HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Regularisation, optimisation, subregularity

Valkonen, T.

2021-03-16

Valkonen , T 2021 , ' Regularisation, optimisation, subregularity ' , Inverse Problems , vol. 37 , no. 4 , 045010 . https://doi.org/10.1088/1361-6420/abe4aa

http://hdl.handle.net/10138/341673
https://doi.org/10.1088/1361-6420/abe4aa

acceptedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

REGULARISATION, OPTIMISATION, SUBREGULARITY

Tuomo Valkonen*

Abstract Regularisation theory in Banach spaces, and non–norm-squared regularisation even in finite dimensions, generally relies upon Bregman divergences to replace norm convergence. This is comparable to the extension of first-order optimisation methods to Banach spaces. Bregman divergences can, however, be somewhat suboptimal in terms of descriptiveness. Using the concept of *(strong) metric subregularity*, previously used to prove the fast local convergence of optimisation methods, we show norm convergence in Banach spaces and for non–norm-squared regularisation. For problems such as total variation regularised image reconstruction, the metric subregularity reduces to a geometric condition on the ground truth: flat areas in the ground truth have to compensate for the fidelity term not having second-order growth within the kernel of the forward operator. Our approach to proving such regularisation results is based on optimisation formulations of inverse problems. As a side result of the regularisation theory that we develop, we provide *regularisation complexity* results for optimisation methods: how many steps N_δ of the algorithm do we have to take for the approximate solutions to converge as the corruption level $\delta \searrow 0$?

1 INTRODUCTION

Let $A \in C^1(X;Y)$ between a Banach space X of unknowns and a Hilbert space Y of measurements. A common approach to solving the inverse problem $A(x) = \hat{b}$, having access only to corrupted measurements b^δ of the true data $\hat{b}$, is to solve for some regularisation parameter $\alpha_\delta > 0$ and a convex regularisation functional R the Tikhonov-style regularised problem

$$\min_{x\in X} J_\delta(x) + \alpha_\delta R(x) \quad \text{where} \quad J_\delta(x) := \frac{1}{2}\|A(x) - b_\delta\|_Y^2. \tag{1.1}$$

We then want to know whether solutions x^δ to this problem converge to a solution (or ground-truth) of the original equation as $\delta \searrow 0$. One typically fixes a specific solution

$$\hat{x} \in \operatorname*{arg\,min}_x \{R(x) \mid A(x) = \hat{b}\}. \tag{1.2}$$

Conventional regularisation theory [19] treats the case $R(x) = \frac{1}{2}\|x\|_X^2$ with X also a Hilbert space. In this case $\hat{x}$ is the minimum-norm solution. Norm convergence $x^\delta \to \hat{x}$ can be shown provided $\hat{x} \in \operatorname{ran} A'(\hat{x})^*$ and $\alpha_\delta \searrow 0$ as well as $\delta^2/\alpha_\delta \searrow 0$ as $\delta \searrow 0$.

How about Banach spaces X, and more complicated regularisation functionals R, such as non-differentiable sparsity-inducing L^1-norm regularisation, total variation regularisation, and its generalisations [9]? Let $C := \{x \in X \mid A(x) = \hat{b}\}$. The first-order optimality conditions[1] or Fermat principle

This research has been supported by the Academy of Finland grants 314701 and 320022.

*Department of Mathematics and Statistics, University of Helsinki, Finland *and* ModeMat, Escuela Politécnica Nacional, Quito, Ecuador, tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572

[1]Necessary and sufficient if R is convex, proper, and lower semicontinuous, and A is linear. More generally necessary; see Section 5.

for (1.2) requires that $0 \in \partial[\delta_C + \alpha_\delta \partial R](\hat{x})$, where ∂ denotes the convex subdifferential, and δ_C the $\{0, \infty\}$-valued indicator function of C. This condition can for some $w \in Y^*$ be expanded as

$$0 \in A'(\hat{x})^* w + \partial R(\hat{x}),$$

which is known in the inverse problems literature as a *source condition*. It encodes the existence of an R-minimising ground-truth. If the source condition holds, then it is shown in [10] that the *Bregman divergence*

$$B_R^{\bar{d}}(x, \bar{x}) := \langle \bar{d} | \bar{x} - x \rangle + R(x) - R(\bar{x}) \quad (x, \bar{x} \in X;\ \bar{d} \in \partial R(\bar{x})) \tag{1.3}$$

satisfies $B_R^{-A'(\hat{x})^* w}(x^\delta, \hat{x}) \to 0$ provided the noise and regularisation parameters $\delta > 0$ and $\alpha_\delta > 0$ convergence as in the conventional norm-squared case. We refer to [33] for the use of Bregman divergences in Banach space regularisation theory.

Convergence of Bregman divergences is, however, a relatively weak result. It does not imply norm convergence unless the Bregman divergence is *elliptic* in the words of [34], i.e., $B_R^{\bar{d}}(x, \bar{x}) \geq \gamma \|x - \bar{x}\|^2$ for some $\gamma > 0$. This simply means that R is strongly convex, something that is not satisfied by L^1 and total variation regularisation. Often, however, not the regulariser itself but the entire objective of (1.1) admits a type of *local* strong convexity.

