0 \in \partial_F[\text{Sup}_{J \circ S} + R](\bar{x})$  for the problem (3.1) holds if  $S$  is inner Lipschitz at  $\bar{x}$  relative to  $\{\bar{u}\}$ , and

$$(3.12) \quad (-J'(\bar{u}), \bar{x}^*) \in \text{ran } \widehat{D}^*G(\bar{u}, \bar{x}|0) \quad \text{with} \quad 0 \in \bar{x}^* + \partial_F R(\bar{x}) \quad \text{and} \quad \bar{u} \in S^b(\bar{x}).$$

*Proof.* Combine Theorem 3.3 (i) with Lemma 3.6. □

We next provide a necessary optimistic optimality condition for the problem (3.1).

**Theorem 3.11.** Consider  $F := J \circ S + R$ , where  $J, S$ , and  $R$  are as in (3.1) and (3.2) with  $J$  continuously differentiable. Suppose  $\text{Inf}_{J \circ S} + R$  is minimised at  $\bar{x} \in X$ , and  $\bar{u} \in S^\sharp(\bar{x})$ . If  $R$  is Lipschitz differentiable (or, more generally  $\partial_{|\text{dom} R} R \neq \emptyset$ ) with  $\text{dom } R = X$ , and the qualification condition (3.9) holds, then

$$(3.13) \quad (-J'(\bar{u}), \bar{x}^*) \in \text{ran } D^*G(\bar{u}, \bar{x}|0) \quad \text{with} \quad 0 \in \bar{x}^* + \partial_F R(\bar{x}) \quad \text{and} \quad \bar{u} \in S^\sharp(\bar{x})$$

*Proof.* Combine Corollary 3.2 with Lemma 3.6. □

To work with either the Fréchet coderivative or the limiting coderivative, we now recall the notation  $\check{D}^*$  to refer to either  $\widehat{D}^*$  or  $D^*$ .

**Remark 3.12.** Consider for  $\check{D}^* = D^*, \widehat{D}^*$  and  $\square = \sharp, b$ , the sets

$$\begin{aligned} \check{d}_1^\square F(x) &:= \{x^* \in X^* \mid x^* \in \check{D}^*S(x|u)(J'(u)), u \in S^\square(x)\} \quad \text{and} \\ \check{d}_2^\square F(x) &:= \{x^* \in X^* \mid (-J'(u), x^*) \in \text{ran } \check{D}^*G(u, x|0), u \in S^\square(x)\}. \end{aligned}$$

The following table indicates our results on when  $0 \in [\mathcal{d}F + \partial R](x)$  for  $\mathcal{d} = \check{d}_i^\square$  is a meaningful optimality condition for the problem (3.1) and (3.2).

		Fréchet		Limiting	
		$\widehat{d}_1^\square F$	$\widehat{d}_2^\square F$	$d_1^\square F$	$d_2^\square F$
suff.-for-nec. pessimistic	$b$	Theorem 3.3	Theorem 3.10		
necessary optimistic	$\sharp$	Theorem 3.1		Corollary 3.2	Theorem 3.11

Practically, we will relax  $u \in S^\sharp(x)$  and  $u \in S^b(x)$  to  $u \in S(x)$ , i.e.,  $0 \in G(u, x)$ , and work with

$$(3.14) \quad \check{d}_2 F(x) := \{x^* \in X^* \mid (-J'(u), x^*) \in \text{ran } \check{D}^* G(u, x|0), u \in S(x)\}$$

that contains both  $\check{d}_2^\sharp F(x)$  and  $\check{d}_2^b F(x)$ . Then  $0 \in [\check{d}_2 F + \partial R](x)$  can also be written as

$$(3.15) \quad 0 \in \check{H}(u, x, x^*, w) := \begin{pmatrix} G(u, x) \\ \check{D}^* G(u, x|0)(w) + (J'(u), -x^*) \\ x^* + \partial R(x) \end{pmatrix}.$$

For the limiting coderivative, subject to the conditions of [Theorem 3.11](#), this still remains a necessary optimalistic optimality condition. If  $S(x)$  is a singleton, then the version with the Fréchet coderivative provides a sufficient-for-necessary optimality condition. In the next section, we will build algorithms to solve this system based on combining standard algorithms for each component inclusion.

**Remark 3.13.** For smooth functions, (3.15) differs from [43, 44] through a different lower-dimensional adjoint variable. Indeed, (3.15) directly gives the dimension-reduced adjoint inclusion as used in [18, Example 2.1] for bilevel optimisation, or classically derived for PDE-constrained optimisation in, e.g., [27, §1.6.2] or [15, §1.2]. It directly calculates  $x^*$  that corresponds to  $J'(S(x))S'(x)$  in the smooth case, instead of first, separately, calculating  $p = S'(x)$  as in [43, 44].

**Remark 3.14 (Proximal optimality conditions in Hilbert spaces).** Suppose  $R$  is convex, proper, and lower semicontinuous, and  $X$  is a Hilbert space. Then, following standard results, [13, Lemma 6.22], we can for any  $\tau > 0$  reinterpret (3.12) as

$$\bar{x} = \text{prox}_{\tau R}(\bar{x} - \tau \mathcal{R} \bar{x}^*) \quad \text{for some } \bar{x}^* \in \check{d}_2^b F(\bar{x}),$$

and (3.13) as

$$\bar{x} = \text{prox}_{\tau R}(\bar{x} - \tau \mathcal{R} \bar{x}^*) \quad \text{for some } \bar{x}^* \in d_2^\sharp F(\bar{x}).$$

Likewise, (3.5) and (3.5) can be given the proximal form in terms of  $\check{d}_1^\square F(x)$ .

## 4 ALGORITHM

We now build a general approach to single-loop algorithms for (1.1), based on the optimality conditions (3.15). First in [Section 4.1](#), we lay out the approach. Then in [Section 4.2](#), we recall the *tracking theory* of [18] for the construction of single-loop algorithms based on differential estimation. This involves so-called *inner tracking*, *adjoint tracking*, and *differential transformation* conditions, which we state in [Assumption 4.3](#), and further interpret in [Sections 4.3 to 4.5](#). In [Section 4.6](#) we then show how these components give convergence to an outer forward-backward method.

### 4.1 AN GENERAL SINGLE-LOOP ALGORITHM

Suppose  $U, W_*$  and  $X$  are normed spaces, and denote  $W = (W_*)^*$ . Let  $\check{D}^* = \widehat{D}^*$  or  $\check{D}^* = D^*$ . We assume that  $J : U \rightarrow \mathbb{R}$  is continuously differentiable and  $R : X \rightarrow \mathbb{R}$  convex. We divide the solution of (3.15) into the following rough components:

1. an *inner algorithm*, which, on each iteration  $k \in \mathbb{N}$ , given the history of (outer, inner, adjoint, approximate subderivative) iterates  $\{(x^j, u^j, w^j, \tilde{x}_j^*)\}_{j=0}^k$ , produces  $u^{k+1}$  that approximately solves the first line, i.e.,

$$(4.1a) \quad 0 \approx \Delta w_*^{k+1} \in G(u^{k+1}, x^k);$$

2. an *adjoint algorithm*, which, given the previous iterates and  $u^{k+1}$ , produces  $w^{k+1}$  and  $\tilde{x}_{k+1}^*$  that approximately solve the second line, i.e., for  $\check{D}^* = \widehat{D}^*$  or  $\check{D}^* = D^*$ ,

$$(4.1b) \quad 0 = \check{D}^* G(u^{k+1}, x^k | \Delta w_*^{k+1})(w^{k+1}) + (J'(u^{k+1}) - \Delta u_{k+1}^*, -\tilde{x}_{k+1}^*) \quad \text{for some } \Delta u_{k+1}^* \approx 0;$$

3. an *outer algorithm*, which, given the previous iterates, produces  $x^{k+1}$  that approximately solves the third line, i.e.,

$$(4.1c) \quad 0 \approx \Delta x_{k+1}^* \in \tilde{x}_{k+1}^* + \partial R(x^{k+1}).$$

The variable  $\tilde{x}_{k+1}^*$  estimates some element of  $\check{\partial} \text{Inf}_{J \circ S}(x^k)$  or  $\check{\partial} \text{Sup}_{J \circ S}(x^k)$ . Adopting the terminology of [18], we call it a *differential estimate*, despite possible lack of differentiability. The perturbations  $\Delta w_*^{k+1} \approx 0$ ,  $\Delta u_{k+1}^* \approx 0$ , and  $\Delta x_{k+1}^* \approx 0$  arise from different algorithms for the realisation of these steps, as demonstrated in the following examples. We will make the allowed level of perturbation more precise in [Assumption 4.3](#). In (4.1b), for simplicity, and for reasons that will become apparent in [Section 5](#), we only include inexactness in the first component. We could also include a perturbation to  $\tilde{x}_{k+1}^*$ , but, in fact,  $\tilde{x}_{k+1}^*$  is already an estimate of  $x_{k+1}^* \in \check{\partial}_2 F(x^k)$  that solves the steps with  $\Delta w_*^{k+1} = 0$  and  $\Delta u_{k+1}^* = 0$ , i.e., the first two lines of (3.15).

**Example 4.1 (Outer forward-backward splitting).** Assuming  $X$  to be Hilbert, for a forward-backward algorithm, as an instance of (4.1c) we would solve  $-(x^{k+1} - x^k)/\tau \in \mathcal{R}[\tilde{x}_{k+1}^* + \partial R(x^{k+1})]$  for a step length parameter  $\tau > 0$ . Here we recall that  $\mathcal{R} \in \mathbb{L}(X^*; X)$  is the Riesz map. This corresponds to

$$x^{k+1} = \text{prox}_{\tau R}(x^k - \tau \mathcal{R} \tilde{x}_{k+1}^*).$$

**Example 4.2 (Inner primal-dual proximal splitting, PDPS).** Let  $U_p, U_d$  and  $X$  be Hilbert spaces,  $f := f_0 + e$  for  $f_0, e : U_p \times X \rightarrow \overline{\mathbb{R}}$  and  $g : U_d \times X \rightarrow \overline{\mathbb{R}}$  that are convex, proper, and lower semicontinuous in their first parameter. Moreover, let  $e$  have  $L$ -Lipschitz gradient with respect to its first parameter and  $K \in \mathbb{L}(U_p; U_d)$ . We consider the problem

$$\min_u f(u, x) + g(Ku; x),$$

which can be equivalently written as the saddle point problem

$$\min_{u_p \in U_p} \max_{u_d \in U_d^*} f(u_p; x) + \langle u_d | Ku_p \rangle - g^*(u_d; x).$$

Following [26, 13, Theorem 5.11] and writing  $u = (u_p, u_d) \in U_p \times U_d^*$ , this problem has primal-dual optimality conditions of the form  $0 \in G(u, x)$  with

$$(4.2) \quad G(u; x) := \begin{pmatrix} \partial_{u_p} f_0(u_p; x) + \nabla_{u_p} e(u_p; x) + K^* u_d \\ \partial_{u_d} g^*(u_d; x) - Ku_p \end{pmatrix}.$$

Defining for step length parameters  $\tau_x, \tau_y > 0$  the preconditioning operator

$$(4.3) \quad M := \begin{pmatrix} \tau_p^{-1} \text{Id} & -K^* \\ -K & \tau_d^{-1} \text{Id} \end{pmatrix},$$

and taking  $\Delta w_*^{k+1} = -M(u^{k+1} - u^k)$ , the inner algorithm (4.1a) then becomes the primal-dual proximal splitting (PDPS) method of [10], here with an additional forward step with respect to  $e$ :

$$(4.4) \quad \begin{cases} u_p^{k+1} &= \text{prox}_{\tau_p f_0(\cdot; x)}(u_p^k - \tau_p(K^* u_d^k + \nabla_{u_p} e(u_p^k; x))), \\ u_d^{k+1} &= \text{prox}_{\tau_d g^*(\cdot; x)}(u_d^k + \tau_d K(2u_p^{k+1} - u_p^k)). \end{cases}$$

The method converges (for fixed  $x$ ) if  $\tau_p L/2 + \tau_p \tau_d \|K\|^2 \leq 1$ ; see, e.g., [13].

## 4.2 TRACKING INEQUALITIES

We next discuss conditions that the inner and adjoint algorithms (4.1a) and (4.1b) have to satisfy, for us to be able to prove the convergence of standard outer methods. We assume that the inner and adjoint algorithms satisfy the following “tracking inequalities”, adapted from [18, 44]. The idea in [18] is to construct single-loop estimates  $\tilde{x}_{k+1}^*$  of differentials  $F'(x^k)$  of smooth functions  $F = J \circ S$ . These tracking inequalities ensure that the accumulated errors remain controlled and that the approximate differential remains meaningful for descent.

To adapt the theory to our nonsmooth and multi-valued case, we replace  $F'(x^k)$  by a “target”  $x_{k+1}^*$  in one of the sets  $\check{\mathcal{A}}_i$  of Remark 3.12, modelling Fréchet or limiting subdifferentials or their relaxations, depending on the exactness of the calculus rules used in their derivation. Thus,  $0 \in x_{k+1}^* + \partial R(x^k)$  models a notion of (relaxed) criticality of  $x^k$  for the problem  $\min(F + R)$ . We estimate this target by  $\tilde{x}_{k+1}^*$  as generated by the general algorithm (4.1). We also work with single-valued selections  $S_u(x^k)$  of the inner solution mapping  $S(x^k)$ , and a single-valued selection  $S_w(x^k)$  of the adjoint solution mapping  $\check{Z}(S_u(x^k), x^k, 0)$ , where we set

$$(4.5) \quad \check{Z}(u, x, w_*) := \{w \in W \mid 0 \in \check{D}^* G(u, x | w_*)(w) + (J'(u), -x^*), x^* \in X^*\} \quad \text{when } w_* \in G(u, x).$$

**Assumption 4.3.** Let  $X, X^*, U$ , and  $W_*$  be normed spaces,  $W = (W_*)^*$ , and both  $\Omega_X \subset X$  and  $\Omega_U \subset U$ . For some norms  $\|\cdot\|_*$  in  $W_*$  and  $\|\cdot\|_{**}$  in  $U^*$ , define

$$(4.6a) \quad d_{k+1}^u := \|u^{k+1} - S_u(x^k)\|_U + \|\Delta w_*^{k+1}\|_*,$$

$$(4.6b) \quad d_{k+1}^w := \|w^{k+1} - S_w(x^k)\|_W + \|\Delta u_{k+1}^*\|_{**}, \quad \text{and}$$

$$(4.6c) \quad \varrho_k := \|x^k - x^{k-1}\|_X.$$

Then on each iteration  $k \geq 0$ , given  $\{(x^n, u^n, w^n)\}_{n=0}^k \subset \Omega_X \times \Omega_U \times W$  from the previous iterations, we are given:

- (i) **(Inner tracking property)** An *inner algorithm* that produces  $u^{k+1} \in U$  satisfying, when  $k \geq 1$ , for some  $\pi_u > 0$  and  $\kappa_u > 1$  that

$$\kappa_u d_{k+1}^u \leq d_k^u + \pi_u \varrho_k.$$

- (ii) **(Adjoint tracking property)** An *adjoint algorithm* that produces  $w^{k+1} \in W$  satisfying, when  $k \geq 1$ , for some  $\mu_w, \pi_w > 0$  and  $\kappa_w > 1$  that

$$\kappa_w d_{k+1}^w \leq d_k^w + \mu_w d_{k+1}^u + \pi_w \varrho_k.$$

- (iii) **(Differential transformation)** A *differential transformation* that produces  $\tilde{x}_{k+1}^* \in X^*$  that, for a target  $x_{k+1}^* \in X^*$ , satisfies for some  $\alpha_u, \alpha_w \geq 0$  the bound

$$\|\tilde{x}_{k+1}^* - x_{k+1}^*\|_{X^*} \leq \alpha_u d_{k+1}^u + \alpha_w d_{k+1}^w.$$

The condition (i) formalises what the rough inner algorithm (4.1a) has to satisfy, and (ii) and (iii) formalises what the rough adjoint algorithm (4.1b) has to satisfy. The inner and adjoint tracking conditions (i) and (ii) are parameter change aware contractivity conditions for the inner and adjoint algorithms: if  $x^k = x^{k-1}$ , the former reduces to a standard contractivity condition. Both are trivial for exact inner and adjoint solutions  $u^{k+1} = S_u(x^k)$  and  $w^{k+1} = S_w(x^k)$ . The condition (iii) allows converting the construction error of  $\tilde{x}_{k+1}^*$  to the tracking errors of the inner and adjoint algorithms.

We next discuss how Assumption 4.3 can be satisfied, then in Theorem 4.18 we show how it can be used to prove the convergence of an outer optimisation method (4.1c), especially the forward-backward splitting of Example 4.1. The proof depends on scalar tracking results from [18], quoted in Appendix C.

**Remark 4.4 (Exact solutions and iterative solutions).** Both the adjoint and inner tracking properties Assumption 4.3 (i) and (ii) are always satisfied if we solve the adjoint and inner inclusion exactly, as then  $u^{k+1} \in S(x^k)$  and  $w^{k+1} \in Z(x^k)$  so we can take  $S_u(x^k) = u^{k+1}$  and  $S_w(x^k) = w^{k+1}$ . Given contractive inner and adjoint algorithms for the fixed parameter  $x^k$ , we can also for any  $\pi_u, \pi_w > 0$  take sufficiently many iterations of those algorithms, until Assumption 4.3 (i) and (ii) are satisfied.

**Remark 4.5 (Relaxed distances).** It is possible to relax the distance in  $X$  to any semi-norm, in which case the norm  $\|\cdot\|_{X^*}$  in (iii) is replaced by its support function. We refer to [18] for details.

### 4.3 INNER TRACKING PROPERTY

We now provide simpler component conditions to satisfy the inner tracking Assumption 4.3 (i). We split it into a regularity property of the solution mapping together with the simpler non-parametric **inner algorithm contractivity**

$$(4.7) \quad \kappa_u \operatorname{dist}(u^{k+1}, S(x^k)) + \kappa_u \|\Delta w_*^{k+1}\|_* \leq \operatorname{dist}(u^k, S(x^k)) \quad \text{for a } \kappa_u > 1$$

for some (semi-)norm  $\|\cdot\|_*$  on  $W_*$ . In practical algorithms it is common to have

$$\kappa_u \|\Delta w_*^{k+1}\|_* = \lambda \|u^{k+1} - u^k\|_M$$

for some  $\lambda > 0$  and a self-adjoint positive (semi-)definite operator  $M$ . This follows from the following theorems for forward-backward splitting and primal-dual proximal splitting. In both results, the first factor  $\kappa_u = 2\sqrt{(1+\rho)/(2+\alpha)} > 1$  if  $0 < \alpha < 2$ . The second factor may be small, but can be absorbed into the factor  $\lambda$  in the definition of the auxiliary norm  $\|\cdot\|_*$  employed in (4.7).

We recall from [1] (a theme also treated in [13, §28.2]) that convex subdifferentials are metrically subregular when they satisfy a version of strong local subdifferentiability where the growth term is relaxed to be with respect to a set, instead of a single point.

**Theorem 4.6 (Contractivity of forward-backward splitting under metric subregularity).** *For convex functions  $g, f : U \rightarrow \overline{\mathbb{R}}$  on a Hilbert space  $U$ , where  $f$  has an  $L$ -Lipschitz gradient, suppose  $T := \partial g + \nabla f$  is metrically subregular at a  $\hat{u} \in T^{-1}(0)$  for  $\hat{w} = 0$  with the factor  $\kappa > 0$  and radius  $\delta > 0$ . Given  $u^k \in B(\hat{u}, \delta)$ , let  $u^{k+1} := \operatorname{prox}_{\tau g}(u^k - \tau \nabla f(u^k))$  be generated through the forward-backward method with  $\tau > 0$  satisfying  $L\tau \leq 2(1 - \varepsilon)$  for an  $\varepsilon > 0$ . Then  $u^{k+1} \in B(\hat{u}, \delta)$ , and, for  $\rho = (1 - L\tau/2 - \varepsilon)\tau^2 / (2\kappa^2(1 + L^2\tau^2)) > 0$  and any  $\alpha > 0$ , we have*

$$(4.8) \quad 2\sqrt{\frac{1+\rho}{2+\alpha}} \operatorname{dist}(u^{k+1}, T^{-1}(0)) + 2\sqrt{\frac{\varepsilon}{2+\alpha^{-1}}} \|u^{k+1} - u^k\| \leq \operatorname{dist}(u^k, T^{-1}(0)).$$

See proof in Appendix A.

**Theorem 4.7 (Contractivity of primal-dual proximal splitting under metric subregularity).** Let  $G$  and  $M$  be as in [Example 4.2](#). For a fixed  $x$ , suppose  $T = G(\cdot; x)$  is metrically subregular with respect to the  $M$  and  $M^{-1}$ -norms at  $\widehat{u}$  for  $\widehat{w} = 0$  with the modulus  $\kappa_M > 0$  and radius  $\delta > 0$ , i.e.,

$$\text{dist}_M(u, T^{-1}(0)) \leq \kappa_M \text{dist}_{M^{-1}}(0, T(u)) \quad (u \in B_M(\widehat{u}, \delta)).$$

Given  $u^k = (u_p^k, u_d^k) \in B_M(\widehat{u}, \delta)$ , let  $u^{k+1}$  be generated through the PDPS [\(4.4\)](#) with  $\tau_p, \tau_d > 0$  satisfying  $\frac{1}{2(1-\varepsilon)}L\tau_p + \tau_p\tau_d\|K\|^2 \leq 1$  for an  $\varepsilon \in (0, 1)$ . Then  $u^{k+1} \in B_M(\widehat{u}, \delta)$  and

$$2\sqrt{\frac{1+\rho}{2+\alpha}} \text{dist}_M(u^{k+1}, T^{-1}(0)) + 2\sqrt{\frac{\varepsilon}{2+\alpha^{-1}}} \|u^{k+1} - u^k\|_M \leq \text{dist}_M(u^k, T^{-1}(0))$$

for  $\rho = (1 - L_M/2 - \varepsilon)/(2(1 + L_M^2)\kappa_M^2) > 0$ ,  $L_M = L\tau_p/(1 - \tau_p\tau_d\|K\|^2)$  and any  $\alpha > 0$ .

See [proof in Appendix A](#).

**Remark 4.8 (Compatibility of neighbourhoods and moduli).** To apply [Theorems 4.6](#) and [4.7](#) (with  $T = G(\cdot, x^k)$ , hence  $T^{-1}(0) = S(x^k)$ ) to proving [\(4.7\)](#), it should be verified, in a case-by-case basis for each problem of interest that:

- (i) The moduli of metric subregularity, hence the factors  $\kappa_u$ , can be chosen uniformly over  $k \in \mathbb{N}$ .
- (ii) The neighbourhoods of metric subregularity of  $G(\cdot, x^k)$  and  $G(\cdot, x^{k+1})$  are compatible:  $u^{k+1} \in B(\widehat{u}^k, \delta_k)$  guarantees  $u^{k+1} \in B(\widehat{u}^{k+1}, \delta_{k+1})$ , where  $0 \in G(\widehat{u}^k, x^k)$  and  $\delta_k > 0$  is the corresponding radius of metric subregularity.

The theorems guarantee that  $u^{k+1} \in B(\widehat{u}^k, \delta_k)$  when  $u^k \in B(\widehat{u}^k, \delta_k)$ , but since the parametric problem changes on each iteration, the neighbourhood of metric subregularity may change. Obviously, it is enough that the neighbourhoods are global,  $\delta_k \equiv \infty$ , such as when  $G(\cdot, x^k)$  is strongly monotone with a uniform factor; see [Lemmas A.3](#) and [A.4](#).

We can now split the inner tracking property into several simpler conditions: the inner algorithm contractivity that we have just treated, a Lipschitz-like condition on the inner solution mapping, and the optimality of the inner selection.

**Theorem 4.9.** Assume the setup of [Section 4.1](#). Let  $u^{k+1} \in U$  generated by [\(4.1a\)](#) satisfy the **inner algorithm contractivity** [\(4.7\)](#). Suppose the **inner solution mapping one-sided Lipschitz-like property**

$$(4.9) \quad \text{dist}(S(x^k), u^k) \leq \text{dist}(S(x^{k-1}), u^k) + \pi_u \|x^{k-1} - x^k\|_X$$

for some constants  $\pi_u > 0$ ,  $\kappa_u > 1$ , and that  $S_u(x^k) \in S(x^k)$  is such that this **inner selection is optimal**,

$$\text{dist}(S(x^k), u^{k+1}) = \|u^{k+1} - S_u(x^k)\|_U \quad \text{for all } k \in \mathbb{N}.$$

Then [Assumption 4.3 \(i\)](#) holds.

*Proof.* For the proof, recall the abbreviations [\(4.6\)](#). We combine [\(4.7\)](#) and [\(4.9\)](#) to obtain

$$\begin{aligned} \kappa_u d_{k+1}^u &= \kappa_u \|u^{k+1} - S_u(x^k)\|_U + \kappa_u \|\Delta w_*^{k+1}\|_* \\ &\leq \text{dist}(u^k, S(x^k)) \\ &\leq \text{dist}(S(x^{k-1}), u^k) + \pi_u \|x^{k-1} - x^k\|_X \\ &= \|u^k - S_u(x^{k-1})\|_U + \pi_u \|x^{k-1} - x^k\|_X \\ &\leq \|u^k - S_u(x^{k-1})\|_U + \|\Delta w_*^k\|_* + \pi_u \|x^{k-1} - x^k\|_X \\ &= d_k^u + \pi_u Q_k. \end{aligned}$$

□

It can be shown that the Pompeii–Hausdorff distance between sets  $A$  and  $B$ ,

$$\text{dist}(A, B) = \sup_{u \in U} |\text{dist}(A, u) - \text{dist}(B, u)|.$$

Therefore, in the worst case, if the inner iterates  $u^k$  are unbounded, (4.9) becomes a Pompeii–Hausdorff–Lipschitz property. However, usually, we can assume that  $u^k$  belongs to a bounded set  $\Omega_U$ , relaxing the requirement. The next lemma shows that (4.9) arises from the Aubin property of  $S$  in a suitable set. However – in the bounded setting – it is conceivable that there are cases where (4.9) holds but the Aubin property of  $S$  does not hold.

**Lemma 4.10.** *Suppose  $S : X \rightrightarrows U$  has the Aubin property in  $\Omega_X$  and  $\Omega_U$  with the factor  $\pi_u > 0$  in the form<sup>3</sup>*

$$\text{dist}(S(x), \tilde{u}) \leq \pi_u \text{dist}(S^{-1}(\tilde{u}) \cap \Omega_X, x) \quad \text{for all } x \in \Omega_X \text{ and } \tilde{u} \in \Omega_U,$$

then, for any  $u \in U$ ,  $x, \tilde{x} \in \Omega_X$ ,

$$\text{dist}(S(x), u) \leq \text{dist}(S(\tilde{x}) \cap \Omega_U, u) + \pi_u \|x - \tilde{x}\|.$$

*Proof.* Expanding the definitions of  $\text{dist}$ , we have

$$\begin{aligned} d &:= \text{dist}(S(x), u) - \text{dist}(S(\tilde{x}) \cap \Omega_U, u) = \sup_{\tilde{u} \in S(\tilde{x}) \cap \Omega_U} \inf_{u' \in S(x)} \|u' - u\|_U - \|\tilde{u} - u\|_U \\ &\leq \sup_{\tilde{u} \in S(\tilde{x}) \cap \Omega_U} \inf_{u' \in S(x)} \|u' - \tilde{u}\|_U = \sup_{\tilde{u} \in S(\tilde{x}) \cap \Omega_U} \text{dist}(S(x), \tilde{u}). \end{aligned}$$

Now, since  $x \in \Omega_X$ , and  $S(\tilde{x}) \cap \Omega_U \subset \Omega_U$ , the Aubin property gives

$$d \leq \sup_{\tilde{u} \in S(\tilde{x}) \cap \Omega_U} \pi_u \text{dist}(S^{-1}(\tilde{u}) \cap \Omega_X, x) \leq \pi_u \|\tilde{x} - x\|. \quad \square$$

**Remark 4.11 (Single-valued Lipschitz solution mappings).** In particular, if  $S$  is single-valued and Lipschitz, then (4.9) holds. It is not uncommon that a primal-only inner solution mapping is single-valued and Lipschitz, compare [46, Theorem 2] and [8, Theorem 4.51], or combine [13, Theorem 28.6] with a strict convexity assumption.

#### 4.4 ADJOINT TRACKING PROPERTY

Similarly to the inner tracking property, we split the adjoint tracking property into regularity properties of the solution mappings together with the simpler non-parametric contractivity condition on the algorithm. Recall the definition of  $\check{Z}$  in (4.5).

**Theorem 4.12.** *Assume the setup of Section 4.1. Let  $\check{Z} : U \times X \times W_* \rightrightarrows W$ , abbreviating  $Z(x) := \check{Z}(S_u(x), x, 0)$ , satisfy for some  $\pi_w, \mu_u > 0$  and  $\kappa_w > 1$  the **adjoint solution mapping one-sided bi-Lipschitz-like property***

$$\begin{aligned} (4.10) \quad \kappa_w \text{dist}(Z(x^k), w^{k+1}) - \text{dist}(Z(x^{k-1}), w^k) \\ \leq \kappa_w \text{dist}(\check{Z}(u^{k+1}, x^k, \Delta w_*^{k+1}), w^{k+1}) - \text{dist}(\check{Z}(u^{k+1}, x^k, \Delta w_*^{k+1}), w^k) \\ + \pi_w \|x^k - x^{k-1}\| + \mu_u \left( \|u^{k+1} - S_u(x^k)\|_U + \|\Delta w^{k+1}\|_* \right). \end{aligned}$$

<sup>3</sup>For spherical neighbourhoods  $\Omega_X = B(\bar{x}, \delta)$ , the equivalence of this form to the the standard form  $\text{dist}(S(x), \tilde{u}) \leq \pi_u \text{dist}(S^{-1}(\tilde{u}), x)$  is proved in [13, Theorem 27.3]. This latter superficially stricter form always implies the one we assume.

Moreover, suppose  $w^{k+1} \in W$  generated by (4.1b) satisfy **adjoint algorithm contractivity**

$$(4.11) \quad \kappa_w \operatorname{dist}(\check{Z}(u^{k+1}, x^k, \Delta w_*^{k+1}), w^{k+1}) + \kappa_w \|\Delta u_{k+1}^*\|_{**} \leq \operatorname{dist}(\check{Z}(u^{k+1}, x^k, \Delta w_*^{k+1}), w^k)$$

and that the **adjoint selection is optimal**, i.e.,  $S_w(x^k) \in Z(x^k)$  is such that

$$(4.12) \quad \operatorname{dist}(Z(x^k), w^{k+1}) = \|w^{k+1} - S_w(x^k)\|_W \quad \text{for all } k \in \mathbb{N}.$$

Then **Assumption 4.3 (ii)** holds.

*Proof.* For the proof, recall the abbreviations (4.6). Combining (4.10) with (4.11), we obtain

$$\kappa_w \|\Delta u_{k+1}^*\|_{**} + \kappa_w \operatorname{dist}(Z(x^k), w^{k+1}) - \operatorname{dist}(Z(x^{k-1}), w^k) \leq \pi_w \|x^k - x^{k-1}\| + \mu_u d_{k+1}^u.$$

Using (4.6b), (4.6c) and (4.12) we find this equivalent to  $\kappa_w d_{k+1}^w \leq d_k^w + \mu_u d_{k+1}^u + \pi_w \varrho_k$ , i.e., **Assumption 4.3 (ii)**.  $\square$

As in the case of (4.9), if  $w^k$  and  $w^{k+1}$  are unbounded, the respective distance differences in (4.10) can be reformulated as Pompeii–Hausdorff distances. Then the property holds if  $\check{Z}$  is Pompeii–Hausdorff–Lipschitz, and the inner selection  $S_u$  is Lipschitz. Next lemma shows, that, if the latter holds, then similarly to **Lemma 4.10**, the condition (4.10) can also be verified through two applications of the Aubin property. However, there are cases where condition (4.10) holds but the assumptions of next lemma does not hold. We encounter such an example in **Section 4.7**. In particular, the Lipschitz assumption on  $S_u$ , where  $S_u(x) \subset S(x)$ , in practise forces  $S$  to be single-valued. This is not the case with (4.10).

**Lemma 4.13.** *Suppose  $U, W_*$  and  $X$  are normed spaces,  $W = (W_*)^*$ . Let  $J : U \rightarrow \mathbb{R}$  and  $G : U \times X \rightrightarrows W_*$ . Moreover, let  $S_u : X \rightarrow U$  be  $L$ -Lipschitz and satisfy  $0 \in G(S_u(x), x)$  for all  $x \in X$ ,  $\check{Z} : U \times X \times W_* \rightrightarrows W$  be given by (4.5) and abbreviate  $Z(x) := \check{Z}(S_u(x), x, 0)$ . Assume  $\check{Z}$  has the Aubin property in  $\Omega := \{(u, x, w_*) \in \Omega_U \times \Omega_X \times W_* \mid w_* \in G(u, x)\}$  and  $\Omega_W$  with the factor  $C_w > 0$  in the form*

$$\operatorname{dist}(\check{Z}(u, x, w_*), \tilde{w}) \leq C_w \operatorname{dist}(\check{Z}^{-1}(\tilde{w}) \cap \Omega, (u, x, w_*)) \quad \text{for all } (u, x, w_*) \in \Omega \text{ and } \tilde{w} \in \Omega_W.$$

Then, for any  $w, \bar{w} \in W, u \in \Omega_U, \tilde{x} \in \Omega_X, x \in \Omega_X \cap S_u^{-1}(\Omega_U), \kappa_w > 0$  and  $\pi_w := C_w(1 + L), \mu_u := C_w(1 + \kappa_w)$

$$(4.13) \quad \begin{aligned} \kappa_w \operatorname{dist}(Z(x), w) - \operatorname{dist}(Z(\tilde{x}) \cap \Omega_W, \bar{w}) \\ \leq \kappa_w \operatorname{dist}(\check{Z}(u, x, w_*) \cap \Omega_W, w) - \operatorname{dist}(\check{Z}(u, x, w_*), \bar{w}) \\ + \pi_w \|x - \tilde{x}\| + \mu_u (\|u - S_u(x)\|_U + \|w_*\|_*). \end{aligned}$$

*Proof.* Again expanding the definitions of  $\operatorname{dist}$  like in the proof of **Lemma 4.10**, we have

$$\begin{aligned} d_1 := \operatorname{dist}(Z(x), w) - \operatorname{dist}(\check{Z}(u, x, w_*) \cap \Omega_W, w) &= \sup_{\tilde{w} \in \check{Z}(u, x, w_*) \cap \Omega_W} \inf_{w' \in Z(x)} \|w' - w\|_W - \|\tilde{w} - w\|_W \\ &\leq \sup_{\tilde{w} \in \check{Z}(u, x, w_*) \cap \Omega_W} \inf_{w' \in Z(x)} \|w' - \tilde{w}\|_W = \sup_{\tilde{w} \in \check{Z}(u, x, w_*) \cap \Omega_W} \operatorname{dist}(\check{Z}(S_u(x), x, 0), \tilde{w}) \end{aligned}$$

and

$$\begin{aligned} d_2 := \operatorname{dist}(\check{Z}(u, x, w_*), \bar{w}) - \operatorname{dist}(Z(\tilde{x}) \cap \Omega_W, \bar{w}) &= \sup_{\tilde{w} \in Z(\tilde{x}) \cap \Omega_W} \inf_{w' \in \check{Z}(u, x, w_*)} \|w' - \bar{w}\|_W - \|\tilde{w} - \bar{w}\|_W \\ &\leq \sup_{\tilde{w} \in Z(\tilde{x}) \cap \Omega_W} \inf_{w' \in \check{Z}(u, x, w_*)} \|w' - \tilde{w}\|_W = \sup_{\tilde{w} \in \check{Z}(S_u(\tilde{x}), \tilde{x}, 0) \cap \Omega_W} \operatorname{dist}(\check{Z}(u, x, w_*), \tilde{w}). \end{aligned}$$

Now, since  $x \in \Omega_X \cap S_u^{-1}(\Omega_U), u \in \Omega_U$ , and  $\tilde{w} \in \Omega_W$  in both cases, the Aubin property gives

$$(4.14) \quad \kappa_w d_1 \leq \kappa_w \sup_{\tilde{w} \in \check{Z}(u, x, w_*) \cap \Omega_W} C_w \operatorname{dist}(\check{Z}^{-1}(\tilde{w}) \cap \Omega, (S_u(x), x, 0)) \leq \kappa_w C_w (\|u - S_u(x)\|_U + \|w_*\|_*)$$

and

$$(4.15) \quad d_2 \leq \sup_{\tilde{w} \in \check{Z}(S_u(\tilde{x}), \tilde{x}, 0) \cap \Omega_W} C_w \text{dist}(\check{Z}^{-1}(\tilde{w}) \cap \Omega, (u, x, w_*)) \\ \leq C_w (\|u - S_u(\tilde{x})\|_U + \|\tilde{x} - x\| + \|w_*\|_*).$$

The claim follows from summing up and rearranging (4.14) and (4.15), and applying Lipschitz-continuity of  $S_u$

$$\|u - S_u(\tilde{x})\|_U \leq \|u - S_u(x)\|_U + \|S_u(x) - S_u(\tilde{x})\|_U \leq \|u - S_u(x)\|_U + L\|x - \tilde{x}\|. \quad \square$$

#### 4.5 DIFFERENTIAL TRANSFORMATION

To motivate the next construction, recall that in the algorithm of Section 4.1, specifically (4.1b), the differential estimate  $\tilde{x}_{k+1}^*$  arises jointly with the adjoint variable  $w^{k+1}$  from the adjoint inclusion

$$(4.16) \quad 0 \in \check{D}^* G(u^{k+1}, x^k | \Delta w_*^{k+1})(w^{k+1}) + (J'(u^{k+1}) - \Delta u_{k+1}^*, -\tilde{x}_{k+1}^*)$$

for either  $\check{D}^* = \widehat{D}^*, D^*$ . We can equivalently write this as

$$(4.17) \quad \tilde{x}_{k+1}^* \in \check{T}(u^{k+1}, w^{k+1}, \Delta u_{k+1}^*, \Delta w_*^{k+1}; x^k),$$

where  $\check{T} : U \times W \times U^* \times W_* \rightarrow X^*$ ,

$$\check{T}(u, w, \Delta u^* \Delta w_*; x) := \{x^* \in X^* \mid 0 \in \check{D}^* G(u, x | \Delta w_*)(w) + (J'(u) - \Delta u^*, -x^*)\}.$$

Likewise, we pick the target  $x_{k+1}^* \in \check{d}_2 F(x^k)$  such that

$$(4.18) \quad x_{k+1}^* \in \check{T}(S_u(x^k), S_w(x^k), 0, 0; x^k).$$

Such a choice exists if there exists an adjoint solution  $S_w(x^k) \in \check{Z}(x^k) := \check{Z}(S_u(x^k), x^k, 0)$ . (We do not have this implicit expression for  $\tilde{x}_{k+1}^*$ , as  $\check{Z}$  lacks the perturbation parameter  $\Delta u_{k+1}^*$ .) We can, thus, use Lipschitz-like properties of  $\check{T}$  to construct the target  $x_{k+1}^*$ , and to ensure Assumption 4.3 (iii):

**Theorem 4.14.** *Let  $U, X$  and  $W_*$  be normed spaces,  $W = (W_*)^*$ ,  $\Omega_X \subset X$ ,  $\Omega_U \subset U$ , and  $G : U \times X \rightrightarrows W_*$ . Let  $S_u$  and  $S_w$  be the single-valued selections of the inner and adjoint solution mapping as in Section 4.2. Assume that there exist exact and algorithmic adjoint solutions, i.e., (4.18) and (4.17). Moreover, assume whenever  $x^k \in \Omega_X$  and  $u^{k+1} \in \Omega_U$  that  $T(\cdot; x^k)$  is inner Lipschitz at  $v := (u^{k+1}, w^{k+1}, \Delta u_{k+1}^*, \Delta w_*^{k+1})$  with the factor  $L_T$  and radius  $\rho > 0$  such that  $(S_u(x^k), S_w(x^k), 0, 0) \in B(v, \delta)$ , more precisely,*

$$(4.19) \quad \text{dist}(\check{T}(S_u(x^k), S_w(x^k), 0, 0; x^k), \tilde{x}_{k+1}^*) \\ \leq L_T (\|S_u(x^k) - u^{k+1}\|_U + \|\Delta w_*^{k+1}\|_* + \|S_w(x^k) - w^{k+1}\|_W + \|\Delta u_{k+1}^*\|_*).$$

Then Assumption 4.3 (iii) holds for any  $\alpha_u = \alpha_w > L_T$  and some  $x_{k+1}^* \in \check{T}(S_u(x^k), S_w(x^k), 0; x^k)$ .

*Proof.* The claim is immediate from the assumed Lipschitz-like property and the existence of the exact and algorithmic adjoint solutions. (We take any  $\alpha_u = \alpha_w > L_T$  to avoid having to assume any compactness of  $\check{T}(S_u(x^k), S_w(x^k), 0, 0; x^k)$ .)  $\square$

**Remark 4.15 (Existence of solutions).** Subject to the conditions of Theorem 3.11, for the limiting coderivative, (4.18) holds for a choice of  $S_u(x^k)$  when  $x^k$  (locally) minimises the tilted functional  $\text{Inf}_{J \circ S} + x^*$  for some  $R = x^* \in X^*$ . Likewise, (4.17) holds when  $x^k$  (locally) minimises  $\text{Inf}_{[J + \Delta u_{k+1}^*] \circ S} + x^*$  for the inner problem  $0 \in G(u, x) - \langle \Delta w_*^{k+1} | u \rangle$ .

#### 4.6 A FORWARD-BACKWARD TYPE OUTER METHOD

We now prove the convergence of a forward-backward type outer method, assuming inner and adjoint methods that satisfy [Assumption 4.3](#). The next convergence result does not depend on our specific problem structure; it applies to general problems of the form

$$(4.20) \quad \min_{x \in X} E(x) + R(x),$$

for which [Assumption 4.3](#) holds with  $R : X \rightarrow \overline{\mathbb{R}}$  is convex and possibly nonsmooth, and  $E : X \rightarrow \mathbb{R}$ . Example choices for  $E$  are the optimistic and pessimistic selections  $\text{Inf}_{J \circ S}$  and  $\text{Sup}_{J \circ S}$ , which are defined in [Section 2](#), with  $J : U \rightarrow \mathbb{R}$  and a solution mapping  $S$  as in [\(3.2\)](#).

Given a step length parameter  $\tau > 0$ , we take as the (inexact) outer algorithm [\(4.1c\)](#) the forward-backward splitting

$$(4.21) \quad x^{k+1} := \arg \min_{z \in X} \left\{ \tau \langle \tilde{x}_{k+1}^* | z - x^k \rangle + \tau R(z) + \frac{1}{2} \|z - x^k\|^2 \right\},$$

In Hilbert spaces this reduces to the familiar

$$x^{k+1} := \text{prox}_{\tau R}(x^k - \tau \mathcal{R} \tilde{x}_{k+1}^*).$$

In our analysis, we do not use any specific coderivative or subdifferential of  $E$ , instead introducing a “target set”  $\mathcal{d}E(x^k) \ni x_{k+1}^*$ , where we recall that  $x_{k+1}^*$  is the tracking target from [Assumption 4.3](#). Informally,  $0 \in [\mathcal{d}E + \partial R](\bar{x})$  should be a meaningful optimality condition for [\(4.20\)](#). We have discussed meaningful choices of  $\mathcal{d}E$  of  $E = J \circ S$  in [Remark 3.12](#). A trivial choice is  $\mathcal{d}E(x) = \{x_{k+1}^*\}$ , although we will impose further technical restrictions that need to be satisfied. In particular, we require the set-valued descent inequality

$$(4.22) \quad \inf_{x^* \in \mathcal{d}E(x^k)} \langle x^* | x - x^k \rangle \geq E(x) - E(x^k) - \frac{L}{2} \|x - x^k\|^2 \quad \text{for all } k \in \mathbb{N} \text{ and } x \in \text{dom } R.$$

This property – that obviously benefits from a smaller choice of  $\mathcal{d}E(x^k)$  – reduces to the standard descent inequality for functions with  $L$ -Lipschitz continuous Fréchet differential [[13](#), [Lemma 7.1](#)] by taking  $\mathcal{d}E(x^k) = \{E'(x^k)\}$ . In [Section 4.7](#) we study a positive nonsmooth example.

With all of the above constructions and assumptions at hand, we can finally show the convergence of subdifferentials to zero. The result is global – even for nonconvex  $E$  – if we can take  $\Omega_X = X$  in [Assumption 4.3](#). It does not say anything about the existence of limiting points of  $\{x^k\}_{k \in \mathbb{N}}$ , or their optimality. However, the result of the theorem proves that under appropriate compactness and outer semicontinuity assumptions, any limiting point of the sequence of iterates will satisfy the relevant condition discussed in [Remark 3.12](#). We prove first a lemma, and then the convergence theorem.

**Lemma 4.16.** *Let  $X$  and  $U$  be normed spaces,  $E : X \rightarrow \mathbb{R}$ , and  $R : X \rightarrow \overline{\mathbb{R}}$  convex, proper and lower semicontinuous. Given an initial iterate  $x^0 \in \Omega_X$ ,  $N \geq 1$  and  $\tau > 0$ , generate  $\{x^k\}_{k=1}^{N+1}$  through [\(4.21\)](#), choosing for each iteration corresponding target sets  $\mathcal{d}E(x^k)$  for which [\(4.22\)](#) holds for an  $L > 0$ , as well as differential estimates  $\tilde{x}_{k+1}^*$  for which [Assumption 4.3](#) holds until  $N$ . Assume, moreover, that  $\Omega_X$  contains the  $([E + R](x^0) + \Psi_1)$ -sublevel set of  $E + R$ , i.e.,*

$$(4.23) \quad \{x \in X \mid [E + R](x) \leq [E + R](x^0) + \Psi_1\} \subset \Omega_X.$$

Then, for each  $k \in \{0, \dots, N + 1\}$ , we have  $x^k \in \Omega_X$  and

$$(4.24) \quad [E + R](x^k) + \tau \left( 1 - \frac{\tau(L + 1 + \zeta_1^2)}{2} \right) \sum_{j=0}^{k-1} \|\tilde{x}_{j+1}^* + q^{j+1}\|^2 \leq [E + R](x^0) + \sum_{j=0}^{k-1} e_j$$

for some  $q^{j+1} \in \partial R(x^{j+1})$  and  $e_j$ , ( $j = 0, \dots, N$ ), satisfying

$$(4.25) \quad \max_{k \in \{0, \dots, N\}} \sum_{j=0}^k e_j \leq \Psi_1 := \frac{(d_1^u)^2}{\pi_u} \left( \frac{\varsigma_1 \alpha_u \kappa}{\kappa - 1} + \frac{\varsigma_1 \alpha_w \mu_u}{(\kappa - 1)^2} \right) + \frac{(d_1^w)^2}{\pi_w} \left( \frac{\varsigma_1 \alpha_w \kappa}{\kappa - 1} \right),$$

where  $\kappa := \min\{\kappa_u, \kappa_w\} > 1$ , and  $\bar{\kappa} := \max\{\kappa_u, \kappa_w\} > 1$ , the first-step distances  $d_1^u$  and  $d_1^w$  are given in (4.6), and the constant

$$(4.26) \quad \varsigma_1 \leq \frac{(\alpha_u \pi_u + \alpha_w \pi_w) \bar{\kappa}}{\kappa - 1} + \frac{\alpha_w \mu_u \pi_u \bar{\kappa}}{(\kappa - 1)^2}.$$

Finally, suppose, additionally, that the step length  $\tau$  satisfies  $\tau(L+1+\varsigma_1^2) < 2$  and  $\inf[E+R] > -\infty$ . Then  $\sup_{k \leq N} \sum_{j=0}^{k-1} \|x^{j+1} - x^j\|^2 < \infty$ . If additionally  $\sup_{x \in \Omega_X} \|S_u(x)\|_U < \infty$ , then also  $\sup_{k \leq N} \|u^{k+1}\|_U < \infty$ , with

$$(4.27) \quad r_U := \sup_{x \in \Omega_X} \|S_u(x)\|_U + \|u^1 - S(x^0)\|_U + \|\Delta w_*^1\|_* + \pi_u \sqrt{\frac{2\tau([E+R](x^0) - \inf[E+R] + \Psi_1)}{(1 - \kappa_u^{-2})(2 - \tau(L+1 + \varsigma_1^2))}}.$$

*Proof.* We use the results of Appendix c from [18] with the choices (4.6), and  $\tilde{d}_k = \|\tilde{x}_{k+1}^* - x_{k+1}^*\|$ . To verify Assumption c.1 through Assumption 4.3, we need to show that  $\{x^j\}_{j=0}^k \subset \Omega_X$  for each  $k \in \{0, \dots, N+1\}$ . We do this inductively along with the inductive proof of (4.24). Clearly both hold for  $k = 0$ , since by assumption,  $x^0 \in \Omega_X$ . This takes care of the inductive basis.

For the inductive step, let  $k \geq 0$  and  $k \leq N$ , and suppose we have proved  $\{x^j\}_{j=0}^k \subset \Omega_X$  and (4.24). By [47, Lemma 3.8], we have

$$(4.28) \quad \|x^{k+1} - x^k\|^2 = \tau^2 \|\tilde{x}_{k+1}^* + q^{k+1}\|^2 = -\tau \langle \tilde{x}_{k+1}^* + q^{k+1} | x^{k+1} - x^k \rangle \quad \text{for some } q^{k+1} \in \partial R(x^{k+1}).$$

Young's inequality gives

$$\langle \tilde{x}_{k+1}^* - x_{k+1}^* | x^{k+1} - x^k \rangle \geq -\frac{1}{2} \|\tilde{x}_{k+1}^* - x_{k+1}^*\|^2 - \frac{1}{2} \|x^{k+1} - x^k\|^2$$

Continuing with Lemma c.3 gives

$$\langle \tilde{x}_{k+1}^* - x_{k+1}^* | x^{k+1} - x^k \rangle \geq -\frac{1 + \varsigma_1^2}{2} \|x^{k+1} - x^k\|^2 - e_k.$$

Now (4.22) implies

$$(4.29) \quad \begin{aligned} \langle \tilde{x}_{k+1}^* | x^{k+1} - x^k \rangle &= \langle \tilde{x}_{k+1}^* - x_{k+1}^* | x^{k+1} - x^k \rangle + \langle x_{k+1}^* | x^{k+1} - x^k \rangle \\ &\geq -\frac{1 + \varsigma_1^2}{2} \|x^{k+1} - x^k\|^2 - e_k + \inf_{x^* \in dE(x^k)} \langle x^* | x^{k+1} - x^k \rangle \\ &\geq E(x^{k+1}) - E(x^k) - \frac{L+1 + \varsigma_1^2}{2} \|x^{k+1} - x^k\|^2 - e_k. \end{aligned}$$

Using the definition of the convex subdifferential as well as (4.28) and (4.29), we get

$$\begin{aligned} [E+R](x^{k+1}) - [E+R](x^k) + \tau \|\tilde{x}_{k+1}^* + q^{k+1}\|^2 &= [E+R](x^{k+1}) - [E+R](x^k) - \langle \tilde{x}_{k+1}^* + q^{k+1} | x^{k+1} - x^k \rangle \\ &\leq E(x^{k+1}) - E(x^k) - \langle \tilde{x}_{k+1}^* | x^{k+1} - x^k \rangle \\ &\leq \frac{L+1 + \varsigma_1^2}{2} \|x^{k+1} - x^k\|^2 + e_k \\ &= \frac{\tau^2(L+1 + \varsigma_1^2)}{2} \|\tilde{x}_{k+1}^* + q^{k+1}\|^2 + e_k. \end{aligned}$$

Therefore,

$$[E + R](x^{k+1}) - [E + R](x^k) + \tau \left( 1 - \frac{\tau(L + 1 + \zeta_1^2)}{2} \right) \|\tilde{x}_{k+1}^* + q^{k+1}\|^2 \leq e_k.$$

Adding this to (4.24), we get the same for  $k + 1$ . Then (4.23) imply  $x^{k+1} \in \Omega_X$ , which concludes the induction. Thus (4.24) holds for each  $k \in \{0, \dots, N + 1\}$ , and  $\{x^k\}_{k=0}^{N+1} \subset \Omega_X$ . Consequently, Assumption 4.3 implies Assumption c.1 for each  $k \in \{0, \dots, N + 1\}$ . Therefore Lemma c.3 gives (4.25). The exact expression for  $\zeta_1$  is in (c.2), and satisfies (4.26).

Using (4.24), (4.25) and (4.28) we obtain

$$(4.30) \quad [E + R](x^k) + \tau^{-1} \left( 1 - \frac{\tau(L + 1 + \zeta_1^2)}{2} \right) \sum_{j=0}^{k-1} \|x^{j+1} - x^j\|^2 \leq [E + R](x^0) + \sum_{j=0}^{k-1} e_j < [E + R](x^0) + \Psi_1.$$

We have  $\tau^{-1} (1 - \tau(L + 1 + \zeta_1^2)/2) > 0$  by assumption  $\tau(L + 1 + \zeta_1^2) < 2$ . Therefore (4.30) implies  $\sup_{k \leq N} \sum_{j=0}^{k-1} \|x^{j+1} - x^j\|^2 < \infty$ .

Within the aforementioned inductive step we have  $\{x^j\}_{j=0}^k \subset \Omega_X$  for some  $k \in \{1, \dots, N\}$ . Therefore we can iterate Assumption 4.3 (i), the inner tracking property, to obtain

$$(4.31) \quad \begin{aligned} \sup_{k \leq N} \|u^{k+1}\|_U &\leq \sup_{k \leq N} \{ \|S_u(x^k)\|_U + \|u^{k+1} - S_u(x^k)\|_U + \|\Delta w_*^{k+1}\|_* \} \\ &\leq \sup_{k \leq N} \|S_u(x^k)\|_U + \sup_{k \leq N} \kappa_u^{-k} (\|u^1 - S(x^0)\|_U + \|\Delta w_*^1\|_*) + \sup_{k \leq N} \sum_{j=1}^k \kappa_u^{-(k+1-j)} \pi_u \|x^j - x^{j-1}\| \\ &\leq \sup_{x \in \Omega_X} \|S_u(x)\|_U + \|u^1 - S(x^0)\|_U + \|\Delta w_*^1\|_* + \sup_{k \leq N} \sum_{j=1}^k \kappa_u^{-(k+1-j)} \pi_u \|x^j - x^{j-1}\|. \end{aligned}$$

Hölder's inequality and the fact  $\{\kappa_u^{-j}\}_{j=1}^\infty$  is a converging geometric sequence due to  $\kappa_u > 1$  imply

$$\sum_{j=1}^k \kappa_u^{-(k+1-j)} \pi_u \|x^j - x^{j-1}\| \leq \pi_u \sqrt{\sum_{j=1}^k \kappa_u^{-2j}} \sqrt{\sum_{j=1}^k \|x^j - x^{j-1}\|^2} < \frac{\pi_u}{\sqrt{1 - \kappa_u^{-2}}} \sqrt{\sum_{j=0}^{k-1} \|x^{j+1} - x^j\|^2}.$$

Combining this with (4.31), the assumption  $\sup_{x \in \Omega_X} \|S_u(x)\|_U < \infty$  and (4.30) imply  $\sup_{k \leq N} \|u^{k+1}\|_U < r_U$  with  $r_U$  defined in (4.27).  $\square$

**Remark 4.17 (Bounds on the inner variable).** Lemma 4.16 requires that the differential transformation Assumption 4.3 (iii) holds. We can obtain it for a special case using Theorem 4.14. Its assumption (4.19) may depend on  $\Omega_U$  to be bounded, as we will see in the applications of Section 5.3. By an inductive argument based on Lemma 4.16, we can, indeed, choose  $\Omega_U = B(0, r_U)$ , where  $r_U$  is the bound (4.27) provided the lemma, independent of  $N$ , for which  $\sup_{k \leq N} \|u^{k+1}\|_U < r_U$ .

To avoid circular reasoning, we have to, however, be careful when using  $r_U$  to construct  $\Omega_U$  and prove Assumption 4.3: the radius depends on factors from Assumption 4.3, that may depend on  $\Omega_U$ . However, studying (4.27) in more detail, independent of these factors, we see that we can make  $r_U$  arbitrarily close to  $\sup_{x \in \Omega_X} \|S_u(x)\|_U$  by (a) either or both taking  $\tau > 0$  small enough, or initialising  $x^0$  close to a minimiser, and (b) making high-quality first steps. In (a), the closeness and smallness still depend on these factors, or rough estimates for them.

**Theorem 4.18.** *Let  $X$  and  $U$  be normed spaces,  $E : X \rightarrow \mathbb{R}$ , and  $R : X \rightarrow \overline{\mathbb{R}}$  convex, proper and lower semicontinuous. Assume additionally that  $\inf[E + R] > -\infty$  and (4.23) hold. Given an initial*

iterate  $x^0 \in \Omega_X$  and  $\tau > 0$ , generate  $\{x^k\}_{k \in \mathbb{N}}$  through (4.21), choosing for each iteration corresponding target sets  $\mathcal{d}E(x^k)$  for which (4.22) holds for an  $L > 0$ , as well as differential estimates  $\widetilde{x}_{k+1}^*$  and targets  $x_{k+1}^* \in \mathcal{d}E(x^k)$  for which Assumption 4.3 holds. Suppose the step length  $\tau$  satisfies  $\tau(L + 1 + \zeta_1^2) < 2$ . Then  $\|u^{k+1} - S_u(x^k)\|_U \rightarrow 0$ . Moreover:

(i) If  $X$  is a Hilbert space, then (recall Remark 3.14)

$$(4.32) \quad \inf_{x^* \in \mathcal{d}E(x^k)} \|\text{prox}_{\tau R}(x^k - \tau \mathcal{R}x^*) - x^k\| \rightarrow 0.$$

(ii) Suppose that either

(a) For all  $\varepsilon > 0$  there exists  $\delta > 0$  such that, for some  $k_N \in \mathbb{N}$  for any  $k \geq k_N$ ,

$$(4.33) \quad \|x^{k+1} - x^k\| \leq \delta \implies \text{dist}(\mathcal{d}E(x^{k+1}), x_{k+1}^*) \leq \varepsilon.$$

(b)  $E$  is bounded from below and  $\partial R$  is single-valued Lipschitz on the restricted sublevel sets  $\text{sub}_R(c|\Omega_X) := \{x \in \Omega_X \mid R(x) \leq c\}$  for any  $c \in \mathbb{R}$ .

Then also  $\inf_{x^* \in [\mathcal{d}E + \partial R](x^{k+1})} \|x^*\| \rightarrow 0$ .

*Proof.* We first prove that

$$(4.34) \quad \|x^{k+1} - x^k\| \rightarrow 0, \quad \|\widetilde{x}_{k+1}^* + q^{k+1}\| \rightarrow 0, \quad \text{and} \quad \|\widetilde{x}_{k+1}^* - x_{k+1}^*\| \rightarrow 0.$$

To do so, we start by applying Lemma 4.16. We have  $\sup_{N \in \mathbb{N}} \sum_j^N e_j \leq \Psi_1 < \infty$ , from (4.25), and (4.24). Since, by assumption,  $\tau(L + 1 + \zeta_p^2) < 2$  and  $\inf[E + R] > -\infty$ , it follows from (4.24) that  $\|\widetilde{x}_{k+1}^* + q^{k+1}\| \rightarrow 0$ . We have again  $\tau\|\widetilde{x}_{k+1}^* + q^{k+1}\| = \|x^{k+1} - x^k\|$  by [47, Lemma 3.8], and thus also  $\|x^{k+1} - x^k\| \rightarrow 0$ . By Lemma C.2 we have  $\|\widetilde{x}_{k+1}^* - x_{k+1}^*\|^2 \leq \check{e}_{p,k}$  for some  $\check{e}_{p,k} \geq 0$ . Since, by Lemma 4.16,  $\sup_{N \in \mathbb{N}} \sum_{k=0}^{N-1} \|x^{k+1} - x^k\|^2 < \infty$ , by Lemma C.4, these quantities satisfy  $\sum_{k=0}^{\infty} \check{e}_{p,k} < \infty$ . The final convergence of (4.34) follows.

Next we prove  $\|u^{k+1} - S_u(x^k)\|_U \rightarrow 0$ . Iterating the inner tracking Assumption 4.3 (i), yields

$$(4.35) \quad \|u^{k+1} - S_u(x^k)\|_U + \|\Delta w_*^{k+1}\|_* \leq \kappa_u^{-k} (\|u^1 - S(x^0)\|_U + \|\Delta w_*^1\|_*) + \sum_{j=1}^k \kappa_u^{-(k+1-j)} \pi_u \|x^j - x^{j-1}\|$$

where  $\kappa_u^{-k} (\|u^1 - S(x^0)\|_U + \|\Delta w_*^1\|_*) \rightarrow 0$  since  $\kappa_u > 1$ . Let  $\varepsilon > 0$ . We have  $\|x^j - x^{j-1}\| < \varepsilon$  for all  $j \geq n_\varepsilon$  for some  $n_\varepsilon \in \mathbb{N}$  because  $\|x^j - x^{j-1}\| \rightarrow 0$ . The norms  $\|x^j - x^{j-1}\|$  are bounded for any  $j \in \mathbb{N}$  since  $\sup_{k \leq N} \sum_{j=0}^{k-1} \|x^{j+1} - x^j\|^2 < \infty$  by Lemma 4.16. Using additionally the fact that  $\{\kappa_u^{-j}\}_{j=1}^{\infty}$  is a converging geometric sequence we obtain for any  $k > n_\varepsilon$  that

$$\begin{aligned} \sum_{j=1}^k \kappa_u^{-(k+1-j)} \|x^j - x^{j-1}\| &= \sum_{j=1}^{n_\varepsilon} \kappa_u^{-(k+1-j)} \|x^j - x^{j-1}\| + \sum_{j=n_\varepsilon+1}^k \kappa_u^{-(k+1-j)} \|x^j - x^{j-1}\| \\ &\leq n_\varepsilon \kappa_u^{-(k+1-n_\varepsilon)} \sup_{1 \leq j \leq n_\varepsilon} \|x^j - x^{j-1}\| + \frac{\varepsilon}{1 - \kappa_u^{-1}}. \end{aligned}$$

Therefore,  $\sum_{j=1}^k \kappa_u^{-(k+1-j)} \|x^j - x^{j-1}\| \rightarrow 0$  and  $\|u^{k+1} - S_u(x^k)\|_U \rightarrow 0$  follow from (4.35).

We now divide the argument into cases for the alternative assumptions.

**Case (i):** Since proximal maps are 1-Lipschitz continuous [13, Corollary 6.18] with factor 1, we have

$$\begin{aligned} \|\text{prox}_{\tau R}(x^k - \mathcal{R}x_{k+1}^*) - x^k\| &\leq \|\text{prox}_{\tau R}(x^k - \mathcal{R}\widetilde{x}_{k+1}^*) - x^k\| + \|\mathcal{R}\widetilde{x}_{k+1}^* - \mathcal{R}x_{k+1}^*\| \\ &= \|x^{k+1} - x^k\| + \|\widetilde{x}_{k+1}^* - x_{k+1}^*\|. \end{aligned}$$

Using (4.34), the claim follows.

**Case (ii)(a):** We have

$$\begin{aligned}
 (4.36) \quad \inf_{x^* \in [\mathcal{d}E + \partial R](x^{k+1})} \|x^*\| &\leq \|\tilde{x}_{k+1}^* + q^{k+1}\| + \inf_{x^* \in [\mathcal{d}E + \partial R](x^{k+1})} \|x^* - (\tilde{x}_{k+1}^* + q^{k+1})\| \\
 &\leq \|\tilde{x}_{k+1}^* + q^{k+1}\| + \inf_{x^* \in \mathcal{d}E(x^{k+1})} \|x^* - \tilde{x}_{k+1}^*\| \\
 &\leq \|\tilde{x}_{k+1}^* + q^{k+1}\| + \text{dist}(\mathcal{d}E(x^{k+1}), x_{k+1}^*) + \|x_{k+1}^* - \tilde{x}_{k+1}^*\|.
 \end{aligned}$$

We have  $\text{dist}(\mathcal{d}E(x^{k+1}), x_{k+1}^*) \rightarrow 0$  from (4.33) and  $\|x^{k+1} - x^k\| \rightarrow 0$ . The latter holds by (4.34), which also establishes that the remaining terms converge to zero.

**Case (ii)(b):** We have

$$\begin{aligned}
 \inf_{x^* \in [\mathcal{d}E + \partial R](x^k)} \|x^*\| &\leq \inf_{q \in \partial R(x^k)} \|x_{k+1}^* + q\| \\
 &\leq \|\tilde{x}_{k+1}^* - x_{k+1}^*\| + \|\tilde{x}_{k+1}^* + q^{k+1}\| + \inf_{q \in \partial R(x^k)} \|q^{k+1} - q\|.
 \end{aligned}$$

Since  $E$  is bounded from below, (4.24) and (4.25) establish a bound  $c \geq R(x^k)$  for all  $k \in \mathbb{N}$ . By Lemma 4.16, we have  $x^k \in \Omega_X$  for all  $k \in \mathbb{N}$ . Since  $\partial R$  is single-valued Lipschitz<sup>4</sup> on  $\text{sub}_R(c|\Omega_X)$ ,  $\|q^{k+1} - q^k\| \rightarrow 0$  follows from  $\|x^{k+1} - x^k\| \rightarrow 0$ . The latter holds by (4.34), which also establishes that the remaining terms converge to zero.  $\square$

**Remark 4.19.** The Lipschitz property on sublevel sets in (ii)(b) holds for, e.g., barrier functions, which are Lipschitz on their sublevel sets, but whose differentials' Lipschitz factors blow up towards the boundary.

**Remark 4.20.** If we are content with subsequential convergence,  $\liminf_{k \rightarrow \infty} \inf_{x^* \in [\mathcal{d}E + \partial R](x^{k+1})} \|x^*\| = 0$ , then the argument of (ii)(b) can be generalised to functions  $R$  that have a Lipschitz differential on their domain, provided the domain has a Lipschitz boundary, and we ensure that  $\{\tilde{x}_{k+1}^*\}$  is bounded. This includes, in particular, indicator functions of Lipschitz domains.

Indeed, we only need to show that  $\liminf_{k \rightarrow \infty} \inf_{q \in \partial R(x^k)} \|q^{k+1} - q\| \rightarrow 0$ . Now, the boundedness of  $\{\tilde{x}_{k+1}^*\}$  and (4.34) ensure that also  $\{q^k\}$  is bounded. Using this boundedness when on the boundary of  $\text{dom } R$  with the assumed Lipschitz properties in a case by case analysis of transitions between  $\text{bd dom } R$  or  $\text{int dom } R$ , we see that only when  $x^{k+1} \in \text{bd dom } R$  and  $x^k \in \text{int dom } R$ , we cannot bound  $\inf_{q \in \partial R(x^k)} \|q^{k+1} - q\| \leq L\|x^{k+1} - x^k\|$  for some  $L > 0$ . This can happen at most on every second step.

#### 4.7 EXAMPLE

We now study the assumptions of Theorem 4.18 in a simple one-dimensional example, reminiscent of optimally choosing the regularisation parameter for the Lasso problem. That is, we consider the problem

$$(4.37) \quad \min_{x,u} \underbrace{\frac{1}{2}(u-2)^2}_{J(u)} + \underbrace{\delta_{\geq \varepsilon}(x)}_{R(x)} \quad \text{subject to} \quad u = \arg \min_v \underbrace{\frac{1}{2}(v-5)^2}_{f(v;x)} + \underbrace{|x|v|}_{g(v;x)}$$

<sup>4</sup>We could virtually relax this assumption to a “uniform semi-Lipschitz” property, but such a property does not appear to hold for any nonsmooth functions of interest.

with a small  $\varepsilon > 0$ . The inner objective is (strongly) convex for any  $x \geq 0$ . Therefore, the inner problem can equivalently be written as  $0 = G(u, x)$  for

$$(4.38) \quad G(u, x) = (u - 5) + x \begin{cases} \{-1\} & \text{if } u < 0, \\ [-1, 1] & \text{if } u = 0, \\ \{1\} & \text{if } u > 0. \end{cases}$$

The next lemma provides the inner solution mapping  $S_u$ .

**Lemma 4.21.** *Let  $G : \mathbb{R} \times \mathbb{R}_+ \rightrightarrows \mathbb{R}$  given by (4.38) and define the solution mapping  $S_u : \mathbb{R}_+ \rightrightarrows \mathbb{R}$  through the satisfaction of  $0 \in G(S_u(x), x)$ . Then  $S_u(x) = \max(0, 5 - x)$ .*

*Proof.* Studying the optimality condition  $0 \in G(u, x)$  yields the following conclusions:

- If  $u < 0$ , we must have  $u = 5 + x$ . Given that we restrict  $x \geq 0$ , it follows that  $u < 0$  cannot be a solution for any admissible  $x$ .
- If  $u > 0$ , we must have  $u = 5 - x$ . Such an  $u$  is indeed a solution when  $0 \leq x < 5$ .
- If  $u = 0$ , we must have  $5 \in x[-1, 1]$ . Thus  $u = 0$  is a solution for any  $x \geq 5$ .

Therefore we obtain the (single-valued inner) solution mapping  $S_u(x) = \max(0, 5 - x)$ .  $\square$

**Remark 4.22.** Similarly to Lemma 4.21,  $\tilde{S}_u : \mathbb{R}_+ \rightrightarrows \mathbb{R}$  solving  $\Delta w_*^{k+1} \in G(\tilde{S}_u(x^k), x^k)$  can be shown to be  $\tilde{S}_u(x) = \max(0, 5 - x + \Delta w_*^{k+1})$ . When  $x - \Delta w_*^{k+1} > 0$  it holds  $\max(0, 5 - x + \Delta w_*^{k+1}) = S_u(x - \Delta w_*^{k+1})$ , which demonstrates that  $\tilde{S}_u(x)$  and  $S_u$  as translation relationship, and thus we can deduce properties related to  $\Delta w_*^{k+1} \in G(u^{k+1}, x^k)$  from the ones related to  $0 \in G(u, x)$ .

Thus taking  $J(u) = \frac{1}{2}(u - 2)^2$  simplifies the first term in the outer objective into the form

$$(4.39) \quad E(x) = [J \circ S_u](x) = \frac{1}{2}(\max(0, 5 - x) - 2)^2 = \begin{cases} \frac{1}{2}(x - 3)^2 & \text{if } 0 \leq x \leq 5 \\ 2 & \text{if } x > 5. \end{cases}$$

The functions  $S_u$  and  $E$  are visualised in Figure 4.1. Since  $S_u$  is Lipschitz continuous (and single-valued), Theorem 4.9 implies the inner tracking property provided the inner algorithm satisfies the basic contractivity property (4.7). This is the case for, e.g., forward-backward splitting by Theorem A.2. Moreover, because  $S_u$  is single valued and Lipschitz, Theorem 2.6 gives

$$\widehat{D}^*[J \circ S_u](x|J(u))(1) = \widehat{D}^*S_u(x|u)(J'(u)), \quad \text{for } x > 0, u = S_u(x).$$

This is also visualised in Figure 4.1. We take next a closer look at the adjoint based on  $G$ , as formulated in (3.15), and used as basis of our algorithms.

**Lemma 4.23.** *Let  $G$  be given by (4.38). Suppose  $u, x \in \mathbb{R}$  satisfy  $0 \in G(u, x)$  and  $x > 0$ . Let  $u^* = J'(u) = u - 2$  and  $x^*, w \in \mathbb{R}$ . Then*

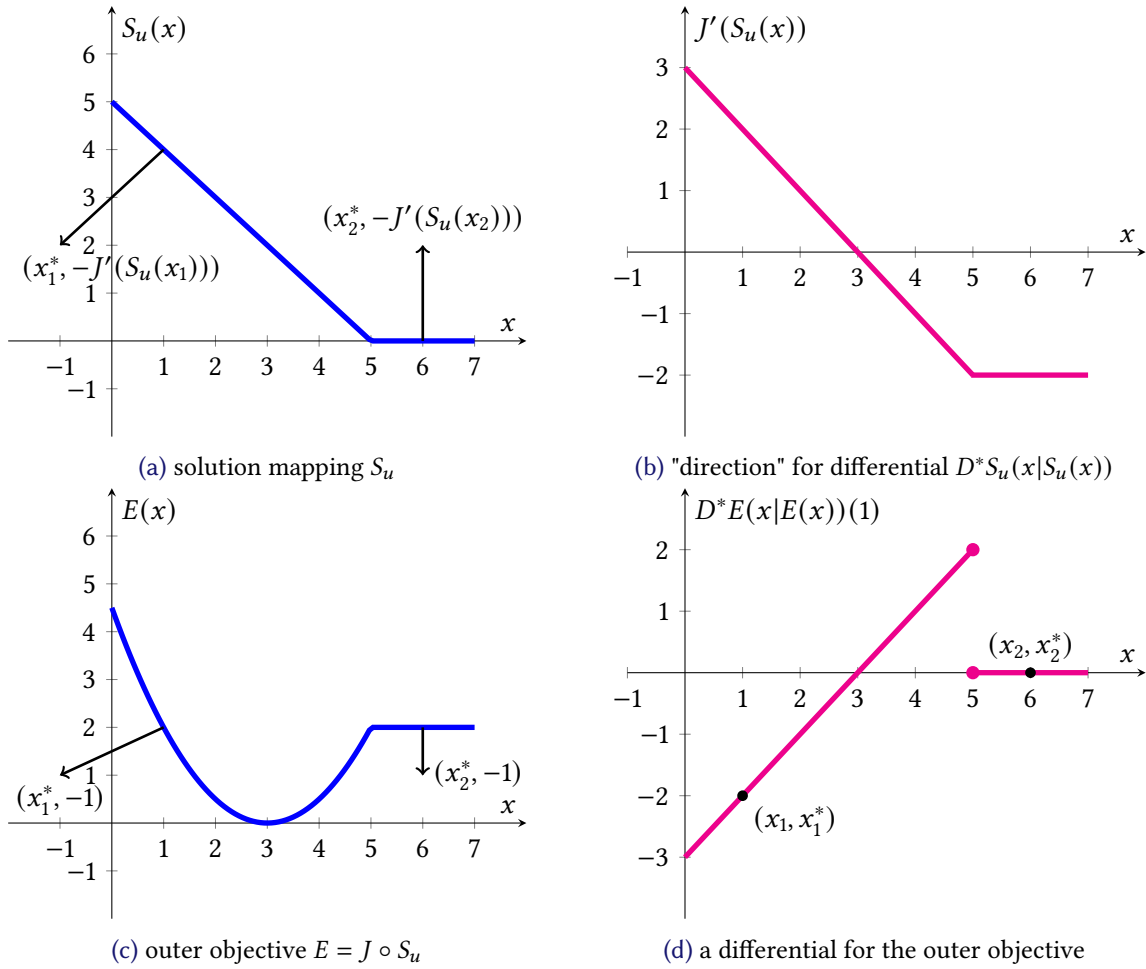
$$(4.40) \quad (-u^*, x^*) \in \widehat{D}^*G(u, x|0)(w)$$

is equivalent to

$$(4.41) \quad x^* = w \in \begin{cases} \{-u^*\} = \{2 - u\} & \text{if } x \in (0, 5) \text{ and } u = 5 - x \\ \{0\} & \text{if } x > 5 \text{ and } u = 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

Likewise, for the limiting coderivative,  $(-u^*, x^*) \in D^*G(u, x|0)(w)$  is equivalent to

$$(4.42) \quad x^* = w \in \begin{cases} \{-u^*\} = \{2 - u\} & \text{if } x \in (0, 5) \text{ and } u = 5 - x \\ \{0\} & \text{if } x > 5 \text{ and } u = 0, \\ \{0, 2\} & \text{if } x = 5 \text{ and } u = 0 \\ \emptyset & \text{otherwise.} \end{cases}$$



**Figure 4.1:** Illustrations for Section 4.7. The left subfigures visualise single-valued Lipschitz continuous solution mapping  $S_u$  and corresponding outer objective  $E = J \circ S_u$  with important normal vectors for  $x_1 = 1$  and  $x_2 = 6$ . The  $y$ -directions of the normal vectors are such that for  $i = 1, 2$  it holds  $x_i^* \in D^*E(x_i|E(x_i))(1)$  and  $x_i^* \in D^*S_u(x_i|S_u(x_i))(J'(S_u(x_i)))$ . In this case different versions of  $\mathcal{d}E(x)$  from Remark 3.12 are equal. It is also visually clear that at  $x = 5$  there does not exist vector orthogonal to  $\text{Graph } S_u$  with positive  $y$ -coordinate or orthogonal to  $\text{Graph } E$  with negative  $y$ -coordinate. Therefore  $\widehat{D}^*E(5|E(5))(1) = \emptyset$  and  $\widehat{D}^*S_u(5|S(5))(J'(S_u(5))) = \emptyset$ .

*Proof.* The condition  $0 \in G(u, x)$  implies  $u = \max(0, 5 - x)$  by Lemma 4.21. Therefore  $u > 0$  when  $x \in (0, 5)$  and  $u = 0$  when  $x \geq 5$ . We consider the case  $u > 0$  first. Then  $G$  is differentiable at  $(u, x)$ , so by [13, Theorem 20.14] we have  $\widehat{D}^*G(u, x|0)(w) = D^*G(u, x|0)(w) = \{G'(u, x)^*w\} = \{(w, w)\}$ . Therefore, as claimed,  $x^* = w = -u^*$ .

We next find  $(-u^*, x^*) \in \widehat{D}^*G(u, x|w_*)(w)$  for  $u = 0$  and any  $w_* \in G(u, x)$  close to zero. By  $w_* \in G(u, x)$ , this requires  $-x \leq w_* + 5 \leq x$ . We separate the study into two cases:  $-x < w_* + 5 < x$  and  $w_* + 5 = x$ , the lower bound not being possible for  $w_* \approx 0$ . In either case, by the definition of the Fréchet coderivative

$$(4.43) \quad \limsup_{x^k \rightarrow x, u^k \rightarrow u, G(u^k, x^k) \ni w_*^k \rightarrow 0} \frac{-\langle w|w_*^k - w_* \rangle + \langle x^*|x^k - x \rangle - \langle u^*|u^k - u \rangle}{\|(x^k - x, u^k - u, w_*^k - 0)\|} \leq 0.$$

We consider the case  $-x < w_* + 5 < x$  first. First of all, taking  $u^k \rightarrow u = 0$  and  $x^k \rightarrow x$ , we can only have  $G(u^k, x^k) \ni w_*^k \rightarrow w_*$  in (4.38) if  $u^k = 0$  and  $x^k > 0$  for all  $k$  large enough. Taking also  $w_*^k = w_*$  and passing to the limit, it follows that, necessarily,  $x^* = 0$ , as claimed. On the other hand, we can take  $w_*^k \in G(u^k, x^k) = x^k[-1, 1] - 5$ ,  $w_*^k \rightarrow w_*$  such that either  $w_*^k > w_*$  or  $w_*^k < w_*$  for all  $k \in \mathbb{N}$ . Therefore for any  $w \in \mathbb{R}$  we have  $\sup -\langle w | w_*^k - w_* \rangle = |w(w_*^k - w_*)|$  and thus (4.43) can hold only if also  $w = 0$ . Clearly, on the other hand, the choice  $x^* = w = 0$  satisfies (4.43). Thus, taking  $w_* = 0$ , we prove the Fréchet coderivative expression for  $u = 0$  and  $x \in (0, 5)$ .

Consider the case  $u = 0$  and  $w_* + 5 = x$ . Take  $x^k > x$  and  $u^k = 0$ . Then  $0 \in G(u^k, x^k)$ , so we can take  $w_*^k = 0$  in (4.43). It follows that (4.43) can hold only if  $\limsup_{k \rightarrow \infty} \langle x^* | x^k - x \rangle / \|x^k - x\| \leq 0$ . Since, we have taken any  $x < x^k \rightarrow x$ , we must have  $x^* \leq 0$ . On the other hand, taking  $x^k < x$  and  $u^k = w_* + 5 - x^k$  we have  $w_*^k \in G(u^k, x^k) = \{0\}$ . Therefore we get that the limsup in (4.43) is at least

$$\lim_{x > x^k \rightarrow x} \frac{\langle x^* | x^k - x \rangle - \langle u^* | u^k \rangle}{\|(x^k - x, u^k, 0)\|} = \lim_{5 > x^k \rightarrow x} \frac{(x^* + u^*)(x^k - x)}{\|(x^k - x, x - x^k, 0)\|} = -\frac{1}{\sqrt{2}}(x^* + u^*),$$

which is non-positive only if  $x^* \geq -u^* = 2 - 0 = 2$ . However, this can not hold simultaneously with  $x^* \leq 0$ . Therefore, no solution exists for  $u = 0$  and  $x = w_* + 5$ .

Finally, we consider the case  $x = 5$  for the limiting coderivative, i.e., we find which  $x^*$  and  $w$  satisfy  $(-u^*, x^*) \in D^*G(u, x|0)(w)$ . By the definition of limiting coderivative this is equivalent to the existence of  $(u_k^*, x_k^*) \in \widehat{D}^*G(u^k, x^k | w_*^k)(w^k)$  such that  $(-u_k^*, x_k^*, w_k^*) \rightarrow (-u^*, x^*, 0)$  with  $u^k \rightarrow u$ ,  $x^k \rightarrow x$  and  $w^k \rightarrow w$ . By the proof above, the only possibilities are

- $(-u_k^*, x_k^*) = (2, 0)$  with  $u^k = 0$ ,  $x^k > 5 + w_*^k$ ,  $w_*^k \rightarrow 0$ , and  $w^k = 0$ .
- $(-u_k^*, x_k^*) = (2 - u^k, 2 - u^k) \rightarrow (2, 2)$  with  $u^k = 5 - x^k \rightarrow 0$ ,  $x^k < 5$  and  $w^k = 2 - u^k \rightarrow 2$ .

Therefore,  $x^* = w = 0$  and  $x^* = w = 2$  are the solutions that satisfy  $(-u^*, x^*) \in D^*G(u, x|0)(w)$ .  $\square$

Recalling definitions of  $Z$  and  $\mathcal{d}_2F$  for the limiting coderivative from (3.14) and (4.5) we obtain the following corollary.

**Corollary 4.24.** *Let  $G$  be given by (4.38). Then the  $\mathcal{d}_2E$  and  $Z$  corresponding to the limiting coderivative satisfies*

$$\mathcal{d}_2E(x) = Z(x) = \begin{cases} \{x - 3\} & \text{if } 0 < x < 5, \\ \{0, 2\} & \text{if } x = 5, \\ \{0\} & \text{if } x > 5, \end{cases}$$

as well as

$$Z(u^{k+1}, x^k, \Delta w_*^{k+1}) = \begin{cases} \{2 - u^{k+1}\} & \text{if } 0 < x < 5 + \Delta w_*^{k+1} \text{ and } u^{k+1} > 0, \\ \{0, 2\} & \text{if } x = 5 + \Delta w_*^{k+1} \text{ and } u^{k+1} = 0, \\ \{0\} & \text{if } x > 5 + \Delta w_*^{k+1} \text{ and } u^{k+1} = 0. \end{cases}$$

We next show that  $E$  satisfies the set-valued descent condition (4.22).

**Lemma 4.25.** *Let function  $R = \delta_{[\varepsilon, \infty)}$ . The function  $E$  given in (4.39), satisfies the inequality (4.22) for  $\mathcal{d}E = \mathcal{d}_2E$  and  $L = 1$ .*

*Proof.* We divide the verification into several cases, depending on the values of the base point  $x^k$ , and of  $x$ :

- If  $x^k \neq 5$  and 5 is not in the interval between  $x$  and  $x^k$ , then  $E$  is Lipschitz-differentiable, and we obtain (4.22) for given  $x^k$  and  $x$  by [13, Lemma 7.1].

- If  $x^k > 5$  and  $x < 5$ , then  $x^* \in \mathcal{d}E(x^k) = \{0\}$  by [Corollary 4.24](#) and [\(4.22\)](#) reduces to

$$(4.44) \quad 0 \geq E(x) - E(x^k) - \frac{1}{2}(x - x^k)^2.$$

When  $x < 1$  we have, as required,

$$E(x) - E(x^k) - \frac{1}{2}(x - x^k)^2 < \frac{9}{2} - 2 - \frac{1}{2}(4)^2 < 3 - 8 < 0.$$

When  $x \geq 1$ , [\(4.44\)](#) also holds since  $E(x) = \frac{1}{2}(x - 3)^2 \leq 2 = E(x^k)$ .

- If  $x^k < 5$  and  $x > 5$ , we verify the claim from

$$\begin{aligned} 0 &\geq 2 - \frac{1}{2}(x - 3)^2 = 2 - \frac{1}{2}(x^k - 3 - (x^k - x))^2 \\ &= 2 - \frac{1}{2}(x^k - 3)^2 - \frac{1}{2}(x - x^k)^2 + (x^k - 3)(x^k - x). \\ &= E(x) - E(x^k) - \frac{1}{2}(x - x^k)^2 - \inf_{x^* \in \mathcal{d}E(x^k)} \langle x^* | x - x^k \rangle \end{aligned}$$

where the last equality follows from the definition of  $E$  and [Corollary 4.24](#).

- If  $x^k = 5$ , then it holds

$$\inf_{x^* \in \mathcal{d}E(x^k)} \langle x^* | x - x^k \rangle = \begin{cases} 0 & \text{if } x \geq 5, \\ 2(x - 5) & \text{if } x < 5 \end{cases}$$

by [Corollary 4.24](#). When  $x \geq 5$ , [\(4.22\)](#) again reads as [\(4.44\)](#) and holds since  $E(x) = 2 = E(x^k)$ . When  $x < 5$ , [\(4.22\)](#) reads as

$$2(x - 5) \geq \frac{1}{2}(x - 3)^2 - 2 - \frac{1}{2}(x - 5)^2.$$

We verify this condition and finish the proof with the help of

$$\frac{1}{2}(x - 3)^2 - \frac{1}{2}(x - 5)^2 = \frac{1}{2}([x - 3] + [x - 5])([x - 3] - [x - 5]) = \frac{1}{2}(2x - 8)2 = 2(x - 4). \quad \square$$

The next lemma proves that the adjoint tracking [Assumption 4.3 \(ii\)](#) holds with the forward-backward inner steps and exact adjoint solver.

**Lemma 4.26.** *Let  $f$ ,  $g$ , and  $J$  be given by [\(4.37\)](#) as well as  $G$  be given by [\(4.38\)](#). Moreover, let  $u^0 \geq 0$ ,  $x^0 \geq \varepsilon > 0$  and*

- *the inner iterate  $u^{k+1}$  be produced by the forward-backward method with step length  $\tau \in (0, 1)$ , i.e.  $u^{k+1} = \text{prox}_{\tau g(\cdot, x^k)}(u^k - \tau \nabla_u f(u^k, x^k))$ ,*
- *the adjoint iterate  $w^{k+1} \in Z(u^{k+1}, x^k, \Delta w_*^{k+1})$  for  $Z$  given by [Corollary 4.24](#) and  $\Delta w_*^{k+1} \in G(u^{k+1}, x^k)$ ,*
- *the outer iterate  $x^k = \text{proj}_{[\varepsilon, \infty)}(x^{k-1} - \sigma \tilde{x}_k^*)$  be produced with the forward-backward method for  $\sigma > 0$  and the differential estimate  $\tilde{x}_k^*$  satisfying  $(-J'(u^k), \tilde{x}_k^*) \in D^*G(u^k, x^{k-1} | \Delta w_*^k)(w^k)$ .*

*Then the adjoint tracking [Assumption 4.3 \(ii\)](#) holds for  $\mu_u, \pi_w > 0$  and  $\kappa_w > 1$  satisfying  $\mu_u \geq 1 + \kappa_w$ ,  $(1 + \kappa_w) \|\cdot\|_U \leq \mu_u \tau^{-1} \|\cdot\|_*$  and  $\pi_w \geq \sigma^{-1} \kappa_w$ .*

*Proof.* To prove that [Assumption 4.3 \(ii\)](#), it suffices to use [Theorem 4.12](#), for which we need to verify [\(4.10\)](#), which we can equivalently write as

$$(4.45) \quad \begin{aligned} & \kappa_w \operatorname{dist}(Z(x^k), w^{k+1}) + \operatorname{dist}(Z(u^{k+1}, x^k, \Delta w_*^{k+1}), w^k) \\ & \leq \operatorname{dist}(Z(x^{k-1}), w^k) + \pi_w \|x^k - x^{k-1}\| + \mu_u \left( \|u^{k+1} - S_u(x^k)\|_U + \|\Delta w_*^{k+1}\|_* \right). \end{aligned}$$

The implicit form of the forward-backward method

$$\Delta w^{k+1} = \tau^{-1}(u^{k+1} - u^k) \in \partial_u g(u^{k+1}, x^k) + \nabla_u f(u^k, x^k).$$

and the assumption  $(1 + \kappa_w) \|\cdot\|_U \leq \mu_u \tau^{-1} \|\cdot\|_*$  let us show [\(4.45\)](#) holds by proving

$$(4.46) \quad \begin{aligned} & \kappa_w \operatorname{dist}(Z(x^k), w^{k+1}) + \operatorname{dist}(Z(u^{k+1}, x^k, \Delta w_*^{k+1}), w^k) \\ & \leq \operatorname{dist}(Z(x^{k-1}), w^k) + \pi_w \|x^k - x^{k-1}\| + \mu_u \|u^{k+1} - S_u(x^k)\|_U + (1 + \kappa_w) \|u^{k+1} - u^k\|_U. \end{aligned}$$

We use the following facts throughout the proof:

(i)  $S_u(x) = \max(0, 5 - x)$  from [Lemma 4.21](#), which combined with [Corollary 4.24](#) gives the form

$$Z(x) = \begin{cases} \{2 - S_u(x)\} & \text{if } 0 < x < 5, \\ \{0, 2 - S_u(x)\} & \text{if } x = 5, \\ \{0\} & \text{if } x > 5. \end{cases}$$

(ii) [Lemma 4.23](#) proves that  $x^* = w$  for  $x^*$  and  $w$  satisfying  $(-u^*, x^*) \in D^*G(u, x|0)(w)$ . Since  $w^k \in Z(u^k, x^{k-1}, \Delta w_*^k)$ , we obtain similarly that the differential estimate  $\tilde{x}_k^* = w^k$ . Therefore we get for  $x^k > 5$  that the outer step  $x^k = \operatorname{proj}_{[\varepsilon, \infty)}(x^{k-1} - \sigma \tilde{x}_k^*) = x^{k-1} - \sigma w^k$ ,

(iii) [Theorem A.2](#) implies  $\kappa_u \|u^{k+1} - S_u(x^k)\|_U \leq \|u^k - S_u(x^k)\|_U$  for some  $\kappa_u > 1$ , since  $f$  is Lipschitz-differentiable with  $L = 1$ ,

(iv) Since the adjoint is solved exactly,  $\operatorname{dist}(Z(u^{k+1}, x^k, \Delta w_*^{k+1}), w^{k+1}) = 0$ .

(v) If  $u^k \geq 0$  and  $\tau \in (0, 1)$ , then  $u^k - \tau(u^k - 5) \geq 0$ . Thus the forward-backward step reads as

$$\begin{aligned} u^{k+1} &= \operatorname{prox}_{\tau x^k|\cdot|}(u^k - \tau(u^k - 5)) \\ &= \max(0, u^k - \tau(u^k - 5) - \tau x^k) \\ &= \max(0, (1 - \tau)u^k + \tau(5 - x^k)). \end{aligned}$$

Therefore  $u^{(0)} > 0$  implies that  $u^k \geq 0$  for all  $k \in \mathbb{N}$ .

We now prove [\(4.46\)](#) through various cases:

(a)  $u^k = 0$ , in which case we consider the sub-cases:

(a.1)  $w^k = 0$ , then we have  $x^k = x^{k-1}$  by the structure of the forward-backward updates, as observed above. We consider the following two sub-cases:

(a.1.1)  $x^k \geq 5$ , then  $S_u(x^{k-1}) = S_u(x^k) = 0$  by [\(i\)](#), and also  $u^{k+1} = 0$  because [\(iii\)](#) implies

$$\|u^{k+1} - 0\|_U = \|u^{k+1} - S_u(x^k)\|_U \leq \|u^k - S_u(x^{k-1})\|_U = \|0 - 0\|_U = 0.$$

Since  $u^{k+1} = S_u(x^k)$  and  $\Delta w_*^{k+1} = \tau^{-1}(u^{k+1} - u^k) = 0$ , we have  $Z(u^{k+1}, x^k, \Delta w_*^{k+1}) = Z(x^k)$ . Therefore, we have  $w^{k+1} = 0 \in Z(x^k)$  and  $w^k \in Z(u^{k+1}, x^k, \Delta w_*^{k+1})$  implying that the left side of [\(4.46\)](#) is zero, verifying the condition.

- (a.1.2)  $x^k < 5$ , then  $S_u(x^k) > 0$  by (i), and thus also  $u^{k+1} > 0$  since it is closer to  $S_u(x^k) > 0$  than  $u^k = 0$ . Therefore,  $w^{k+1} = 2 - u^{k+1} = Z(u^{k+1}, x^k, \Delta w_*^{k+1})$  and  $Z(x^k) = 2 - S_u(x^k)$ . The left side of the inequality (4.46) is now

$$\begin{aligned} & \kappa_w \|2 - S_u(x^k) - (2 - u^{k+1})\|_U + \|2 - u^{k+1} - 0\|_U \\ & \leq (1 + \kappa_w) \|u^{k+1} - S_u(x^k)\|_U + \|2 - S_u(x^k)\|_U \\ & = (1 + \kappa_w) \|u^{k+1} - S_u(x^k)\|_U + \|2 - S_u(x^{k-1})\|_U, \end{aligned}$$

and the right side of (4.46) is

$$\|2 - S_u(x^{k-1})\|_U + \mu_u \left( \|u^{k+1} - S_u(x^k)\|_U + \tau \|u^{k+1} - u^k\|_U \right).$$

Therefore, (4.46) holds with  $\mu_u \geq (1 + \kappa_w)$ .

- (a.2)  $w^k = 2 - u^k = 2$ , then  $x^k = x^{k-1} - \sigma 2 < x^{k-1}$ . We again consider two cases:

- (a.2.1)  $x^k < 5$ , then (i) again gives  $S_u(x^k) > 0$  implying  $u^{k+1} > 0$ . The left side of (4.46) is

$$\kappa_w \|2 - S_u(x^k) - (2 - u^{k+1})\|_U + \|2 - u^{k+1} - (2 - u^k)\|_U = \kappa_w \|u^{k+1} - S_u(x^k)\|_U + \|u^{k+1} - u^k\|_U,$$

implying the inequality holds for  $\mu_u \geq \kappa_w$ .

- (a.2.2)  $x^k \geq 5$ , then  $x^{k-1} > 5$  implying  $S_u(x^{k-1}) = S_u(x^k) = 0$ . Thus also  $u^{k+1} = 0$ . Since  $u^{k+1} = S_u(x^k)$  and  $\Delta w_*^{k+1} = 0$ , we have  $Z(u^{k+1}, x^k, \Delta w_*^{k+1}) = Z(x^k)$  implying  $w^{k+1} \in Z(x^k)$ . We can have  $Z(u^{k+1}, x^k, \Delta w_*^{k+1}) = 0$  and thus the left side of (4.46) equal 2. However, also  $Z(x^k) = \{0\}$  implying  $\text{dist}(Z(x^{k-1}), w^k) = 2$  and that the right side of (4.46) is larger than 2.

- (b)  $u^k \in (0, 2]$ , implies  $w^k = 2 - u^k \leq 0$ , and thus  $x^k \leq x^{k-1}$ . We check again sub-cases separately:

- (b.1)  $x^k \leq 5$ , then  $2 - S(x^k) \in Z(x^k)$ . We obtain

$$u^{k+1} \geq (1 - \tau)u^k + \tau(5 - x^k) \geq (1 - \tau)u^k > 0$$

from the inner step. Thus  $w^{k+1} = 2 - u^{k+1}$  and this case is analogous to (a.2.1).

- (b.2)  $x^k > 5$ , then  $Z(x^k) = 0$  and  $Z(x^{k-1}) = 0$ . Since  $w^{k+1} \in Z(u^{k+1}, x^k, \Delta w_*^{k+1}) \subset \{0, 2 - u^{k+1}\}$  by Corollary 4.24, we can have only either:

- (b.2.1)  $w^{k+1} = 0$ , then the left side of (4.46) is at most  $\|2 - u^k\|_U$ , which is less than the right side that involves  $\text{dist}(Z(x^{k-1}), w^k) = \|2 - u^k\|_U$ .

- (b.2.2)  $w^{k+1} = 2 - u^{k+1}$ , then the left side of (4.46) is

$$\kappa_w \|2 - u^{k+1}\|_U + \|u^{k+1} - u^k\|_U \leq \kappa_w \|2 - u^k\|_U + (1 + \kappa_w) \|u^{k+1} - u^k\|_U.$$

The right side is

$$\|2 - u^k\|_U + \pi_w \|x^k - x^{k-1}\| + \mu_u \|u^{k+1} - S_u(x^k)\|_U + (1 + \kappa_w) \|u^{k+1} - u^k\|_U.$$

Using the fact (ii) we have  $x^k - x^{k-1} = \sigma w^k = \sigma(2 - u^k)$ , which verifies (4.46) for  $\pi_w \geq \sigma^{-1}(\kappa_w - 1)$ .

- (c)  $u^k > 2$ , then  $w^k = 2 - u^k < 0$ . We consider the sub-cases:

- (c.1)  $x^k \leq 5$ , which is analogous to (b.1).

(c.2)  $x^k > 5$ , which we divide into two further sub-cases based on the implication of [Corollary 4.24](#),  $Z(u^{k+1}, x^k, \Delta w_*^{k+1}) \subset \{0, 2 - u^{k+1}\}$ :

(c.2.1)  $w^{k+1} \in Z(u^{k+1}, x^k, \Delta w_*^{k+1}) = \{0\}$ , then either  $u^{k+1} = 0$  or  $u^{k+1} = 2$ . In both cases

$$\kappa_w \text{dist}(Z(x^k), w^{k+1}) + \text{dist}(Z(u^{k+1}, x^k, \Delta w_*^{k+1}), w^k) = 0 + \|2 - u^k\|_U \leq \|u^{k+1} - u^k\|_U,$$

which immediately verifies [\(4.46\)](#).

(c.2.2)  $(2 - u^{k+1}) \in Z(u^{k+1}, x^k, \Delta w_*^{k+1})$ , then (if  $0 \notin Z(u^{k+1}, x^k, \Delta w_*^{k+1})$ ) the left side of [\(4.46\)](#) is at most

$$\begin{aligned} \kappa_w \|2 - u^{k+1}\|_U + \|u^{k+1} - u^k\|_U &\leq \kappa_w \|2 - u^k\|_U + (1 + \kappa_u) \|u^{k+1} - u^k\|_U \\ &= \kappa_w \sigma^{-1} \|x^k - x^{k-1}\| + (1 + \kappa_u) \|u^{k+1} - u^k\|_U \end{aligned}$$

where the last equality follows from (ii). Therefore, [\(4.46\)](#) holds for  $\pi_w \geq \sigma^{-1} \kappa_w$ .  $\square$

**Remark 4.27.** If the inner iterates  $u^{k+1}$  are good enough approximations of solutions  $S_u(x^k)$ , the error is less than 2, the case can't happen. Then we can take less strict assumptions for  $\pi_w$ , i.e.  $\pi_w = \sigma^{-1}(\kappa_w - 1)$  is enough.

Finally, we can prove the assumptions of our main convergence [Theorem 4.18](#).

**Theorem 4.28.** Let  $J, R, f, g$ , and  $G$  be as in the example problem [\(4.37\)](#) and [\(4.38\)](#), and  $E$  be given by [\(4.39\)](#), i.e.  $E = J \circ S_u$  for the inner solution mapping  $S_u$  satisfying  $0 \in G(S_u(x), x)$ . Moreover, let

- the inner iterate  $u^{k+1}$  be produced by the forward-backward method with step length  $\tau \in (0, 1)$ , i.e.,  $u^{k+1} = \text{prox}_{\tau g(\cdot, x^k)}(u^k - \tau \nabla_u f(u^k, x^k))$ ,
- the adjoint iterate  $w^{k+1} \in Z(u^{k+1}, x^k, \tau^{-1}(u^{k+1} - u^k))$  for  $Z$  given by [Corollary 4.24](#),
- the outer iterate  $x^k = \text{proj}_{[0, \infty)}(x^{k-1} - \sigma x_k^*)$  be produced with the forward-backward method for  $\sigma > 0$  and the differential estimate  $\tilde{x}_k^*$  arising from  $(-J'(u^k), \tilde{x}_k^*) \in D^*G(u^k, x^{k-1} | \tau^{-1}(u^{k+1} - u^k))(w^k)$ .

Assume the step lengths  $\tau, \sigma$  and the constants  $\mu_u, \pi_w > 0$  and  $\kappa_w > 1$  to satisfy  $\mu_u \geq 1 + \kappa_w$ ,  $(1 + \kappa_w) \|\cdot\|_U \leq \mu_u \tau^{-1} \|\cdot\|_*$ ,  $\pi_w \geq \sigma^{-1} \kappa_w$ , and  $\sigma(2 + \zeta_1^2) < 2$  where  $\kappa := \min\{\kappa_u, \kappa_w\} > 1$  and  $\zeta_1 > 0$  satisfies the bound [\(4.26\)](#). Then  $\|u^{k+1} - S_u(x^k)\|_U \rightarrow 0$  and  $\inf_{x^* \in [\mathcal{d}E + \partial R](x^{k+1})} \|x^*\| \rightarrow 0$ .

*Proof.* The condition  $\inf[E + R] > -\infty$  holds because  $E, R \geq 0$ . We can take  $\Omega_X = [\varepsilon, \infty)$ , which implies the condition [\(4.23\)](#). Moreover, the set-valued descent inequality [\(4.22\)](#) holds for  $\mathcal{d}E = \mathcal{d}_2E$  and  $L = 1$  by [Lemma 4.25](#). Therefore, we can take  $\mathcal{d}_2E(x^k)$  given by [Corollary 4.24](#) as the target sets. Our  $X = \mathbb{R}$  is also Hilbert space.

We next consider [Assumption 4.3](#):

- Since  $S_u$  is Lipschitz continuous (and single-valued) by [Lemma 4.21](#) and the forward-backward method satisfies the basic contractivity property [\(4.7\)](#) by [Theorem A.2](#), [Theorem 4.9](#) implies the inner tracking property, [Assumption 4.3 \(i\)](#)
- We verify the adjoint tracking property, [Assumption 4.3 \(ii\)](#), in [Lemma 4.26](#).
- [Lemma 4.23](#) and [Corollary 4.24](#) demonstrate that we have  $\tilde{x}_{k+1}^* = w^{k+1} \in Z(u^{k+1}, x^k, \Delta w_*^{k+1})$  and  $\mathcal{d}_2E(x^k) \ni x_{k+1}^* = S_w(x^k) \in Z(x^k)$ . Therefore, taking  $d_{X^*} = \|\cdot\|$ , we see that

$$\|\tilde{x}_{k+1}^* - x_{k+1}^*\| \leq \|w^{k+1} - S_w(x^k)\|_W.$$

This shows that the differential transformation [Assumption 4.3 \(iii\)](#) holds for  $\alpha_u = 0$  and  $\alpha_w = 1$ .

Therefore, [Assumption 4.3](#) holds.

We finally verify the condition (ii)(a) of [Theorem 4.18](#), i.e., for all  $\varepsilon > 0$  there exists  $\delta > 0$  such that, for some  $k_N \in \mathbb{N}$  for any  $k \geq k_N$ ,

$$(4.47) \quad \|x^{k+1} - x^k\| \leq \delta \implies \text{dist}(\mathcal{d}E(x^{k+1}), x_{k+1}^*) \leq \varepsilon.$$

If either  $x^k > 5$  or  $x^k < 5$  for all  $k \geq k_N$ , the implication (4.47) holds since  $\mathcal{d}E$  is Lipschitz-continuous by [Corollary 4.24](#). We consider next the case  $\mathcal{d}E(x^{k+1}) = 0$  and  $x_{k+1}^* = x - 3 = 2 - S_u(x^k)$ , i.e.  $x^{k+1} \geq 5$  and  $x^k \leq 5$  with  $x^{k+1} \neq x^k$ . By [Lemma 4.16](#) we have  $u^{k+1} - S_u(x^k) \rightarrow 0$  and thus for  $k$  large enough also  $\bar{x}_{k+1}^* \approx x_{k+1}^*$ . However, this implies  $\|x^{k+1} - x^k\| = \|\sigma \bar{x}_{k+1}^*\| \approx \sigma(5 - 3) = 2\sigma$ , which is contradiction to  $\|x^{k+1} - x^k\| \leq \delta$ . The case  $\mathcal{d}E(x^{k+1}) = x - 3 = 2 - S_u(x^k)$  and  $x_{k+1}^* = 0$  produces contradiction similarly. Therefore, (4.47) must hold for some  $k_N \in \mathbb{N}$  for any  $k \geq k_N$ . We have proved that the assumptions of [Theorem 4.18](#) hold, and thus the claim follows from it.  $\square$

**Remark 4.29.** Note, however, that since  $E = J \circ S_u$  is nonconvex, the convergence given by [Theorem 4.18](#),  $\inf_{\bar{x}^* \in [\mathcal{d}E + \partial R](x^{k+1})} \|\bar{x}^*\| \rightarrow 0$ , could be satisfied in local optima, in  $[5, \infty)$  where  $E$  is constant.

## 5 TOTAL VARIATION REGULARISED INVERSE PROBLEMS

In this section, we base the inner problem on the ill-posed inverse problem

$$A_x u = m + v,$$

where  $A_x$  is a forward operator corresponding to the (hyper)parameters  $x$ . It maps the unknown  $u$  into corresponding measurable data. The measurement is  $m$ , however, it may deviate from the true data by noise  $v$ . Since the dimension of the measurements  $m$  is likely smaller than the dimension of  $x$ , the system will not have a unique solution. The solution may also be highly sensitive to the noise. Incorporating a data fidelity  $d$  to model the noise, and adding (discretised) total variation regularisation to introduce our preconceptions of a good solution, we obtain the inner optimisation problem

$$(5.1) \quad \min_u d(A_x u, m) + C(x) \|Ku\|_{2,1}$$

Here  $K$  is a discretised differential operator. The bilevel problem with (5.1) as the inner problem arises from the need to choose an optimal regularisation parameter  $C(x)$  and a measurement operator  $A_x$  with respect to the outer objective. The corresponding outer problem reads as

$$(5.2) \quad \min_x J(u) + R(x) \quad \text{such that } u \text{ solves (5.1).}$$

We recall that we write  $F := J \circ S$ , where  $S : x \mapsto u$  is the solution mapping of (5.1).

The regularisation term  $\|Ku\|_{2,1}$  is nonsmooth in  $u$ , and although convex, its proximal map is hard to evaluate. We therefore use the primal-dual algorithm of [Example 4.2](#) for the solution of (5.1). [Examples 4.1](#) and [4.2](#) provide generic inner and outer algorithms for the bilevel problem (5.2). However, the adjoint algorithm needs to be specifically developed for this case. That is the goal of this section.

First, in [Section 5.1](#) we provide general results on the adjoint inclusion, where both the data and regularisation terms in the inner problem can depend on the parameter  $x$ . Then, in [Section 5.2](#), we refine the results for the case where only the data term depends on  $x$ . This makes the differential transformation [Assumption 4.3 \(iii\)](#) easier to satisfy as discussed in [Section 4.5](#). All the spaces throughout this section are Hilbert spaces, so we are justified in identifying the spaces with their dual spaces. In particular, the inner optimality condition  $G : U \times X \rightrightarrows W_*$  where we can take  $W_* = W = U$ , because  $G(\cdot, x)$  is a subdifferential in a Hilbert space. Thus also the coderivatives  $\check{D}^*G(u, x|w) : U \rightrightarrows U \times X$ .

### 5.1 GENERAL RESULTS FOR THE ADJOINT INCLUSION

We consider first more general inner problem (primal) formulation that encompasses (5.1),

$$(5.3) \quad \min_{u_p} f(u_p, x) + g(Ku_p, x),$$

with  $f$  and  $g$  convex functions. The corresponding bilevel problem is

$$(5.4) \quad \min_{u \in U_p \times U_d, x \in X} J(u) + R(x) \quad \text{subject to} \quad 0 \in G(u, x) := \begin{pmatrix} \nabla_{u_p} f(u_p, x) + K^* u_d \\ -K u_p + \partial_{u_d} g^*(u_d, x) \end{pmatrix},$$

where the primal-dual optimality condition  $0 \in G(u, x)$  arises from (5.3) as discussed in Example 4.2.

The next lemma analyses the adjoint inclusion

$$(5.5) \quad (-u^*, x^*) \in \check{D}^* G(u, x | \Delta w_*)(w)$$

for  $\check{D}^* = \widehat{D}^*$  or  $\check{D}^* = D^*$ . For the basic adjoint inclusion in (3.15), we take

$$(5.6) \quad u^* = J'(u).$$

For the general step (4.1b) with  $u = u^{k+1}$ , we would have  $u^* = J'(u^{k+1}) - \Delta u_{k+1}^*$  for a perturbation  $\Delta u_{k+1}^* \approx 0$ . In practise, we will have  $\Delta u_{k+1}^*$  arise implicitly from the inexact solution of (5.5) for  $u^* = J'(u^{k+1})$ .

**Lemma 5.1.** *Let  $U = U_p \times U_d$  and  $X$  be Hilbert spaces,  $K \in \mathbb{L}(U_p; U_d)$  and  $G : U \times X \rightrightarrows U$ . Let  $G$  be as in (5.4) for  $f : U_p \times X \rightarrow \mathbb{R}$  twice continuously differentiable and  $g^* : U_d \times X \rightrightarrows \overline{\mathbb{R}}$  convex, proper, and lower semicontinuous. Pick  $x, x^* \in X$ ,*

$$u = (u_p, u_d) \in U, \quad \Delta w_* = (\Delta w_p, \Delta w_d) \in G(u, x), \quad \text{and} \quad w = (w_p, w_d) \in U.$$

*Then, for any  $u^* \in U$  and  $\check{D}^* = \widehat{D}^*$  or  $\check{D}^* = D^*$ , the adjoint inclusion (5.5) holds, i.e.,  $(-u^*, x^*) \in \check{D}^* G(u, x | \Delta w_*)(w)$ , if and only if*

$$(5.7a) \quad 0 = u^* + \begin{pmatrix} \nabla_{u_p}^2 f(u_p, x) w_p - K^* w_d \\ K w_p \end{pmatrix} + \begin{pmatrix} 0 \\ u_g^* \end{pmatrix} \quad \text{and}$$

$$(5.7b) \quad x^* = \nabla_{x u_p} f(u_p, x) w_p + x_g^*$$

for some

$$(5.7c) \quad (u_g^*, x_g^*) \in \check{D}^* \partial_{u_d} g^*(u_d, x | \Delta w_d + K u_p)(w_d) \subset U_d \times X.$$

*Proof.* Split

$$G = G_1 + G_2 \quad \text{for} \quad G_1(u, x) := \begin{pmatrix} \nabla_{u_p} f(u_p, x) + K^* u_d \\ -K u_p \end{pmatrix} \quad \text{and} \quad G_2(u, x) := \begin{pmatrix} 0 \\ \partial_{u_d} g^*(u_d, x) \end{pmatrix}.$$

The coderivative sum rule [37, Theorem 1.62] gives

$$(5.8) \quad \check{D}^* G(u, x | \Delta w_*)(w) = \check{D}^* G_2(u, x | \Delta w_* - G_1(u, x))(w) + [G_1'(u, x)]^* w.$$

We have

$$[G_1'(u, x)]^* w = ([G_1^{(u)}(u, x)]^* w, [G_1^{(x)}(u, x)]^* w)$$

with

$$(5.9) \quad [G_1^{(u)}(u, x)]^* w = \begin{pmatrix} \nabla_{u_p}^2 f(u_p, x) & -K^* \\ K & 0 \end{pmatrix} \begin{pmatrix} w_p \\ w_d \end{pmatrix} = \begin{pmatrix} \nabla_{u_p}^2 f(u_p, x) w_p - K^* w_d \\ K w_p \end{pmatrix} \quad \text{and}$$

$$(5.10) \quad [G_1^{(x)}(u, x)]^* w = \begin{pmatrix} \nabla_{x u_p} f(u_p, x) & 0 \end{pmatrix} \begin{pmatrix} w_p \\ w_d \end{pmatrix} = \nabla_{x u_p} f(u_p, x) w_p.$$

We finish the proof by showing

$$\check{D}^* G_2(u, x | \Delta w_* - G_1(u, x))(w) = \{((0, u_g^*), x_g^*) \in U \times X \mid (u_g^*, x_g^*) \in \check{D}^* \partial_{u_d} g^*(u_d, x | \Delta w_d + K u_p)(w_d)\},$$

which combined with (5.8) to (5.10) yields the claim (5.7).

By the definition of Fréchet coderivative  $((u_p^*, u_g^*), x_g^*) \in \check{D}^* G_2(u, x | \Delta w_* - G_1(u, x))(w)$  is equivalent to

$$(5.11) \quad \limsup_{x^k \rightarrow x, u^k \rightarrow u, G_2(u^k, x^k) \ni w_*^k \rightarrow w_*} \frac{-\langle w, w_*^k - w_* \rangle_U + \langle x_g^*, x^k - x \rangle_X + \langle (u_p^*, u_g^*), u^k - u \rangle_U}{\|(x^k - x, u^k - u, w_*^k - w_*)\|} \leq 0$$

with  $w_* = \Delta w_* - G_1(u, x)$ . The operator  $G_2$  doesn't depend on  $u_p^*$  and thus  $w_* \in G_2(u, x)$  implies  $w_* \in G_2(u + \tau_k(u_p^*), x)$ . Therefore we can take  $u^k = u + \tau_k(u_p^*)$  with  $\tau_k \rightarrow 0$ ,  $x^k = x$  and  $w_*^k = w_*$ . We get that the limsup in (5.11) is at least  $\langle u_p^*, \tau_k u_p^* \rangle / \|\tau_k u_p^*\| = \|u_p^*\|$ , and thus  $u_p^* = 0$  is necessary for (5.11) to hold.

Because  $\Delta w_* \in G(u, x)$  it holds  $w_* \in G(u, x) - G_1(u, x) = G_2(u, x)$  and  $w_* = (0, \Delta w_d + K u_p)$ . Similarly the primal component of every  $w_*^k$  has to be zero. Thus we obtain (5.11) is equivalent to

$$\limsup_{\substack{x^k \rightarrow x, u_d^k \rightarrow u_d, \\ \partial_{u_d} g^*(u_d^k, x^k) \ni w_{*,d}^k \rightarrow \Delta w_d + K u_p}} \frac{-\langle w_d, w_{*,d}^k - (\Delta w_d + K u_p) \rangle_{U_d} + \langle x_g^*, x^k - x \rangle_X + \langle u_g^*, u_d^k - u_d \rangle_{U_d}}{\|(x^k - x, u_d^k - u_d, w_{*,d}^k - (\Delta w_d + K u_p))\|} \leq 0,$$

which by the definition equals  $(u_g^*, x_g^*) \in \widehat{D}^* \partial_{u_d} g^*(u_d, x | \Delta w_d + K u_p)(w_d)$ . Since  $u_p^* = 0$  for all Fréchet coderivatives (of  $G_2$ ) it also has to hold for limiting coderivatives.  $\square$

**Remark 5.2.** To solve (5.7), we only need to find  $w = (w_p, w_d)$ ,  $u_g^*$ , and  $x_g^*$  satisfying (5.7a) and (5.7c). Then (5.7b) directly produces  $x^*$ . The former two have a solution if and only if there exists a solution – with the optimal value equal to zero – to the optimisation problem

$$(5.12) \quad \min_{w_p \in U_p, w_d, u_g^* \in U_d, x_g^* \in X} \frac{1}{2} \|u_p^* + \nabla_{u_p}^2 f(u_p, x) w_p - K^* w_d\|_{U_p}^2 + \frac{1}{2} \|u_d^* + K w_p + u_g^*\|_{U_d}^2$$

subject to  $(u_g^*, x_g^*) \in \check{D}^* \partial_{u_d} g^*(u_d, x | \Delta w_d + K u_p)(w_d)$ .

The structure of this problem suggests to solve it with forward-backward splitting. We return to this in Section 5.3, after forming  $\check{D}^* \partial_{u_d} g^*$ .

## 5.2 ADJOINT INCLUSION FOR NON-PARAMETRIC REGULARISATION TERM

In the numerical experiments of Section 6, we concentrate on (5.1) with only the forward operator  $A_x$  dependent on the parameter  $x$ . That is,  $C(x) \equiv C$  is a constant. To apply Section 5.1 to this setting, we take

$$g(u_d; x) := C \|u_d\|_{2,1} := C \sum_{j=1}^n \|u_{d,j}\|_2.$$

Then the Fenchel conjugate

$$(5.13) \quad g^*(u_d, x) = \sum_{j=1}^n \delta_{B_{\mathbb{R}^2}(0, C)}(u_{d,j}).$$

Its subdifferential

$$\partial_{u_{d,j}} g^*(u_d, x) = \begin{cases} \{0\} & \text{if } \|u_{d,j}\| < C, \\ u_{d,j}[0, \infty) & \text{if } \|u_{d,j}\| = C, \\ \emptyset & \text{otherwise.} \end{cases}$$

Recalling Lemma 5.1, we need the coderivatives of  $\partial_u g^*$ . By separability, it suffices to study the component coderivatives.

**Lemma 5.3.** *Let  $z, y_*, y \in \mathbb{R}^2, C > 0$ . Then*

$$(5.14) \quad \widehat{D}^* \partial \delta_{B_{\mathbb{R}^2}(0, C)}(z|y_*)(y) = \begin{cases} \{0\} & \text{if } \|z\| < C, y_* = 0, y \in \mathbb{R}^2, \\ z\mathbb{R} + \frac{\|y_*\|}{\|z\|} y & \text{if } \|z\| = C, y_* \in z(0, \infty), \langle y, z \rangle = 0, \\ z[0, \infty) & \text{if } \|z\| = C, y_* = 0, \langle y, z \rangle \geq 0, \\ \emptyset & \text{otherwise} \end{cases}$$

and

$$(5.15) \quad D^* \partial \delta_{B_{\mathbb{R}^2}(0, C)}(z|y_*)(y) = \begin{cases} \{0\} & \text{if } \|z\| < C, y_* = 0, y \in \mathbb{R}^2, \\ z\mathbb{R} + \frac{\|y_*\|}{\|z\|} y & \text{if } \|z\| = C, y_* \in z[0, \infty), \langle y, z \rangle = 0, \\ z[0, \infty) & \text{if } \|z\| = C, y_* = 0, \langle y, z \rangle > 0, \\ \{0\} & \text{if } \|z\| = C, y_* = 0, \langle y, z \rangle < 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

See proof in Appendix B.

**Lemma 5.4.** *Let  $U_d = \prod_{j=1}^n \mathbb{R}^2, X$  be a Hilbert space,  $u_d, w_*, w_d \in U_d, x \in X$ . Moreover, let function  $g^* : U_d \times X \rightarrow \overline{\mathbb{R}}$  be defined as in (5.13) for some  $C > 0$ . Then*

$$\widehat{D}^* \partial_{u_d} g^*(u_d, x|w_*)(w_d) = \left( \prod_{j=1}^n \widehat{D}^* \partial \delta_{B_{\mathbb{R}^2}(0, C)}(u_{d,j}|w_{*,j})(w_{d,j}) \right) \times \{0\}.$$

*Proof.* Follows from the definition of the Fréchet coderivative and the separable structure □

With Lemmas 5.3 and 5.4 we can expand (5.7c), and thus write the optimisation problem (5.12) in a more explicit form.

**Lemma 5.5.** *Let  $U_p$  and  $X$  be Hilbert spaces,  $U_d = \prod_{j=1}^n \mathbb{R}^2, U = U_p \times U_d$  and  $K \in \mathbb{L}(U_p; U_d)$ . Let  $G$  be as in (5.4) for  $f : U_p \times X \rightarrow \mathbb{R}$  twice continuously differentiable and  $g^* : U_d \times X \rightarrow \overline{\mathbb{R}}$  as in (5.13) for some  $C > 0$ . Suppose  $u_p \in U_p, u_d, w_d \in U_d$  and  $x \in X$  as well as*

$$\Delta w_* = (\Delta w_p, \Delta w_d) \in G(u, x), \quad \text{i.e.,} \quad \Delta w_d \in -K u_p + \partial_{u_d} g^*(u_d, x) \quad \text{and} \quad w_* = \Delta w_d + K u_p.$$

Moreover, let  $\lambda \odot : U_d \rightarrow U_d$  and  $V : U_d \rightarrow U_d$ , for fixed  $u_d$  be defined by

$$(5.16) \quad [\lambda \odot u_d]_j = \lambda_j u_{d,j} \quad \text{and} \quad [V w_d]_j := \begin{cases} \frac{\|w_{*,j}\|}{\|u_{d,j}\|} w_{d,j} & \text{if } \|u_{d,j}\| = C, w_{*,j} \in u_{d,j}(0, \infty), \\ 0 & \text{otherwise.} \end{cases}$$

Then the adjoint inclusion (5.5) holds for  $\check{D}^* = \widehat{D}^*$ , i.e.,  $(-u^*, x^*) \in \widehat{D}^*G(u, x|\Delta w_*)(w)$ , if and only if (5.7a) and (5.7b) hold with

$$(5.17) \quad x_g^* = 0, u_g^* = \lambda \odot u_d + Vw_d \quad \text{with} \quad \begin{cases} \lambda_j = 0 & \text{if } \|u_{d,j}\| < C, w_{*,j} = 0 \\ \lambda_j \in \mathbb{R}, \langle w_{d,j}, u_{d,j} \rangle = 0 & \text{if } \|u_{d,j}\| = C, w_{*,j} \in u_{d,j}(0, \infty), \\ \langle w_{d,j}, u_{d,j} \rangle \geq 0, \lambda_j \geq 0 & \text{if } \|u_{d,j}\| = C, w_{*,j} = 0 \end{cases}$$

For  $\check{D}^* = \widehat{D}^*$ , the adjoint inclusion (5.5) holds if and only if (5.7a) and (5.7b) hold and

$$(5.18) \quad \begin{cases} \lambda_j = 0 & \text{if } \|u_{d,j}\| < C, w_{*,j} = 0 \\ \langle w_{d,j}, u_{d,j} \rangle = 0 & \text{if } \|u_{d,j}\| = C, w_{*,j} \in u_{d,j}(0, \infty), \\ \left\{ \begin{array}{l} \lambda_j \langle w_{d,j}, u_{d,j} \rangle \geq 0 \text{ and} \\ \max(\langle w_{d,j}, u_{d,j} \rangle, \lambda_j) \geq 0 \end{array} \right\} & \text{if } \|u_{d,j}\| = C, w_{*,j} = 0 \end{cases}$$

*Proof.* By Lemma 5.1, we only need to prove the equivalence of (5.7c), i.e.,

$$(u_g^*, x_g^*) \in \check{D}^* \partial_{u_d} g^*(u_d, x|\Delta w_d + Ku_p)(w_d)$$

to (5.17) or (5.18), depending on the choice of the coderivative. By assumption,  $\Delta w_d + Ku_p \in \partial_{u_d} g^*(u_d, x)$ . Therefore, Lemma 5.4 establishes that  $x_g^* = 0$  and  $[u_g^*]_j \in \widehat{D}^* \partial \delta_{B_{\mathbb{R}^2}(0,C)}(u_{d,j}|w_{*,j})(w_{d,j})$  for all  $j = 1, \dots, n$ , when (5.7c) holds. This reduces our task to proving that (5.17) or (5.18) gives for all  $j = 1, \dots, n$  the same expression for  $[u_g^*]_j$  as Lemma 5.3 does.

We start with the Fréchet coderivative.

- (a)  $\|u_{d,j}\| < C$  and  $w_{*,j} = 0$ . Then, by the definition in (5.16),  $[\lambda \odot u_d + Vw_d]_j = \lambda_j u_{d,j}$ . Now (5.17) yields  $x_g^* = 0$  and  $u_g^* = 0$ . The equivalence now follows from Lemmas 5.3 and 5.4, which also give the only option  $[u_g^*]_j = 0$  for (5.7c) to hold.
- (b)  $\|u_{d,j}\| = C$  and  $w_{*,j} \in u_{d,j}(0, \infty)$ . Then, by the definition in (5.16),  $[\lambda \odot u_d + Vw_d]_j = \lambda_j u_{d,j} + \frac{\|w_{*,j}\|}{\|u_{d,j}\|} w_{d,j}$ . Now (5.17) yields in agreement with Lemma 5.3 that  $[u_g^*]_j \in \mathbb{R}u_{d,j} + \frac{\|w_{*,j}\|}{\|u_{d,j}\|} w_{d,j}$  provided that  $\langle w_{d,j}^*, u_{d,j} \rangle = 0$ , with no solution otherwise.
- (c)  $\|u_{d,j}\| = C$  and  $w_{*,j} = 0$ . Then, by the definition in (5.16),  $[\lambda \odot u_d + Vw_d]_j = \lambda_j u_{d,j}$ . Now (5.17) yields in agreement with Lemma 5.3 that  $[u_g^*]_j \in [0, \infty)u_{d,j}$  provided that  $\langle w_{d,j}, u_{d,j} \rangle \geq 0$ , with no solution otherwise.

For the limiting coderivative cases (a) and (b) work out exactly in the same way, while case (c) is somewhat more tedious to prove. Write

$$(5.19) \quad \begin{aligned} A &:= \{[\lambda \odot u_d + Vw_d]_j \mid \lambda_j \langle w_{d,j}, u_{d,j} \rangle \geq 0 \text{ and } \max(\langle w_{d,j}, u_{d,j} \rangle, \lambda_j) \geq 0\} \\ &= \{\lambda_j u_{d,j} \mid \lambda_j > 0 \text{ and } \langle w_{d,j}, u_{d,j} \rangle > 0, \text{ or } \lambda_j = 0, \text{ or } \langle w_{d,j}, u_{d,j} \rangle = 0\} \end{aligned}$$

and

$$B := \{u_{g,j}^* \mid (u_g^*, x_g^*) \in \check{D}^* \partial_{u_d} g^*(u_d, x|\Delta w_d + Ku_p)(w_d)\}.$$

We have to show that  $A \subset B$  and  $B \subset A$ , which both require going through three cases. To show that  $A \subset B$ , let  $t \in A$ , and consider the cases:

- (c.1)  $\lambda_j > 0$  and  $\langle w_{d,j}, u_{d,j} \rangle > 0$ : then  $t = \lambda_j u_{d,j} \in u_{d,j}[0, \infty)$ . Third row of (5.15) now yields  $t \in B$ .
- (c.2)  $\lambda_j = 0$ : then  $t = 0$  and, since  $t \in \{0\}, t \in u_{d,j}[0, \infty)$  and  $t \in u_{d,j}\mathbb{R} + 0$ , (5.15) yields  $t \in B$ .
- (c.3)  $\langle w_{d,j}, u_{d,j} \rangle = 0$ : then  $t = \lambda_j u_{d,j} \in u_{d,j}\mathbb{R}$ , or the second row of (5.15) yields  $t \in B$ .

To show that  $B \subset A$ , let  $t \in B$ , and consider the cases:

- (c.4)  $\langle w_{d,j}, u_{d,j} \rangle = 0$ : because  $w_{*,j} = 0$ , (5.15) gives  $t = \lambda_j u_{d,j}$  for  $\lambda_j \in \mathbb{R}$ . But now  $t \in A$  by (5.19).  
(c.5)  $\langle w_{d,j}, u_{d,j} \rangle > 0$ : again (5.15) gives  $t = \lambda_j u_{d,j}$  for  $\lambda_j \geq 0$ , and therefore  $t \in A$  by (5.19).  
(c.6)  $\langle w_{d,j}, u_{d,j} \rangle < 0$ : then (5.15) gives  $t = 0$  and so (5.19) yields  $t \in A$ .

We have thus proved that  $u_g^* = \lambda \odot u_d + V w_d$  with  $\lambda_j$  and  $w_{d,j}$  constrained as in (5.17) for  $\check{D}^* = \widehat{D}^*$ , and as in (5.18) for  $\check{D}^* = D^*$ .  $\square$

Reformulating the third case of (5.18) into two branches, as well as replacing (5.7a) and (5.7c) with the optimisation problem (5.12), we obtain:

**Corollary 5.6.** *Suppose the assumptions of Lemma 5.5 hold. Define the adjoint objective function  $\mathcal{L}$  as*

$$(5.20) \quad \mathcal{L}(w_p, w_d, \lambda) := \frac{1}{2} \|u_p^* + \nabla_{u_p}^2 f(u_p, x) w_p - K^* w_d\|_{U_p}^2 + \frac{1}{2} \|u_d^* + K w_p + \lambda \odot u_d + V w_d\|_{U_d}^2,$$

the constraint set corresponding to the Fréchet coderivative as

$$(5.21) \quad \check{\mathcal{C}}_d := \left\{ (w_d, \lambda) \in U_d \times \mathbb{R}^n \left| \begin{array}{ll} \lambda_j = 0 & \text{if } \|u_{d,j}\| < C, [\Delta w_d + K u_p]_j = 0 \\ \langle w_{d,j}, u_{d,j} \rangle = 0, & \text{if } \|u_{d,j}\| = C, [\Delta w_d + K u_p]_j \in u_{d,j}(0, \infty) \\ \langle w_{d,j}, u_{d,j} \rangle \geq 0, \lambda_j \geq 0 & \text{if } \|u_{d,j}\| = C, [\Delta w_d + K u_p]_j = 0 \end{array} \right. \right\}$$

and the constraint set corresponding to limiting coderivative as

$$(5.22) \quad \mathcal{C}_d := \left\{ (w_d, \lambda) \in U_d \times \mathbb{R}^n \left| \begin{array}{ll} \lambda_j = 0 & \text{if } \|u_{d,j}\| < C, [\Delta w_d + K u_p]_j = 0 \\ \langle w_{d,j}, u_{d,j} \rangle = 0, & \text{if } \|u_{d,j}\| = C, [\Delta w_d + K u_p]_j \in u_{d,j}(0, \infty) \\ \left\{ \begin{array}{l} \lambda_j \geq 0 \text{ and } \langle w_{d,j}, u_{d,j} \rangle \geq 0; \text{ or} \\ \lambda_j \leq 0 \text{ and } \langle w_{d,j}, u_{d,j} \rangle = 0 \end{array} \right\} & \text{if } \|u_{d,j}\| = C, [\Delta w_d + K u_p]_j = 0 \end{array} \right. \right\}.$$

In the definitions of both constraint sets,  $j$  runs over  $1, \dots, n$  in the qualifier. Then (5.5) holds, i.e.,  $(-u^*, x^*) \in \check{D}^* G(u, x | \Delta w_*)(w)$ , if and only if,  $x^* = \nabla_{x u_p} f(u_p, x) w_p$  for some  $w = (w_p, w_d)$  and  $\lambda \in \mathbb{R}^n$  that achieve

$$(5.23) \quad 0 = \min_{w_p \in U_p, w_d \in U_d, \lambda \in \mathbb{R}^n} \mathcal{L}(w_p, w_d, \lambda) \quad \text{subject to} \quad (w_d, \lambda) \in \check{\mathcal{C}}_d,$$

where  $\check{\mathcal{C}}_d = \mathcal{C}_d$  or  $\check{\mathcal{C}}_d = \check{\mathcal{C}}_d$  according to the choice of the coderivative  $\check{D}^* = D^*$  or  $\check{D}^* = \widehat{D}^*$ .

**Remark 5.7.** For the Fréchet coderivative, (5.23) is a finite-dimensional and convex constrained quadratic problem, so has a minimiser. For the limiting coderivative, the problem may be nonconvex in the third case of (5.22), but again, by considering the two convex branches individually, we obtain the existence of a minimiser. The main question, therefore, is if the minimum value is zero. This is tied to the satisfaction of (4.17), which we use to guarantee the differential transformation Assumption 4.3 (iii). The latter guarantees the existence of a solution to (5.5), hence to (5.23).

**Example 5.8 (Denoising as the inner problem).** Take  $f(u_p, x) = \frac{x}{2} \|u_p - m\|^2$  for some noisy measurement  $m \in U_p$  and therefore  $\nabla_{u_p}^2 f(u_p, x) w_p = x w_p$ . Taking  $J(u) = \frac{1}{2} \|u_p - b\|^2$  for a target “true solution”  $b \in U_p$ , recalling (5.6), we have  $u_p^* = u_p - b$  and  $u_d^* = 0$ . Since the optimal value of the adjoint problem (5.23) is zero when the original adjoint inclusion (5.5) has a solution, we can solve the problem for  $w_p$  by setting the first term in  $\mathcal{L}(w_p, w_d, \lambda)$  equal to zero. This yields

$$w_p = x^{-1} (K^* w_d + b - u_p).$$

**Algorithm 5.1** Nonsmooth Bilevel Primal-Dual method

**Require:** On Hilbert spaces  $U = U_p \times U_d$  and  $X$ , functions  $J : U \rightarrow \mathbb{R}, R : X \rightarrow \overline{\mathbb{R}}, f = f_0 + e$  for  $f_0, e : U_p \times X \rightarrow \overline{\mathbb{R}}$  and  $g^* : U_d \times X \rightrightarrows \overline{\mathbb{R}}$ , as well as linear operator  $K \in \mathbb{L}(U_p; U_d)$ , such that  $J, R, f_0, e, g$  are convex, proper, and lower semicontinuous in their first parameter. Additionally,  $J$  and  $e$  have Lipschitz gradients, the latter with respect to first parameter, with constants  $L$  and  $L(x)$  respectively. For each iteration, a way to construct the adjoint objective functions  $\mathcal{L}_k : U \times \mathbb{R}^n \rightarrow \mathbb{R}$ , and constraint sets  $\check{\mathcal{C}}_{d,k} \subset U_d \times \mathbb{R}^n$ . Moreover, an outer step length parameter  $\tau > 0$ , an adjoint step length parameter  $\theta$ , inner step length parameters  $\tau_p, \tau_d > 0$  satisfying  $\tau_p L(x^k)/2 + \tau_p \tau_d \|K\|^2 \leq 1$  and a parameter  $\omega > 0$ . We are also given iteration numbers  $N_{\text{inner}}, N_{\text{adjoint}} \in \mathbb{N}$  for the inner and adjoint loops respectively.

- 1: Pick initial iterates  $u^0 := (u_p^0, u_d^0) \in U, v^0 := (w_p^0, w_d^0, \lambda^0) \in U \times \mathbb{R}^n$ , and  $x^0 \in X$ .
- 2: **for**  $k \in \mathbb{N}$  **do**
- 3:    $u^{k,0} := u^k$
- 4:   **for**  $i = 0, \dots, N_{\text{inner}} - 1$  **do**  $\triangleright$  inner PDPS
- 5:      $u_p^{k,i+1} := \text{prox}_{\tau_p f_0(\cdot; x^k)}(u_p^{k,i} - \tau_p(K^* u_d^{k,i} + \nabla_{u_p} e(u_p^{k,i}; x^k)))$
- 6:      $u_d^{k,i+1} := \text{prox}_{\tau_d g^*(\cdot; x^k)}(u_d^{k,i} + \tau_d K((1 + \omega)u_p^{k,i+1} - \omega u_p^{k,i}))$
- 7:   **end for**
- 8:    $u^{k+1} := u^{k, N_{\text{inner}}}, v^{k,0} := v^k$
- 9:   **for**  $j = 0, \dots, N_{\text{adjoint}} - 1$  **do**  $\triangleright$  Forward-backward adjoint solver
- 10:      $v^{k,j+1} := \text{proj}_{\check{\mathcal{C}}_{d,k}}(v^{k,j} - \theta \nabla \mathcal{L}_k(v^{k,j}; u^{k+1}))$
- 11:   **end for**
- 12:    $v^{k+1} := v^{k, N_{\text{adjoint}}}$
- 13:    $\tilde{x}_{k+1}^* := [f^{(u_p, 2)}(u_p^{k+1}, x^k)]^* w_p^{k+1}$   $\triangleright$  Form the differential estimate
- 14:    $x^{k+1} := \text{prox}_{\tau R}(x^k - \tau \tilde{x}_{k+1}^*)$   $\triangleright$  Forward-backward outer step
- 15: **end for**

The rest of the adjoint problem (5.23) now reduces to

$$(5.24) \quad \min_{w_d \in U_d, \lambda \in \mathbb{R}^n} \frac{1}{2} \|x^{-1} K(K^* w_d + b - u_p) + \lambda \odot u_d + V w_d\|_{U_d}^2 \quad \text{subject to} \quad (w_d, \lambda) \in \mathcal{C}_d,$$

where  $V$  is defined in (5.16).

### 5.3 SUMMARY

We present in Algorithm 5.1 our overall proposed algorithm for bilevel problems (1.1) where  $G$  arises from the primal-dual optimality conditions of (5.3). In it, on each outer iteration

1. we work towards a solution of the inner problem, by taking (single or multiple) primal-dual proximal splitting (PDPS) steps,
2. work towards a root of the adjoint inclusion (5.7) (with  $\Delta w_*^{k+1} = -M(u^{k+1} - u^k)$  for  $M$  defined in (4.3)) by taking (single or multiple) forward-backward steps on (5.12) (when  $g$  is a non-parametric total variation regularisation term, (5.23)), and follow this by
3. taking forward-backward steps to update the outer variable.

Corresponding to the construction of  $\tilde{x}_{k+1}^* = x^*$  as an approximate solution of (5.5) (equivalently (5.7)) for  $\Delta w_* = -M(u^{k+1} - u^k)$  for  $M$  defined in (4.3) for a choice of coderivative  $\check{D}^* = D^*, \widehat{D}^*$ , we work with the target sets  $\mathcal{A}F(x^k) = \check{\mathcal{A}}_2 F(x^k)$  (see Remark 3.12) and the target  $x_{k+1}^*$ , solved as  $x^*$  from (5.7)

with  $\Delta w_* = 0$ . The convergence of the method, in the sense that  $\inf_{\bar{x}^* \in [\check{d}_2 F + \partial R](x^{k+1})} \|\bar{x}^*\| \rightarrow 0$ , thus follows from [Theorem 4.18](#) subject to the satisfaction of:

- (i) The tracking [Assumption 4.3](#) for the inner PDPS and the forward-backward adjoint solver.
- (ii) The pseudo-smoothness property (4.22) on the optimistic selection  $E = \text{Inf}_F$  or the pessimistic selection  $E = \text{Sup}_F$ , where  $F = J \circ S$  for the inner solution mapping  $S$ .
- (iii) The continuity condition (a) or (b) in the statement of the theorem.
- (iv) The local initialisation and level set boundedness condition (4.23).

These properties need to be analysed for each specific problem, and their complete verification remains outside the scope of the present already long work, where we were able to provide a suggestion of the feasibility of their satisfaction through the example of [Section 4.7](#). The continuity condition (b) can always be made to hold for Lipschitz-differentiable  $R$ , or, e.g. barrier functions (see [Remark 4.19](#)).

As a modification of [Algorithm 5.1](#), recalling [Remark 4.4](#), if both the inner problem and adjoint are solved exactly, both [Assumption 4.3](#) (i) and (ii) hold. We can also adaptively take  $N_{\text{inner}}$  and  $N_{\text{adjoint}}$  sufficiently large that these conditions hold for prescribed parameters.

Moreover,

1. [Assumption 4.3](#) (i) follows from [Theorems 4.7](#) and [4.9](#) subject to

- (I) The inner solution mapping satisfying the Lipschitz-type property (4.9), by, e.g., having the Aubin property or being Pompeii–Hausdorff–Lipschitz; see [Lemma 4.10](#), and the discussion preceding it.

It is not uncommon for the Aubin property to hold for primal-only formulations of the inner problem, see [[13](#), §28] and [Section 4.7](#). However, the primal-dual formulation can add additional difficulties and locality restrictions.

- (II) The mapping  $G(\cdot; x^k)$ , as defined in (5.4), being metrically subregular for all  $k \in \mathbb{N}$  with compatible neighbourhoods and moduli, as discussed in [Remark 4.8](#).

It is enough that each  $G(\cdot, x^k)$  is strongly monotone (outside a kernel subspace) with uniform factors.

2. [Assumption 4.3](#) (ii) follows from [Theorem 4.12](#) subject to:

- (III) The adjoint solution mapping satisfying the bi-Lipschitz-type property (4.10). This requires that exact adjoint solutions exist: that (4.18) holds.

[Lemma 4.13](#) provides one simplified tool if the inner solution mapping  $S$  is single-valued and Lipschitz. The general condition (4.10) makes no such requirement, as we saw in the example of [Section 4.7](#). However, a detailed study of the condition for the total variation regularised problems remains outside the scope of the present, already long, work.

[Remark 5.9](#). Nevertheless, we note that, presently, by [Lemma 5.1](#),

$$\begin{aligned} \check{Z}(u, x, w_*) &:= \{w \in W \mid 0 \in \check{D}^* G(u, x|w_*)(w) + (J'(u), -x^*), x^* \in X^*\} \\ &= \{w \in W \mid (5.7a) \text{ and } (5.7c) \text{ hold with } u^* = J'(u)\}. \end{aligned}$$

In the strictly complementary case, (i.e., only the first two cases of (5.21) or (5.21) occur), this can be reduced to a simple matrix system. If we replace  $g$  by its Moreau–Yosida regularisation (which is less than the second-order regularisation commonly required), i.e., add a quadratic to  $g^*$ , then it is not difficult to see that the system is non-singular provided  $\nabla_{u_p}^2 f(u_p, x)w_p$  and  $Kw_p = 0$  imply  $w_p = 0$ . If  $K$  is a discretised differential operator, this essentially means that  $f$  must be sensitive to each component of  $w_p$  changing by the same constant. Then  $\check{Z}$  is single-valued and Lipschitz.

(IV) The contractivity (4.11) of the adjoint update on Line 10 of Algorithm 5.1. This requires that algorithmic adjoint solutions exist: that (4.17) holds and hence that zero is reached in (5.23).

For the Fréchet coderivative,  $\widehat{\mathcal{C}}_d$  of (5.21) is convex. Since  $\mathcal{L} = \frac{1}{2}\|A \cdot -b\|^2$  for some  $A$  is linear-quadratic,  $P := \nabla \mathcal{L} + \partial \delta_{\mathcal{C}}$  is  $A^*A$ -strongly monotone in the sense of Lemma A.3. Thus, that lemma with  $M = \theta^{-1} \text{Id}$  verifies the metric subregularity of  $P$  at any of its minimisers with the global radius  $\delta = \infty$ . Now Theorem 4.6 and Remark 4.8 prove (4.11).

For the limiting coderivative,  $\mathcal{C}_d$  of (5.22) is nonconvex in the third case. However, it is only the union of two convex sets. We call them here branches. We can, as in the Fréchet case, obtain contractivity on the active branch, where  $\text{dist}(\check{Z}(u^{k+1}, x^k, \Delta w_*^{k+1}), w^k)$  is achieved. The left hand side of (4.11) can then only be improved by taking the minimum over the solutions on both branches, individually. Thus (4.11) still holds.

3. Assumption 4.3 (iii) follows from Theorem 4.14 subject to:

(V) The Lipschitz-type assumption (4.19).

Presently, from Lemma 5.1,

$$(5.25) \quad \check{T}(u, w, \Delta u^*, \Delta w_*; x) := \{x^* \in X^* \mid 0 \in \check{D}^*G(u, x | \Delta w_*)(w) + (J'(u) - \Delta u^*, -x^*)\} \\ = \{\nabla_{xu_p} f(u_p, x) w_p \mid (5.7a) \text{ and } (5.7c) \text{ hold with } u^* = J'(u)\}.$$

Therefore, (4.19) can be verified from the properties of  $f$ . This may require the variables  $u_p$ ,  $w_p$ , and  $x$  to be bounded. In Assumption 4.3, we already allow the restriction  $x \in \Omega_X$  and  $u \in \Omega_U$ , which are then ensured by the convergence Theorem 4.18 (via Lemma 4.16); see Remark 4.17. The variable  $w_p$  can be bounded through the adjoint optimisation problem (5.23) by comparing against  $(w_p, w_d, \lambda) = 0$ , recalling that  $u^* = J'(u^{k+1})$ .

(VI) The existence of exact and algorithmic adjoint solutions, i.e., again, (4.17) and (4.18).

This, again, depends on a careful analysis of the specific problem. In particular, under the special conditions of Remark 5.9 of (III), we can guarantee the existence of  $w = S_w(u^{k+1})$  or  $w = w^{k+1}$ . Then (5.25) immediately establishes (4.17) and (4.18).

Otherwise, the structural assumptions of Theorem 4.14 correspond to those of this section, as stated in Lemma 5.1.

The main challenge is, therefore, (4.22), adjoint existence (4.17) and (4.18), and the Lipschitz-type properties (4.9) and (4.10) of the inner and adjoint solution mappings. These need to be analysed on a case-by-case basis.

## 6 NUMERICAL RESULTS

We present two total variation regularised imaging problems and assess the performance of our Algorithm 5.1 on them. The example problems are:

**denoise** Find an optimal regularisation parameter to denoise a photographic image with simulated (Gaussian) noise.

**deblur** Find an optimal parametrisation of convolution kernel and regularisation parameter to deblur/deconvolve a photographic image with simulated blur and noise.

We call the evaluated algorithms:

**implicit** Algorithm 5.1 with  $N_{\text{inner}} = 500$  and  $N_{\text{adjoint}} = 2000$ . With near exact initialisation of the inner and adjoint variables this algorithm solves the inner and adjoint problems near exactly within each outer iterate.

**single-loop** Algorithm 5.1 with  $N_{\text{inner}} = 1$  and  $N_{\text{adjoint}} = 3$ , i.e. the version that allows inexact solution of the inner and adjoint problems, but it also uses the same near exact initialisation of the inner and adjoint variables.

The goal of these numerical results is to show that an implementation our algorithm works in practice, not to demonstrate the use in actual applications, where the size and the amount of the target images would be higher.

In Section 6.1, we present the inner, adjoint and outer objective functions of our numerical example bilevel problems. We present numerical details in Section 6.2, and discuss the results in Section 6.3.

**Remark 6.1 (Algorithms in the literature).** The “implicit” method roughly corresponds to the best applicable variant of first-order algorithms found in the literature. Aside from stochastic single-loop algorithms for smooth and strongly convex problems developed for machine learning applications (see [43] for an overview), these methods generally depend on exact or near-exact inner and adjoint solutions, by any available means. Apart from [46], the convergence theories, including that of SparseHO [4], generally require additional smoothing.

Also, the specific formulations treated in the literature are not always the same for the same application problem. Therefore, the algorithms are not necessarily directly comparable. In particular, our primal-dual formulation of the inner problem is rarely found in the literature. The closest practical match to our approach is given by [5] that has significant smoothness and strong convexity requirements.

The primal-dual transformation of the inner problem is, however, not critical in case of algorithms that require exact or near-exact inner and adjoint solutions. The allowed placement of the hyperparameters  $x$  is much more critical. In particular, [46] that has the least smoothness assumptions, only applies to inner problems of the form  $\min_u f(u) + \sum_i \psi(x_i)g_i(u)$ , where  $f$  is smooth, and  $g_i$  possibly nonsmooth. This excludes even our formulation (6.1) of the denoising problem, that places  $x$  in front of  $f$ . Our deblurring problem is not covered by the model of [46].

As for second-order methods, [46] also considers a nonsmooth trust region method based on the general locally Lipschitz approach of [40]. The specific approach of [46] is unapplicable for the same reasons as above. The authors also do not compare the forward-backward and trust region methods in terms of CPU time or total computational complexity, only in terms of iterations. The overall method of [40] might be applicable, if we were able to show the required Lipschitz assumptions. Judging from our earlier experience [16, 18], semismooth Newton’s methods have significant performance challenges on the type of problems considered here, even with appropriate smoothing.

## 6.1 OBJECTIVE FUNCTIONS

For both problems we take the outer fidelity term as  $J(u) = \frac{1}{2}\|u_p - b\|_2^2$  for  $b \in \mathbb{R}^{N^2}$  the “ground truth” image of dimensions  $N \times N$ . The latter is a cropped portion of image o7 or o2 – for denoising and deblurring respectively – of the free Kodak dataset [22] converted to grey values in  $[0, 1]$ ; see Figures 6.1a and 6.2b.

For the denoising experiment our inner problem reads as

$$(6.1) \quad \min_{u_p \in \mathbb{R}^{256^2}} \frac{x}{2} \|u_p - m\|^2 + C \|Ku_p\|_{2,1} \quad (x \in \mathbb{R}_+)$$

with simulated measurement  $m$ , see Figure 6.1b, obtained from  $b$  by adding Gaussian noise of standard deviation 0.08 and the matrix  $K$  is a (discrete) backward difference operator with Dirichlet boundary

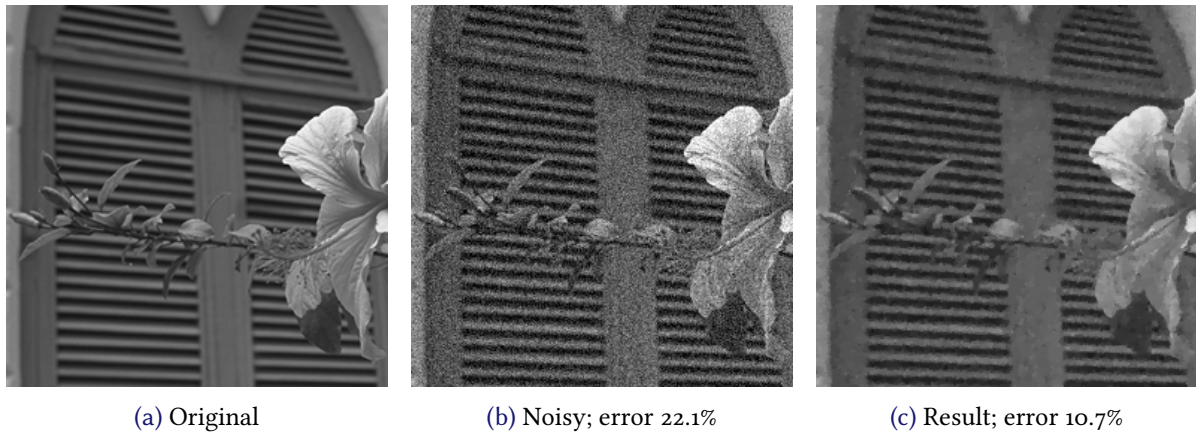


Figure 6.1: Denoising data and result for the "single-loop" method ( $N = 256$ ). The errors in (b) and (c) are  $\|\text{image} - \text{target}\|_2 / \|\text{target}\|_2$ .

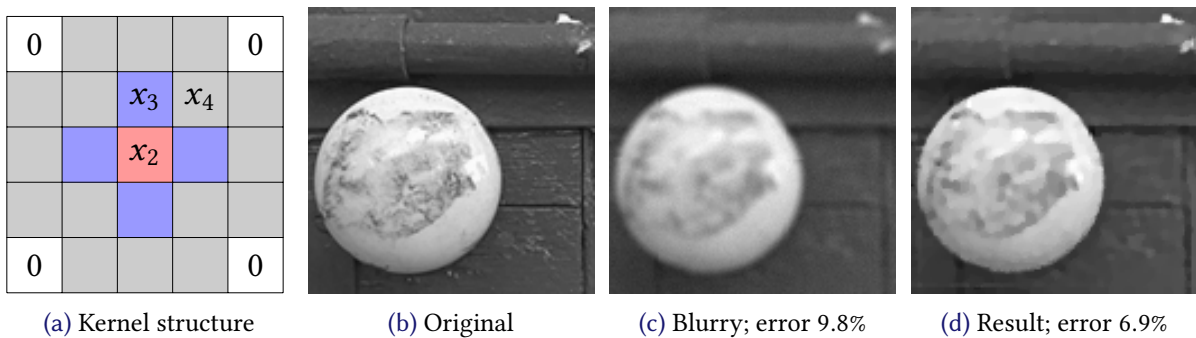


Figure 6.2: Deconvolution kernel parametrisation, data, and result for the "single-loop" method ( $N = 128$ ). The different colours in (a) represent different components of  $x$  such that elements of kernel with same colour have same value. The errors in (c) and (d) are  $\|\text{image} - \text{target}\|_2 / \|\text{target}\|_2$ .

conditions. We used the choice  $C = 0.1$ , and the outer regulariser  $R = 0$ . In Algorithm 5.1 we take  $f_0 = 0$ ,  $e(u_p, x) = \frac{x_1}{2} \|u_p - m\|^2$ , and  $g(u, x) = C \|Ku_p\|_{2,1}$ . The adjoint objective is given by (5.20) with  $u^* = \nabla J(u)$ , for the constraint set of (5.22) corresponding to the limiting coderivative (of  $G$ ).

For the deblurring experiment, we take as the forward operator  $A_x$  the convolution with  $5 \times 5$ -kernel parametrised by  $x$  as visualised in Figure 6.2a. With this and the  $K$  from above, the inner problem reads as

$$(6.2) \quad \min_{u_p \in \mathbb{R}^{128^2}} \frac{x_1}{2} \|A_x u_p - m\|^2 + C \|Ku_p\|_{2,1} \quad (x \in \mathbb{R}_+^4).$$

Write  $r_\varphi$  be the operator for rotating the image  $\varphi$  degrees, clockwise for  $\varphi > 0$  and counterclockwise for  $\varphi < 0$ . We then use in (6.2) the simulated measurement  $m = r_{-1}(A_x r_1(b)) + \varepsilon$ , see Figure 6.2c for the true kernel weights  $x_2 = 0.15$ ,  $x_3 = 0.1$  and  $x_4 = 0.75$ , as well as Gaussian noise  $\varepsilon$  from distribution with standard deviation 0.01. The aim of rotations in simulating the measurement is to provide additional modelling error to avoid inverse crime. We take the constant  $C = 0.1$  and the outer regulariser  $R(x) = \beta(x_2 + x_3 + x_4 - 1)^2$  with  $\beta = 10^4$ . The functions for the adjoint objective (5.20) are analogous to the denoising experiment with the difference of  $f(u_p, x) = f_0(u_p, x) + e(u_p, x) = e(u_p, x) = \frac{x_1}{2} \|A_x u_p - m\|^2$ .

## 6.2 NUMERICAL SETUP

Our algorithm implementation (code files) will be available on Zenodo for the publication of this article. The choices for adjoint and outer step lengths for both experiments are listed in Table 6.1. We picked inner step lengths  $\tau_p, \tau_d > 0$  satisfying  $\tau_p L(x)/2 + \tau_p \tau_d \|K\|^2 \leq 1$  for PDPS with the upper bound of  $\|K\|^2 \leq 8$  from [9]. The Lipschitz constant  $L(x)$  depends on the outer variable (iterates) since the first terms of the inner objectives from (6.1) and (6.2) also has dependency on it. All the other step lengths, numbers of steps for solving inner problem and adjoint system to a high precision in implicit method, and parameters, such as  $\omega = 1$ , are chosen by trial and error to obtain an apparently stable, but as efficient as possible, algorithm. We do not attempt to verify their theoretical conditions.

We choose the adjoint system that corresponds to limiting coderivative (of  $G$ ) and essentially the constraint set  $\mathcal{C}_{d,k}$  given by (5.22). We also tried the system corresponding to Fréchet coderivative, but observed no significant difference. We took the former since we considered the potentially empty outer objective differential more serious issue than the potential issues with the convergence of the adjoint algorithm from nonconvexity of the constraint set.

**Remark 6.2.** To compute the constraint set  $\mathcal{C}_{d,k}$  given by (5.22), we require  $\Delta w_d^{k+1} + K u_p^{k+1}$ . We can avoid calculating this explicitly: We have  $[\Delta w_d^{k+1} + K u_p^{k+1}]_j \in \partial \delta_{B_{\mathbb{R}^2}(0,C)}(u_{d,j}^{k+1})$ , which implies that  $[\Delta w_d^{k+1} + K u_p^{k+1}]_j = v_j u_{d,j}^{k+1}$  for some  $v_j \geq 0$ . From the treatment of the PDPS as inner method Example 4.2, we have  $(\Delta w_p^{k+1}, \Delta w_d^{k+1}) = \Delta w_*^{k+1} = -M(u^{k+1} - u^k)$ . Thus

$$(u_d^k + \tau K(2u_p^{k+1} - u_p^k))_j = (1 + \tau v_j) u_{d,j}^{k+1}.$$

On the other hand, the dual update of the inner PDPS yields

$$u_{d,j}^{k+1} = \text{proj}_{B(0,1)}([u_d^k + \tau K(2u_p^{k+1} - u_p^k)]_j)$$

These two equations allow us to solve for  $v_j$  and then  $[\Delta w_d^{k+1} + K u_p^{k+1}]_j = v_j u_{d,j}^{k+1}$ . This is practical because we have to compute  $\|[u_d^k + \tau K(2u_p^{k+1} - u_p^k)]_j\|$  within the inner step to perform the projection onto  $B(0,1)$ .

The initial iterates for the outer variable were taken as  $x^{(0)} = 4$  for the denoise experiment and  $x^{(0)} = (10, 0.25, 0.25, 0.5)$  for the deblurring experiment. The initial iterates for the inner and adjoint variables were obtained by running one iteration of outer loop of Algorithm 5.1 with zero initialisation,  $N_{\text{inner}} = 10000$  and  $N_{\text{adjoint}} = 50000$ .

To compare algorithm performance, we plot norms of differential estimates and values of outer objective as function of the CPU time value of Matlab on an AMD Ryzen 5 5600H CPU. We call this value “computational resources”, since it measures the use of several CPU cores by Matlab’s internal linear algebra. Thus CPU time is more accurate indicator of the actual resources than the elapsed real time.

## 6.3 RESULTS

The numerical evidence presented in Figure 6.4 suggest that both methods converge in terms of both function values, and the infimal norm of the coderivative. This numerical convergence is observed even if we can’t confirm all of the assumptions for the convergence Theorem 4.18. Especially for the single-loop method the validation of the assumptions would require additional work. The convergence seems to be even linear, at least in the infimal coderivative norm. This is reasonable for denoising since by Figure 6.3 the outer objective resembles a locally strongly convex function around the minimum.

With limiting coderivatives, the possibly nonconvex constraint set  $\mathcal{C}_{d,k}$  could in principle be problematic in the adjoint algorithm. However, in practise, we did not observe any problems: the convergence

**Table 6.1:** Algorithm parametrisation, time multiplier, and total outer steps to reach threshold computational resources (20000 CPU time) value. The time multipliers allow conversion from computational resources to seconds. It varies between algorithms and problems due to different levels of parallelisability. The ‘inner steps’ and ‘adjoint steps’ refer to  $N_{\text{inner}}$  and  $N_{\text{adjoint}}$  in Algorithm 5.1, which indicate the number of iterations taken towards a solution of the inner problem or the adjoint inclusion on every outer iteration. Step lengths for PDPS (inner steps) varied for iterations, because the relevant Lipschitz variable depends on outer iterate.

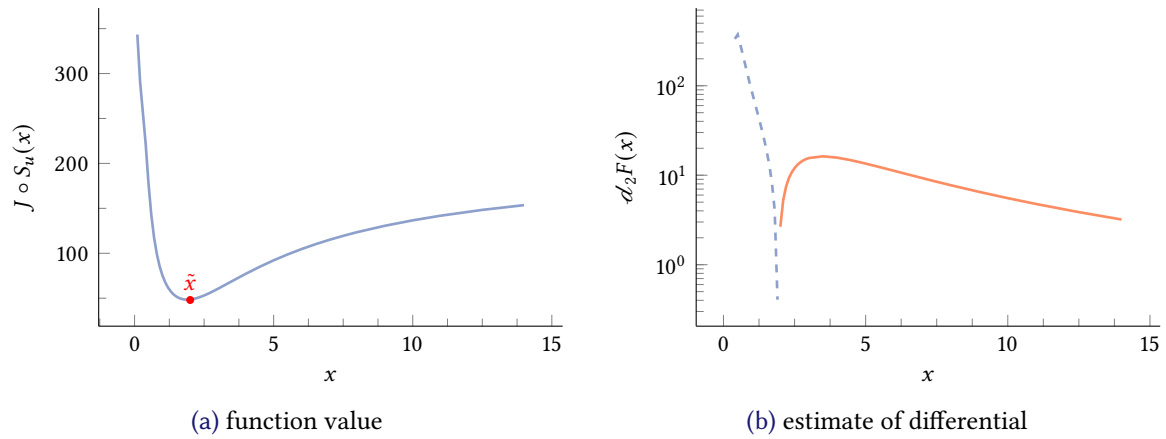
	method	outer steps	inner steps	adjoint steps	time mult.	$\theta$	$\tau$
denoise	implicit	$7.3 \cdot 10^2$	500	$2 \cdot 10^3$	0.31	$5 \cdot 10^{-3}$	$3 \cdot 10^{-4}$
	single-loop	$4.4 \cdot 10^5$	1	3	0.29	$5 \cdot 10^{-3}$	$1 \cdot 10^{-6}$
deblur	implicit	$1.7 \cdot 10^3$	500	$2 \cdot 10^3$	0.23	$5 \cdot 10^{-4}$	$1 \cdot 10^{-2}$
	single-loop	$7.8 \cdot 10^5$	1	3	0.23	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$

behaviour of adjoint algorithms with limiting coderivatives was not significantly different from the Fréchet coderivative, for which the set  $\mathcal{E}_{d,k}$  is convex.

To conclude, our numerical experiments confirm that our proposed single-loop method converges, and that the convergence is faster than that of the (near-exact) implicit method. The theoretical gaps outlined in Section 5.3 are a rich source of research questions for the future.

## REFERENCES

- [1] F.J. A. Artacho and M.H. Geoffroy, Characterization of metric regularity of subdifferentials, *Journal of Convex Analysis* 15 (2008), 365–380.
- [2] M. Benko, On inner calmness\*, generalized calculus, and derivatives of the normal cone mapping, *Journal of Nonsmooth Analysis and Optimization* 2 (2021).
- [3] M. Benko, H. Gfrerer, and J. V. Outrata, Calculus for directional limiting normal cones and subdifferentials, *Set-Valued and Variational Analysis* 27 (2019), 713–745.
- [4] Q. Bertrand, Q. Klopfenstein, M. Massias, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon, Implicit Differentiation for Fast Hyperparameter Selection in Non-Smooth Convex Learning, *Journal of Machine Learning Research* 23 (2022), 1–43, <http://jmlr.org/papers/v23/21-0486.html>.
- [5] L. Bogensperger, M.J. Ehrhardt, T. Pock, M. S. Salehi, and H. S. Wong, An Adaptively Inexact Method for Bilevel Learning Using Primal–Dual-Style Differentiation, *Journal of Mathematical Imaging and Vision* 67 (2025), [doi:10.1007/s10851-025-01262-w](https://doi.org/10.1007/s10851-025-01262-w).
- [6] J. Bolte, T. Le, E. Pauwels, and T. Silveti-Falls, Nonsmooth Implicit Differentiation for Machine-Learning and Optimization, in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan (eds.), volume 34, Curran Associates, Inc., 2021, 13537–13549, [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/70afbf2259b4449d8ae1429e054df1b1-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/70afbf2259b4449d8ae1429e054df1b1-Paper.pdf).
- [7] J. Bolte, E. Pauwels, and A. Silveti-Falls, Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems, *SIAM Journal on Optimization* 34 (2024), 71–97, [doi:10.1137/22m1541630](https://doi.org/10.1137/22m1541630).



**Figure 6.3:** Graph of composed objective for the denoising problem along with near-optimal  $\tilde{x}$  found by recursive subdivision. Subfigure (b) presents the estimate of the differential target set  $\mathcal{d}_2F$  defined in (3.14). The estimate is obtained as  $\tilde{x}_1^*$  given by Algorithm 5.1 with  $N_{\text{inner}} = 10000$ ,  $N_{\text{adjoint}} = 50000$ , and zero initialisation of iterates. The values for composed objective function in subfigure (a) are obtained using the same inner solution estimate than for the differential target estimate. In (b), the dashed blue line indicates negative values, and the red continuous line, positive values. This numerical evidence also confirms that the minimum of  $J \circ S_u$  is characterised by a zero of  $\mathcal{d}_2F$ .

- [8] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer New York, 2000, [doi:10.1007/978-1-4612-1394-9](https://doi.org/10.1007/978-1-4612-1394-9).
- [9] A. Chambolle, An algorithm for total variation minimization and applications, *Journal of Mathematical Imaging and Vision* 20 (2004), 89–97, [doi:10.1023/b:jmiv.0000011325.36760.1e](https://doi.org/10.1023/b:jmiv.0000011325.36760.1e).
- [10] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision* 40 (2011), 120–145, [doi:10.1007/s10851-010-0251-1](https://doi.org/10.1007/s10851-010-0251-1).
- [11] T. Chen, Y. Sun, and W. Yin, A single-timescale stochastic bilevel optimization method, 2021, [arXiv:2102.04671](https://arxiv.org/abs/2102.04671).
- [12] C. Christof, J. C. De Los Reyes, and C. Meyer, A non-smooth trust-region method for locally Lipschitz functions with application to optimization problems constrained by variational inequalities, *SIAM Journal on Optimization* 30 (2020), 2163–2196, [doi:10.1137/18m1164925](https://doi.org/10.1137/18m1164925), [arXiv:1711.03208](https://arxiv.org/abs/1711.03208).
- [13] C. Clason and T. Valkonen, *Introduction to Nonsmooth Analysis and Optimization*, MOS-SIAM Series on Optimization, SIAM, 2026, [doi:10.1137/1.9781611978995](https://doi.org/10.1137/1.9781611978995).
- [14] M. Dagr eou, P. Ablin, S. Vaiter, and T. Moreau, A framework for bilevel optimization that enables stochastic and global variance reduction algorithms, 2022, [arXiv:2201.13409](https://arxiv.org/abs/2201.13409).
- [15] J. C. De Los Reyes, *Numerical PDE-Constrained Optimization*, SpringerBriefs in Optimization, Springer International Publishing, 2015, [doi:10.1007/978-3-319-13395-9](https://doi.org/10.1007/978-3-319-13395-9).
- [16] J. C. De Los Reyes, C. B. Sch onlieb, and T. Valkonen, Bilevel parameter learning for higher-order total variation regularisation models, *Journal of Mathematical Imaging and Vision* 57 (2017), 1–25, [doi:10.1007/s10851-016-0662-8](https://doi.org/10.1007/s10851-016-0662-8), [arXiv:1508.07243](https://arxiv.org/abs/1508.07243).

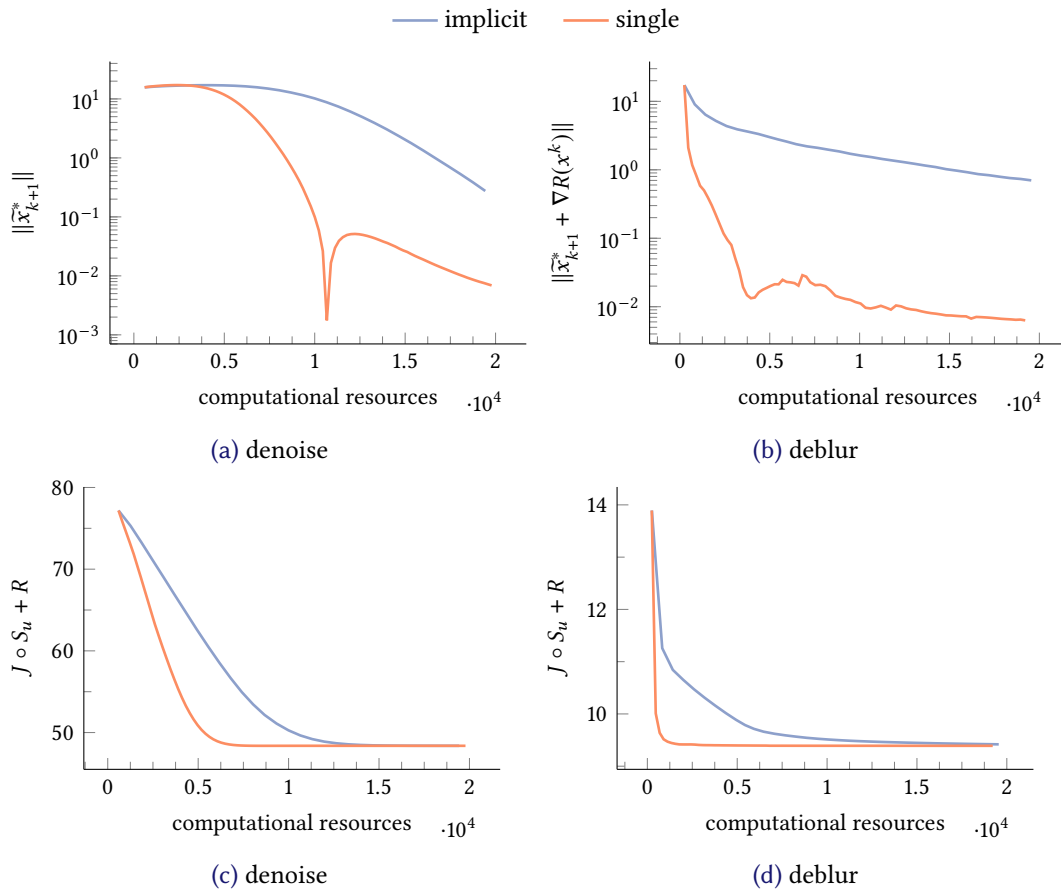


Figure 6.4: Performance of the compared methods. Top row: (norm of) differential estimate convergence, for denoising  $R = 0$ , and bottom row: outer function value convergence. The “computational resources” is the spent CPU time over multiple cores, parallelisation level depending on algorithm.

- [17] S. Dempe and P. Mehlitz, Duality-Based Single-Level Reformulations of Bilevel Optimization Problems, *Journal of Optimization Theory and Applications* 205 (2025), doi:10.1007/s10957-025-02627-2.
- [18] N. Dizon and T. Valkonen, Differential estimates for fast first-order multilevel nonconvex optimisation, 2024, arXiv:2412.01481v3. submitted.
- [19] A. L. Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings: A View from Variational Analysis*, Springer Series in Operations Research and Financial Engineering, Springer, 2014, doi:10.1007/978-1-4939-1037-3.
- [20] M. J. Ehrhardt and L. Roberts, Inexact derivative-free optimization for bilevel learning, *Journal of mathematical imaging and vision* 63 (2021), 580–600, doi:10.1007/s10851-021-01020-8.
- [21] M. C. Ferris, O. L. Mangasarian, and J. S. Pang, *Complementarity: Applications, Algorithms and Extensions*, Springer US, 2001, doi:10.1007/978-1-4757-3279-5.
- [22] R. Franzen, Kodak lossless true color image suite, PhotoCD PCD0992. Lossless, true color images released by the Eastman Kodak Company, 1999, <http://rok.us/graphics/kodak/>.
- [23] S. Ghadimi and M. Wang, Approximation methods for bilevel programming, 2018, arXiv:1802.02246.

- [24] N. T. V. Hang, W. Jung, and E. Sarabi, Role of Subgradients in Variational Analysis of Polyhedral Functions, *Journal of Optimization Theory and Applications* 200 (2024), 1160–1192, doi:10.1007/s10957-024-02378-6.
- [25] N. T. V. Hang and M. E. Sarabi, A Chain Rule for Strict Twice Epi-Differentiability and Its Applications, *SIAM Journal on Optimization* 34 (2024), 918–945, doi:10.1137/22m1520025.
- [26] B. He and X. Yuan, Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective, *SIAM Journal on Imaging Sciences* 5 (2012), 119–149, doi:10.1137/100814494.
- [27] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, number 23 in Mathematical Modelling: Theory and Applications, Springer Netherlands, 2009, doi:10.1007/978-1-4020-8839-1.
- [28] M. Hong, H. T. Wai, Z. Wang, and Z. Yang, A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic, 2020, arXiv:2007.05170.
- [29] A. D. Ioffe, *Variational Analysis of Regular Mappings: Theory and Applications*, Springer Monographs in Mathematics, Springer, 2017, doi:10.1007/978-3-319-64277-2.
- [30] J. Jauhainen, Y. Nabou, and T. Valkonen, Dynamic inverse problems: Single-loop online algorithms, 2026, arXiv:2606.10951. submitted.
- [31] K. Ji and Y. Liang, Lower bounds and accelerated algorithms for bilevel optimization, *Journal of Machine Learning Research* 23 (2022), 1–56.
- [32] K. Ji, J. Yang, and Y. Liang, Bilevel optimization: Convergence analysis and enhanced design, in *International Conference on Machine Learning*, 2021, 4882–4892.
- [33] L. O. Jolaoso, P. Mehlitz, and A. B. Zemkoho, A fresh look at nonsmooth Levenberg–Marquardt methods with applications to bilevel optimization, *Optimization* 74 (2024), 2745–2792, doi:10.1080/02331934.2024.2313688.
- [34] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak, A fully first-order method for stochastic bilevel optimization, in *International Conference on Machine Learning*, 2023, 18083–18113.
- [35] J. Li, B. Gu, and H. Huang, A fully single loop algorithm for bilevel optimization without hessian inverse, in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022, 7426–7434.
- [36] Z. Q. Luo, J. S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, 1996, doi:10.1017/cb09780511983658.
- [37] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330 of Grundlehren der mathematischen Wissenschaften, Springer, 2006, doi:10.1007/3-540-31247-1.
- [38] B. S. Mordukhovich, N. M. Nam, and N. D. Yen, Fréchet subdifferential calculus and optimality conditions in nondifferentiable programming, *Optimization* 55 (2006), 685–708, doi:10.1080/02331930600816395.
- [39] J. V. Outrata, On the numerical solution of a class of Stackelberg problems, *ZOR Zeitschrift für Operations Research Methods and Models of Operations Research* 34 (1990), 255–277, doi:10.1007/bf01416737.

- [40] L. Qi and J. Sun, A trust region algorithm for minimization of locally Lipschitzian functions, *Mathematical Programming* 66 (1994), 25–43, [doi:10.1007/bf01581136](https://doi.org/10.1007/bf01581136).
- [41] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer, 1998, [doi:10.1007/978-3-642-02431-3](https://doi.org/10.1007/978-3-642-02431-3).
- [42] M. S. Salehi, S. Mukherjee, L. Roberts, , and M. J. Ehrhardt, An adaptively inexact first-order method for bilevel optimization with application to hyperparameter learning, 2024, [arXiv:2308.10098](https://arxiv.org/abs/2308.10098).
- [43] E. Suonperä and T. Valkonen, Linearly convergent bilevel optimization with single-step inner methods, *Computational Optimization and Applications* (2023), [doi:10.1007/s10589-023-00527-7](https://doi.org/10.1007/s10589-023-00527-7), [arXiv:2205.04862](https://arxiv.org/abs/2205.04862).
- [44] E. Suonperä and T. Valkonen, Single-loop methods for bilevel parameter learning in inverse imaging, *Journal of Nonsmooth Analysis and Optimization* 5 (2025), [doi:10.46298/jnsao-2025-15577](https://doi.org/10.46298/jnsao-2025-15577), [arXiv:2408.08123](https://arxiv.org/abs/2408.08123).
- [45] T. Valkonen, Preconditioned proximal point methods and notions of partial subregularity, *Journal of Convex Analysis* 28 (2021), 251–278, [arXiv:1711.05123](https://arxiv.org/abs/1711.05123).
- [46] E. Vilches, D. Villacís, and P. Pérez-Aros, A Variational Analysis Approach for Bilevel Hyperparameter Optimization with Sparse Regularization, 2025, <https://optimization-online.org/?p=30877>.
- [47] G. Wachsmuth and D. Walter, Proximal gradient methods in Banach spaces, 2025, [arXiv:2509.24685](https://arxiv.org/abs/2509.24685).
- [48] J. Yang, K. Ji, and Y. Liang, Provably faster algorithms for bilevel optimization, *Advances in Neural Information Processing Systems* 34 (2021), 13670–13682.

## APPENDIX A CONTRACTIVITY UNDER METRIC SUBREGULARITY

With the goal of verifying the inner tracking property for standard inner algorithms, subject to metric subregularity, we prove here the linear convergence of forward-backward type methods to the entire set of minimisers of a convex problem. Our approach includes primal-dual methods, and relaxes the standard assumption of strong convexity. Similar results first appeared in [45], however, focussing on basic primal-dual methods, they did not include forward steps. Our proof here adapts those in [13, Theorem 29.8 and Lemmas 29.6 and 29.7] to (a) treat primal-dual methods, (b) include the residual  $\|u^{k+1} - u^k\|^2$  in the claim, and (c) to remove the squares.

We consider here, in a Hilbert space  $U$ , general forward-backward type methods that solve optimality conditions of the form

$$0 \in T(u^{k+1}) := \partial g(u^{k+1}) + \nabla f(u^{k+1}) + \Xi u^{k+1},$$

and algorithms of the form

$$(A.1) \quad 0 \in \partial g(u^{k+1}) + \nabla f(u^k) + \Xi u^{k+1} + M(u^{k+1} - u^k),$$

where  $\Xi \in \mathbb{L}(U; U)$  is skew-adjoint;  $M \in \mathbb{L}(U; U)$  is self-adjoint and positive definite;  $g : U \rightarrow \overline{\mathbb{R}}$  is convex, proper, and lower semicontinuous; and  $f : U \rightarrow \overline{\mathbb{R}}$  is convex and Fréchet differentiable with an  $L_M$ -Lipschitz gradient with respect to the  $M$  and  $M^{-1}$ -norms:  $\|\nabla f(u) - \nabla f(\tilde{u})\|_{M^{-1}} \leq L_M \|u - \tilde{u}\|_M$ . For the basic forward-backward method,  $M = \tau^{-1} \text{Id}$ . For the PDPS, see [Example 4.2](#).

**Lemma A.1.** *For  $u^{k+1}$  generated by (A.1) from  $u^k$ , we have  $2(1 + L_M^2) \|u^{k+1} - u^k\|_M^2 \geq \text{dist}_{M^{-1}}^2(0, T(u^{k+1}))$ .*

*Proof.* Using (A.1), we have

$$(A.2) \quad \frac{1}{2} \|u^{k+1} - u^k\|_M^2 = \frac{1}{2} \text{dist}_{M^{-1}}^2(0, \{-M(u^{k+1} - u^k)\}) \geq \frac{1}{2} \text{dist}_{M^{-1}}^2(0, \partial g(u^{k+1}) + \nabla f(u^k) + \Xi u^{k+1}).$$

Young's inequality for any  $\alpha \in (0, 1)$  then yields

$$\begin{aligned} & \frac{1}{2} \text{dist}_{M^{-1}}^2(0, \partial g(u^{k+1}) + \nabla f(u^k) + \Xi u^{k+1}) \\ &= \frac{1}{2} \text{dist}_{M^{-1}}^2(\nabla f(u^{k+1}) - \nabla f(u^k), \partial g(u^{k+1}) + \nabla f(u^{k+1}) + \Xi u^{k+1}) \\ &= \inf_{q \in \partial g(u^{k+1})} \frac{1}{2} \|(\nabla f(u^{k+1}) - \nabla f(u^k)) - (q + \nabla f(u^{k+1}) + \Xi u^{k+1})\|_{M^{-1}}^2 \\ &\geq \frac{1 - \alpha^{-1}}{2} \|\nabla f(u^{k+1}) - \nabla f(u^k)\|_{M^{-1}}^2 + \inf_{q \in \partial g(u^{k+1})} \frac{1 - \alpha}{2} \|q + \nabla f(u^{k+1}) + \Xi u^{k+1}\|_{M^{-1}}^2 \\ &\geq \frac{(1 - \alpha^{-1})L_M^2}{2} \|u^{k+1} - u^k\|_M^2 + \frac{1 - \alpha}{2} \text{dist}_{M^{-1}}^2(0, T(u^{k+1})), \end{aligned}$$

where we have used in the last step that  $\alpha \in (0, 1)$  and that  $\nabla f$  is Lipschitz continuous between the  $M^{-1}$  and  $M$ -norms. Combining this estimate with (A.2) and taking  $\alpha = 1/2$ , we obtain that

$$\frac{1 + L_M^2}{2} \|u^{k+1} - u^k\|_M^2 \geq \frac{1}{4} \text{dist}_{M^{-1}}^2(0, T(u^{k+1})). \quad \square$$

**Theorem A.2.** *With the above setup, suppose  $0 < \varepsilon \leq 1 - L_M/2$ , and that  $T$  is metrically subregular with respect to the  $M$  and  $M^{-1}$ -norms at  $\hat{u}$  for  $\hat{w} = 0$  with the modulus  $\kappa_M > 0$  and radius  $\delta > 0$ , i.e.,*

$$(A.3) \quad \text{dist}_M(u, T^{-1}(0)) \leq \kappa_M \text{dist}_{M^{-1}}(0, T(u)) \quad (u \in B_M(\hat{u}, \delta)).$$

Take  $u^k \in B_M(\hat{u}, \delta)$  for a  $\hat{u} \in T^{-1}(0)$ , and generate  $u^{k+1}$  through the satisfaction of (A.1). Then  $u^{k+1} \in B_M(\hat{u}, \delta)$ , and, for  $\rho = (1 - L_M/2 - \varepsilon)/(2(1 + L_M^2)\kappa_M^2) > 0$  and any  $\alpha > 0$ , we have

$$(A.4) \quad 2\sqrt{\frac{1 + \rho}{2 + \alpha}} \text{dist}_M(u^{k+1}, T^{-1}(0)) + 2\sqrt{\frac{\varepsilon}{2 + \alpha^{-1}}} \|u^{k+1} - u^k\|_M \leq \text{dist}_M(u^k, T^{-1}(0)).$$

*Proof. Step 1:* We first show that  $u^{k+1} \in B_M(\hat{u}, \delta)$  using standard F ejer monotonicity. In fact, take any element  $\bar{u} \in T^{-1}(0)$ . Then for some  $\bar{w} \in \partial g(\bar{u})$  we have  $0 = \bar{w} + \nabla f(\bar{u}) + \Xi \bar{u}$ . Also let  $w^{k+1} \in \partial g(u^{k+1})$  be the element that satisfies (A.1) holds. By [13, Theorem 7.2], equipping  $U$  with the norm  $\|\cdot\|_M$  and  $U^*$  with norm  $\|R \cdot\|_{M^{-1}}$ , where  $R : U^* \rightarrow U$  is the Riesz map (for the original inner product of  $U$ , unrelated to  $M$ ), the Lipschitz continuity of  $\nabla f$  implies the three-point monotonicity

$$\langle \nabla f(u^k) - \nabla f(\bar{u}), u^{k+1} - \bar{u} \rangle_U = \langle f'(u^k) - f'(\bar{u}), u^{k+1} - \bar{u} \rangle \geq -\frac{L_M}{4} \|u^{k+1} - u^k\|_M^2.$$

The first equality here holds because the special choice of norms does not affect the duality pairing, hence not the Riesz representation using the original inner product of  $U$ . Combining this with the monotonicity of  $\partial g$ , and the fact that by skew-symmetry  $\langle \Xi u, u \rangle = 0$  for all  $u$ , we obtain

$$(A.5) \quad \langle w^{k+1} + \nabla f(u^k) + \Xi u^{k+1}, u^{k+1} - \bar{u} \rangle = \langle w^{k+1} - \bar{w} + \nabla f(u^k) - \nabla f(\bar{u}), u^{k+1} - \bar{u} \rangle \geq -\frac{L_M}{4} \|u^{k+1} - u^k\|_M^2.$$

Applying (A.1), this gives  $-\langle M(u^{k+1} - u^k), u^{k+1} - \bar{u} \rangle \geq -\frac{L_M}{4} \|u^{k+1} - u^k\|_M^2$ . Pythagoras' identity then establishes

$$(A.6) \quad \frac{1}{2} \|u^k - \bar{u}\|_M^2 \geq \frac{1 - L_M/2}{2} \|u^{k+1} - u^k\|_M^2 + \frac{1}{2} \|u^{k+1} - \bar{u}\|_M^2.$$

Since  $L_M \leq 2$  by assumption, we obtain the Féjer monotonicity  $\|u^{k+1} - \bar{u}\|_M \leq \|u^k - \bar{u}\|_M$ . Taking  $\bar{u} = \hat{u}$ , our assumption  $u^k \in B_M(\hat{u}, \delta)$ , thus, implies  $u^{k+1} \in B_M(\hat{u}, \delta)$ .

**Step 2:** We then prove that

$$(A.7) \quad \frac{1}{2} \text{dist}_M^2(u^k, T^{-1}(0)) \geq \frac{1+\rho}{2} \text{dist}_M^2(u^{k+1}, T^{-1}(0)) + \frac{\varepsilon}{2} \|u^{k+1} - u^k\|_M^2.$$

Indeed, combining Lemma A.1 and the definition of metric subregularity yields the *error bound*

$$2(1 + L_M^2) \|u^{k+1} - u^k\|_M^2 \geq \text{dist}_{M^{-1}}^2(0, T(u^{k+1})) \geq \kappa_M^{-2} \text{dist}_M^2(u^{k+1}, T^{-1}(0)).$$

It follows for all  $\bar{u} \in T^{-1}(0)$  that

$$(A.8) \quad \frac{1 - L_M/2 - \varepsilon}{2} \|u^{k+1} - u^k\|_M^2 + \frac{1}{2} \|u^{k+1} - \bar{u}\|_M^2 \geq \frac{\rho}{2} \text{dist}_M^2(u^{k+1}; T^{-1}(0)) + \frac{1}{2} \|u^{k+1} - \bar{u}\|_M^2 \\ \geq \frac{1+\rho}{2} \text{dist}_M^2(u^{k+1}, T^{-1}(0)).$$

Summing (A.6) and (A.8) yields

$$\frac{1}{2} \|u^k - \bar{u}\|_M^2 \geq \frac{\varepsilon}{2} \|u^{k+1} - u^k\|_M^2 + \frac{1+\rho}{2} \text{dist}_M^2(u^{k+1}, T^{-1}(0)).$$

Taking the infimum over  $\bar{u} \in T^{-1}(0)$ , we obtain (A.7)

**Step 3:** We now remove the squares from (A.7). By Young's inequality, for any  $\alpha > 0$  and  $a, b \in \mathbb{R}$ , we have

$$\frac{1+\rho}{2} a^2 + \frac{\varepsilon}{2} b^2 \geq \left( \sqrt{\frac{1+\rho}{2+\alpha}} a + \sqrt{\frac{\varepsilon}{2+\alpha^{-1}}} b \right)^2.$$

Taking  $a = \text{dist}_M^2(u^{k+1}, T^{-1}(0))$  and  $b = \|u^{k+1} - u^k\|_M$ , and combining with (A.7) finishes the proof.  $\square$

The following lemmas verify the required metric subregularity subject to strong monotonicity on subspaces; see also [45]. We recall that we assume  $M$  to be self-adjoint and positive definite.

**Lemma A.3.** *Suppose  $T$  is  $\Gamma$ -strongly monotone for a self-adjoint  $\Gamma \in \mathbb{L}(U; U)$  at a  $\hat{u} \in T^{-1}(0)$  in the sense that, for some  $\gamma > 0$ ,*

$$(A.9) \quad \|u - \hat{u}\|_\Gamma^2 \leq \langle u - \hat{u}, q - \hat{q} \rangle \quad \text{for all } u \in U, q \in T(u), \hat{q} \in T(\hat{u}).$$

*Suppose, moreover, that  $V := T^{-1}(0) - \hat{u}$  is a closed subspace of  $U$ , and, that we have  $\gamma M \leq \Gamma$  on  $V_M^\perp$  (i.e.,  $\langle v^\perp, [\Gamma - \gamma M]v^\perp \rangle \geq 0$  for all  $v^\perp \in V_M^\perp$ ) for some  $\gamma > 0$  and  $V_M^\perp := \{v^\perp \in U \mid \langle Mv, v^\perp \rangle = 0 \text{ for all } v \in V\}$ . Then the metric subregularity (A.3) holds with  $\delta = \infty$  and  $\kappa_M = \gamma^{-1}$ .*

*Proof.* We can decompose<sup>5</sup> any  $u \in U$  as  $u = u_V + u^\perp \in V + V_M^\perp$ . Hence, for any  $q \in T(u)$ ,

$$\gamma \text{dist}_M(u, T^{-1}(0))^2 = \inf_{\bar{u} \in T^{-1}(0)} \gamma \|u - \bar{u}\|_M^2 = \inf_{v \in V} \gamma \|v + \hat{u} - \bar{u}\|_M^2 = \gamma \|u^\perp - \hat{u}^\perp\|_M^2 \leq \gamma \|u^\perp - \hat{u}^\perp\|_\Gamma^2 \\ = \gamma \|u - \hat{u}\|_\Gamma^2 \leq \langle u - \hat{u}, q - 0 \rangle \leq \|u - \hat{u}\|_M \|q\|_{M^{-1}} = \text{dist}_M(u, T^{-1}(0)) \|q\|_{M^{-1}}.$$

Taking the infimum over  $q \in T(u)$ , and dividing by  $\text{dist}_M(u, T^{-1}(0))$ , we obtain (A.3) with  $\kappa_u = \gamma^{-1}$ .  $\square$

**Lemma A.4.** *Suppose  $T$  is  $\Gamma = \gamma M$ -strongly monotone in the sense (A.9). Then the metric subregularity (A.3) holds with  $\delta = \infty$  and  $\kappa_M = \gamma^{-1}$ .*

<sup>5</sup>Solving the problem  $\min_{v \in V} \|u - v\|_M^2$  gives  $u = v + M^{-1}z$  for some  $z \in N_V(v)$ . Since  $V$  is a closed subspace,  $M^{-1}z \in V_M^\perp$ .

*Proof.* We apply Lemma A.3, where now  $V = \{0\}$ , because (A.9) implies  $T^{-1}(0) = \widehat{u}$  through the positive definiteness of  $M$ .  $\square$

**Theorem 4.6 (Contractivity of forward-backward splitting under metric subregularity).** *For convex functions  $g, f : U \rightarrow \overline{\mathbb{R}}$  on a Hilbert space  $U$ , where  $f$  has an  $L$ -Lipschitz gradient, suppose  $T := \partial g + \nabla f$  is metrically subregular at a  $\widehat{u} \in T^{-1}(0)$  for  $\widehat{w} = 0$  with the factor  $\kappa > 0$  and radius  $\delta > 0$ . Given  $u^k \in B(\widehat{u}, \delta)$ , let  $u^{k+1} := \text{prox}_{\tau g}(u^k - \tau \nabla f(u^k))$  be generated through the forward-backward method with  $\tau > 0$  satisfying  $L\tau \leq 2(1 - \varepsilon)$  for an  $\varepsilon > 0$ . Then  $u^{k+1} \in B(\widehat{u}, \delta)$ , and, for  $\rho = (1 - L\tau/2 - \varepsilon)\tau^2 / (2\kappa^2(1 + L^2\tau^2)) > 0$  and any  $\alpha > 0$ , we have*

$$(4.8) \quad 2\sqrt{\frac{1+\rho}{2+\alpha}} \text{dist}(u^{k+1}, T^{-1}(0)) + 2\sqrt{\frac{\varepsilon}{2+\alpha^{-1}}} \|u^{k+1} - u^k\| \leq \text{dist}(u^k, T^{-1}(0)).$$

*Proof.* We take  $\Xi = 0$  and  $M = \tau^{-1} \text{Id}$  in Theorem A.2, observing that metric subregularity (with respect to the standard norms) with the modulus  $\kappa$  translates to metric subregularity with respect to the  $M$  and  $M^{-1}$  norms with modulus  $\kappa_M = \kappa/\tau$  and radius  $\delta/\tau$ . Likewise, the Lipschitz property of  $\nabla f$  translates to the one with respect to these norms with factor  $L_M = L\tau$ .  $\square$

**Corollary A.5 (Nonconvex  $g$ ).** *Suppose that  $g$  is globally nonconvex, however, convex within  $B(\widehat{u}, \delta')$  for some  $\delta'$ . For a convex  $f : U \rightarrow \overline{\mathbb{R}}$  with an  $L$ -Lipschitz gradient, let*

$$T(u) = \begin{cases} \partial[g + f + \delta_{B(\widehat{u}, \delta')}] (u), & u \in \text{int } B(\widehat{u}, \delta'), \\ \emptyset, & \text{otherwise.} \end{cases}$$

*Assume that  $\widehat{u} \in T^{-1}(0)$ , and that  $T$  is metrically subregular at  $\widehat{u}$  with factor  $\kappa > 0$  and radius  $\delta \in (0, \delta')$ . Also assume the lower bound  $\inf(g + f) \geq -C$  for some  $C \in \mathbb{R}$ , and take  $\tau > 0$  small enough that  $\tau L < 1$  as well as  $2\tau([g + f](u^0) + C) < (1 - \tau L)(\delta' - \delta)^2$ . Given  $u^k \in B(\widehat{u}, \delta)$ , update  $u^{k+1} = \arg \min_u g(u) + \langle \nabla f(u), u - u^k \rangle + \frac{1}{2\tau} \|u - u^k\|^2$ . Then  $u^{k+1} \in B(\widehat{u}, \delta)$  and (4.8) holds with  $\varepsilon = 1/2$ .*

The minimisation-form update here is, in the convex case, equivalent to the subdifferential form (A.1) with  $\Xi = 0$  and  $M = \tau^{-1} \text{Id}$ .

*Proof.* The update gives  $g(u^{k+1}) + \langle \nabla f(u^k), u^{k+1} - u^k \rangle + \frac{1}{2\tau} \|u^{k+1} - u^k\|^2 \leq g(u^k)$ . Using here the descent inequality for  $f$ , equivalent to the  $L$ -Lipschitz gradient (see, e.g., [13, Theorem 7.1]), we obtain  $[g + f](u^{k+1}) + \frac{1-\tau L}{2\tau} \|u^{k+1} - u^k\|^2 \leq [g + f](u^k)$ . Thus  $\{[g + f](u^k)\}_{k \in \mathbb{N}}$  is non-increasing, and, also using our step length bound,  $\|u^{k+1} - u^k\|^2 \leq \frac{2\tau}{1-\tau L} ([g + f](u^0) + C) < (\delta' - \delta)^2$ . Since  $u^k \in B(\widehat{u}, \delta)$ , it follows that  $u^{k+1} \in \text{int } B(\widehat{u}, \delta')$ . We also have  $L\tau \leq 2(1 - \varepsilon)$  for  $\varepsilon = 1/2$ . Now, since all our points of interest,  $\widehat{u}, u^k, u^{k+1}$  are in the interior of the region of convexity  $B(\widehat{u}, \delta')$  “a priori”, the proof of Theorem 4.6 goes through “a posteriori”.  $\square$

**Theorem 4.7 (Contractivity of primal-dual proximal splitting under metric subregularity).** *Let  $G$  and  $M$  be as in Example 4.2. For a fixed  $x$ , suppose  $T = G(\cdot; x)$  is metrically subregular with respect to the  $M$  and  $M^{-1}$ -norms at  $\widehat{u}$  for  $\widehat{w} = 0$  with the modulus  $\kappa_M > 0$  and radius  $\delta > 0$ . Given  $u^k = (u_p^k, u_d^k) \in B_M(\widehat{u}, \delta)$ , let  $u^{k+1}$  be generated through the PDPS (4.4) with  $\tau_p, \tau_d > 0$  satisfying  $\frac{1}{2(1-\varepsilon)} L\tau_p + \tau_p \tau_d \|K\|^2 \leq 1$  for an  $\varepsilon \in (0, 1)$ . Then  $u^{k+1} \in B_M(\widehat{u}, \delta)$  and (A.4) holds for  $L_M = L\tau_p / (1 - \tau_p \tau_d \|K\|^2)$  and any  $\alpha > 0$ .*

*Proof.* We can write the PDPS (4.4) as (A.1) with  $\Xi = \begin{pmatrix} K & 0 \\ -K & 0 \end{pmatrix}$ , and the liftings  $\bar{f}(u) = (e(u_p; x), 0)$  and  $\bar{g}(u) = (f_0(u_p; x), g^*(u_d; x))$ . We have for  $N = (\text{Id} - \tau_d \tau_p K K^*)^{-1}$  that

$$M^{-1} = \begin{pmatrix} A_{11} & \tau_p \tau_d K^* N \\ \tau_p \tau_d N K & \tau_d N \end{pmatrix} \quad \text{for} \quad A_{11} := \tau_p \text{Id} + \tau_p^2 \tau_d K^* N K.$$

We have

$$A_{11} \leq \tau_p [1 + \tau_p \tau_d \|K\|^2 / (1 - \tau_p \tau_d \|K\|^2)] \text{Id} = [\tau_p / (1 - \tau_p \tau_d \|K\|^2)] \text{Id}.$$

We also have  $\|u\|_M^2 \geq \tau_p^{-1} (1 - \tau_p \tau_d \|K\|) \|u_p\|^2$ . Since  $e$  has an  $L$ -Lipschitz gradient, it follows

$$\begin{aligned} \|\nabla \bar{f}(u) - \nabla \bar{f}(\tilde{u})\|_{M^{-1}}^2 &= \|\nabla e(u_p; x) - \nabla e(\tilde{u}_p; x)\|_{A_{11}}^2 \\ &\leq [\tau_p / (1 - \tau_p \tau_d \|K\|^2)] L^2 \|u_p - \tilde{u}_p\|^2 \leq [\tau_p L / (1 - \tau_p \tau_d \|K\|^2)]^2 \|u - \tilde{u}\|_M^2. \end{aligned}$$

That is, the lifted function  $\bar{f}$  is  $L_M$ -Lipschitz in the  $M^{-1}$  and  $M$  norms.

The step length condition guarantees  $\tau_p \tau_d \|K\|^2 < 1$ , hence by [13, Lemma 9.12] the positive definiteness (in addition to self-adjointness by construction) of  $M$ . The result now follows from [Theorem A.2](#).  $\square$

## APPENDIX B PROOFS OF AUXILIARY RESULTS FOR TOTAL VARIATION

Here we prove further auxiliary results that we did not prove in the main text.

**Lemma 5.3.** *Let  $z, y_*, y \in \mathbb{R}^2, C > 0$ . Then*

$$(5.14) \quad \widehat{D}^* \partial \delta_{B_{\mathbb{R}^2}(0,C)}(z|y_*)(y) = \begin{cases} \{0\} & \text{if } \|z\| < C, y_* = 0, y \in \mathbb{R}^2, \\ z\mathbb{R} + \frac{\|y_*\|}{\|z\|} y & \text{if } \|z\| = C, y_* \in z(0, \infty), \langle y, z \rangle = 0, \\ z[0, \infty) & \text{if } \|z\| = C, y_* = 0, \langle y, z \rangle \geq 0, \\ \emptyset & \text{otherwise} \end{cases}$$

and

$$(5.15) \quad D^* \partial \delta_{B_{\mathbb{R}^2}(0,C)}(z|y_*)(y) = \begin{cases} \{0\} & \text{if } \|z\| < C, y_* = 0, y \in \mathbb{R}^2, \\ z\mathbb{R} + \frac{\|y_*\|}{\|z\|} y & \text{if } \|z\| = C, y_* \in z[0, \infty), \langle y, z \rangle = 0, \\ z[0, \infty) & \text{if } \|z\| = C, y_* = 0, \langle y, z \rangle > 0, \\ \{0\} & \text{if } \|z\| = C, y_* = 0, \langle y, z \rangle < 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

*Proof.* We first prove (5.14), and then use it to prove (5.15). By definition, we have  $z^* \in \widehat{D}^* \partial \delta_{B_{\mathbb{R}^2}(0,C)}(z|y_*)(y)$  if and only if

$$(B.1) \quad \limsup_{\text{Graph } \partial \delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle z^*, \tilde{z} - z \rangle - \langle y, \tilde{y}_* - y_* \rangle}{\|\tilde{z} - z\| + \|\tilde{y}_* - y_*\|} \leq 0.$$

We analyse this limit in various cases:

1.  $\|z\| < C$  and  $y_* = 0$ , in which case we need to prove that  $z^* = 0$  and  $y \in \mathbb{R}^2$ : We must have  $\tilde{y}_* = 0$  to stay in  $\text{Graph } \partial \delta_{B_{\mathbb{R}^2}(0,C)}$  and thus  $-\langle y, \tilde{y}_* - y_* \rangle = 0$  for any  $y \in \mathbb{R}^2$ . Since we have  $B(z, \varepsilon) \times \{0\} \subset \text{Graph } \partial \delta_{B_{\mathbb{R}^2}(0,C)}$  for small enough  $\varepsilon$ , and  $z + \frac{1}{n}z^* \in B(z, \varepsilon)$  for  $n > \varepsilon^{-1}\|z^*\|$ , taking  $\tilde{z} = z + \frac{1}{n}z^*$  and  $\tilde{y}_* = 0$  we get

$$\limsup_{\text{Graph } \partial \delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle z^*, \tilde{z} - z \rangle - \langle y, \tilde{y}_* - y_* \rangle}{\|\tilde{z} - z\| + \|\tilde{y}_* - y_*\|} \geq \limsup_{n \rightarrow 0} \frac{\langle z^*, \frac{1}{n}z^* \rangle}{\|\frac{1}{n}z^*\|} = \|z^*\|.$$

Therefore, we must have  $z^* = 0$  for (B.1) to hold.

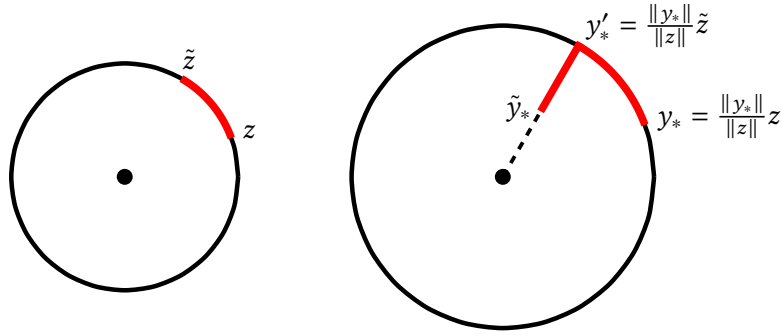


Figure B.1: Demonstrating the paths of the limiting progress  $\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)$ .

2.  $\|z\| = C$  and  $y_* \in z(0, \infty)$ , in which case we need to prove that  $\langle y, z \rangle = 0$  and  $z^* \in z\mathbb{R} + \frac{\|y_*\|}{\|z\|} y$ : We must in (B.1) have  $\langle y, z \rangle = 0$  for (B.1) to hold because otherwise either  $\tilde{y}_* = (1 - \varepsilon)y_*$  or  $\tilde{y}_* = (1 + \varepsilon)y_*$  for some  $\varepsilon > 0$  gives  $-\langle y, \tilde{y}_* - y_* \rangle > 0$  regardless of  $\tilde{z}$ . Since  $\langle y, z \rangle = 0$ , the vectors  $y$  and  $z$  form basis for  $\mathbb{R}^2$  and we can write  $z^* = sz + ty$  for some  $s, t \in \mathbb{R}$ .

Figure B.1 demonstrates the path from  $\tilde{z}$  to  $z$ , along the circle  $\{v \in \mathbb{R}^2 \mid \|v\| = C\}$ , and how the path from  $\tilde{y}_*$  to  $y_*$  can be divided into two components,  $\tilde{y}_* - y'_*$  and  $y'_* - y_*$ . For  $(z, y_*)$ ,  $(\tilde{z}, \tilde{y}_*) \in \text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)}$  it holds that  $y'_* = \frac{\|y_*\|}{\|z\|} z'$  and  $y_* = \frac{\|y_*\|}{\|z\|} z$ . The limiting behaviour  $(\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)$  is such that

$$\begin{aligned} \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle z, \tilde{z} - z \rangle}{\|\tilde{z} - z\|} &= \limsup_{\|\tilde{z}\|=C, \tilde{z} \rightarrow z} \frac{\langle z, \tilde{z} - z \rangle}{\|\tilde{z} - z\|} = 0 \quad \text{and} \\ \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle y, \tilde{y}_* - y'_* \rangle}{\|\tilde{y}_* - y'_*\|} &= \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle y, z \rangle}{\|\tilde{y}_* - y_*\|} = 0. \end{aligned}$$

Therefore we have

$$\begin{aligned} & \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle z^*, \tilde{z} - z \rangle - \langle y, \tilde{y}_* - y_* \rangle}{\|\tilde{z} - z\| + \|\tilde{y}_* - y_*\|} \\ &= \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle sz + ty, \tilde{z} - z \rangle - \langle y, \tilde{y}_* - y'_* + y'_* - y_* \rangle}{\|\tilde{z} - z\| + \|\tilde{y}_* - y_*\|} \\ &= \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle ty, \tilde{z} - z \rangle - \langle y, y'_* - y_* \rangle}{\|\tilde{z} - z\| + \|\tilde{y}_* - y_*\|} \\ &= \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \left( t - \frac{\|y_*\|}{\|z\|} \right) \frac{\langle y, \tilde{z} - z \rangle}{\|\tilde{z} - z\| + \|\tilde{y}_* - y_*\|}. \end{aligned}$$

Since  $\limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle y, \tilde{z} - z \rangle}{\|\tilde{z} - z\|} = \|y\|$  we must have  $t = \|y_*\|/\|z\|$  for (B.1) to hold.

3.  $\|z\| = C$ ,  $x > 0$ , and  $y_* = 0$ , in which case we need to prove that  $\langle y, z \rangle \geq 0$  and  $z^* \in z[0, \infty)$ : Write  $z^* = sz + tv$  for some  $s, t \in \mathbb{R}$  and  $v$  such that  $\langle z, v \rangle = 0$ . Taking  $\tilde{y}_* = y_*$  we obtain that the limsup in (B.1) is bounded below by  $\|tv\| \|\tilde{z} - z\| / \|\tilde{z} - z\| = \|tv\|$  and thus we must have  $t = 0$ . On the other hand, taking  $\tilde{z} = z$  and  $\langle y, \tilde{y}_* - y_* \rangle < 0$  we find again that (B.1) is bounded below and thus we must have  $\langle y, \tilde{y}_* - y_* \rangle \geq 0$ . Since  $\tilde{y}_* - y_* = \tilde{y}_* \in z[0, \infty)$  we get the equivalent condition  $\langle y, z \rangle \geq 0$ . Similarly to  $\tilde{y}_* - y_*$  in the previous case and Figure B.1, we can split  $\tilde{z} - z$

into the components  $\tilde{z} - z'$  and  $z' - z$ . Using this split,  $\tilde{y}_* - y_* \in z[0, \infty)$  and  $\langle y, z \rangle \geq 0$ , we get

$$\limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{\langle z^*, \tilde{z} - z \rangle - \langle y, \tilde{y}_* - y_* \rangle}{\|\tilde{z} - z\| + \|\tilde{y}_* - y_*\|} = \limsup_{\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)} \ni (\tilde{z}, \tilde{y}_*) \rightarrow (z, y_*)} \frac{-s\|z\|\|z' - z\|}{\|\tilde{z} - z\|}.$$

This is less than zero, i.e., (B.1) holds, if and only if  $s \geq 0$ .

4. Otherwise: the point  $(z, y_*)$  does not belong to the closed  $\text{Graph } \partial\delta_{B_{\mathbb{R}^2}(0,C)}$  and thus (B.1) cannot be satisfied.

The limiting coderivative equals the Fréchet coderivative with the following exceptions:

- when  $\|z\| = C$  and  $\langle y, z \rangle = 0$ , we have  $\widehat{D}^* \partial\delta_{B_{\mathbb{R}^2}(0,C)}(z|0)(y) = z[0, \infty)$  but taking  $y_*^k \in z(0, \infty)$  such that  $y_*^k \rightarrow 0$  we obtain  $D^* \partial\delta_{B_{\mathbb{R}^2}(0,C)}(z|0)(y) = z\mathbb{R}$  from the definition of (finite dimensional) limiting coderivative and  $\widehat{D}^* \partial\delta_{B_{\mathbb{R}^2}(0,C)}(z|y_*^k)(y) = z\mathbb{R} + \frac{\|y_*^k\|}{\|z\|} y$ ,
- when  $\|z\| = C$ ,  $y_* = 0$  and  $\langle y, z \rangle < 0$ , the limiting coderivative is the singleton  $\{0\}$  since  $0 \in \widehat{D}^* \partial\delta_{B_{\mathbb{R}^2}(0,C)}(z^k|0)(y)$  for  $z^k$  such that  $\|z^k\| < C$  and  $z^k \rightarrow z$ .  $\square$

## APPENDIX C SCALAR TRACKING RESULTS

We recall from [18] the following scalar tracking results. In our work, we take  $d_k^u = \|u^k - S_u(x^{k-1})\|_U + \|\Delta w_*^k\|_*$ ,  $d_k^w = \|w^k - S_w(x^{k-1})\|_W$ ,  $\varrho_k = \|x^k - x^{k-1}\|$  and  $\tilde{d}_k = \|\tilde{x}_{k+1}^* - x_{k+1}^*\|$ .

**Assumption c.1** ([18, Assumption A.1]). For a given  $k \geq 0$  and scalars  $d_0^u, \dots, d_{k+1}^u, d_0^w, \dots, d_{k+1}^w, \varrho_1, \dots, \varrho_k \geq 0$ , and  $\tilde{d}_0, \dots, \tilde{d}_k \in \mathbb{R}$ , there exist  $\pi_u, \pi_w, \mu_u \alpha_w, \alpha_u > 0$ , such that

- (i)  $\kappa_u d_{j+1}^u \leq d_j^u + \pi_u \varrho_j$ , for all  $j = 1, \dots, k$ ,
- (ii)  $\kappa_w d_{j+1}^w \leq d_j^w + \mu_u d_{k+1}^u + \pi_w \varrho_j$ , for all  $j = 1, \dots, k$ , and
- (iii)  $\tilde{d}_j \leq \alpha_u d_{j+1}^u + \alpha_w d_{j+1}^w$  for all  $j = 0, \dots, k$ .

To simplify the following results, we introduce  $\kappa := \min\{\kappa_u, \kappa_w\}$  and  $\bar{\kappa} := \max\{\kappa_u, \kappa_w\}$ .

**Lemma c.2** ([18, Lemma A.3]). *Suppose Assumption c.1 holds for a  $k \geq 0$ . Then for any  $p \in (0, \kappa)$ , we have*

$$(c.1) \quad \tilde{d}_k^2 \leq (\alpha_u d_{k+1}^u + \alpha_w d_{k+1}^w)^2 \leq \check{e}_{p,k},$$

where, for  $\psi_j := \alpha_u \kappa_u^{-j} \pi_u + \alpha_w [\iota_j \mu_u \pi_u + \kappa_w^{-j} \pi_w]$  and, we set

$$(c.2) \quad \zeta_p := \frac{\bar{\kappa}}{p} \sum_{j=0}^{\infty} p^j \psi_j \leq \frac{(\alpha_u \pi_u + \alpha_w \pi_w) \kappa \bar{\kappa}}{p(\kappa - p)} + \frac{\alpha_w \mu_u \pi_u \bar{\kappa}}{p^2(\kappa - p)^2} \quad \text{and}$$

$$(c.3) \quad \check{e}_{p,k} := \frac{\zeta_p (\alpha_u \kappa_u^{-k} + \alpha_w \iota_k \mu_u)}{\pi_u p^k} (d_1^u)^2 + \frac{\zeta_p \alpha_w \kappa_w^{-k}}{\pi_w p^k} (d_1^w)^2 + \sum_{j=0}^{k-1} \frac{\zeta_p \psi_{k-j}}{p^{k-j}} \varrho_{j+1}.$$

The second inequality of (c.1) holds even if Assumption c.1 (iii) does not.

**Lemma c.3** ([18, Lemma A.4]). *Suppose Assumption c.1 holds for a  $k \geq 0$ . Then for any  $p \in (0, \kappa)$ , we have*

$$(c.4) \quad \widetilde{d}_k^2 \leq \varsigma_p^2 \varrho_{k+1}^2 + e_{p,k},$$

where, for  $\check{e}_{p,k}$  defined (c.3),

$$(c.5) \quad e_{p,k} := \check{e}_{p,k} - \varsigma_p^2 \varrho_{k+1}^2$$

satisfies

$$(c.6) \quad \sup_{N \in \mathbb{N}} \sum_{k=0}^{N-1} p^k e_{p,k} \leq \Psi_p := \frac{(d_1^u)^2}{\pi_u} \left( \frac{\varsigma_p \alpha_u \kappa}{\kappa - 1} + \frac{\varsigma_p \alpha_w \mu_u}{(\kappa - 1)^2} \right) + \frac{(d_1^w)^2}{\pi_w} \left( \frac{\varsigma_p \alpha_w \kappa}{\kappa - 1} \right).$$

**Lemma c.4** ([18, Lemma A.5]). *Suppose Assumption c.1 holds for a  $k \in \mathbb{N}$ . Then, for  $\check{e}_{1,n}$  given in (c.3) we have  $\sum_{n=0}^{k-1} \check{e}_{1,n} \leq \Psi_1 + \varsigma_1^2 \sum_{n=0}^{k-1} \varrho_{n+1}^2$ .*