Example 1.1. On $\mathbb{R}^2$, take $A(x_1, x_2) = x_1$ and $R(x_1, x_2) = |x_2|$. Then, due to the growth properties of the absolute value function at zero, $J_\delta + \alpha_\delta R$ is locally strongly convex at $(x_1, 0)$, but not at (x_1, x_2) for $x_2 \neq 0$.

If A is linear, then $J_\delta + \alpha_\delta R$ is convex. In this case local strong convexity is equivalent to the *strong metric subregularity* of the subdifferential $\partial[J_\delta(x) + \alpha_\delta R]$ [4, 3]. Strong metric subregularity, and (non-strong) metric subregularity introduced in [25, 17], are Lipschitz-like properties of set-valued maps. We briefly recall their definitions and interpretations in Section 2, referring to [32, 24, 14] for more comprehensive introductions. To see how such concepts can be used in optimisation, we refer to [37].

We will exploit strong metric subregularity and an intermediate concept between strong and non-strong metric subregularity to show for general regularisers R the norm convergence of *approximate* solutions to (1.1). We do this first in Section 3 for linear inverse problems, and then in Section 5 for nonlinear inverse problems. In Section 5 we also generalise the results to general data discrepancies E in place of the squared norm. In Section 4 we verify the relevant subregularity—expressed as a *strong source condition*—for ℓ^1 regularisation in finite dimensions, and for total variation regularisation in $L^2(\Omega)$. In both cases, the lack of ellipticity of A^*A (i.e., $A^*A \ngeq \gamma \operatorname{Id}$ for some $\gamma > 0$) has to be compensated for by the regularisation term. In the first case, we obtain this unconditionally, whereas for total variation regularisation our results are more preliminary and require the ground-truth to have "strictly flat areas" that perform this compensation.

We finish in Section 6 with interpretations of the regularisation results of Section 3 as "regularisation complexity" results for optimisation methods: forward-backward splitting and primal-dual proximal splitting. We derive expressions for how many iterations N_δ of the algorithm are needed to produce approximate solutions x^δ to (1.1) that converge to $\hat{x}$ as $\delta \searrow 0$.

Our proof approach is intrinsically based around treating inverse problems as optimisation problems. In this sense, our work is closely related to [23], which also proves convergence in a "hybrid" topology involving the Bregman divergence and the norm in the data space: $B_R^{-A'(\hat{x})^* w}(x^\delta, \hat{x}) + \|A(x^\delta) - A(\hat{x})\| \to 0$. More general optimisation-based formulations of inverse problems are treated in [27], whereas the weak-$*$ convergence of solutions to total generalised variation regularised linear inverse problems is treated in [8]. Norm convergence of solutions to multibang control problems is studied in [16]. There

the metric subregularity R is employed, however, not that of the entire Tikhonov-regularised objective as we will do.

Iterative regularisation methods [28] also closely tie optimisation methods to regularisation, however, this is different from the "regularisation complexity" results of Section 6: iterative regularisation adapts the parameter $\alpha_\delta > 0$ within each step of the optimisation method, whereas we simply want to know how many steps of the method we need to take for fixed α_δ. Numerically efficient iterative regularisation is largely limited to smooth regularisers R as nonsmooth approaches require solving a difficult "inner" optimisation problem in each step of the "outer" method [5].

Finally, metric regularity—a concept stronger than metric subregularity and distinct from strong metric subregularity—has been used in [21] to study the regularisation of set-valued inclusions $0 \in T(x)$ via the Tikhonov-style expression $0 \in [T+\alpha_\delta \operatorname{Id}](x^\delta)$. The incorporation of the identity map Id requires T to be a set-valued map form X to X, whereas subdifferentials are set-valued maps from X to X^*.

NOTATION AND ELEMENTARY RESULTS

We denote the extended reals by $\overline{\mathbb{R}} := [-\infty, \infty]$. We write $H : X \rightrightarrows Y$ when H is a set-valued map from the space X to Y. For Fréchet differentiable $F : X \to R$, we write $F'(x) \in X^*$ for the Fréchet derivative at $x \in X$. Here X^* is the dual space to X. For a convex function $F : X \to \overline{\mathbb{R}}$, we write $\partial F : X \rightrightarrows X^*$ for its subdifferential map. On a normed space X, for a point $x \in X$ and a set $U \subset X$, we write $\operatorname{dist}(x, U) := \inf_{x' \in U} \|x - x'\|_X$, where $\|\cdot\|_X$ is the norm on X. We also write $\operatorname{dist}^2(x, U) := \operatorname{dist}(x, U)^2$. We write $\langle x, x' \rangle$ for the inner product between two elements x and x' of a Hilbert space X, and $\langle x^* | x \rangle := x^*(x)$ for the dual product or dual pairing in a Banach space. We write $\operatorname{Id} : X \to X$ for the identity operator on X and $\delta_A : X \to \overline{\mathbb{R}}$ for the $\{0, \infty\}$-valued indicator function of a set $A \subset X$.

For X a Hilbert space, we will frequently use Pythagoras' three-point identity

$$\langle x - y, x - z \rangle_X = \frac{1}{2}\|x - y\|_X^2 - \frac{1}{2}\|y - z\|_X^2 + \frac{1}{2}\|x - z\|_X^2 \quad (x, y, z \in X) \tag{1.4}$$

and (inner product) Young's inequality

$$\langle x, y \rangle \le \|x\|_X \|y\|_X \le \frac{1}{2\alpha}\|x\|_X^2 + \frac{\alpha}{2}\|y\|_X^2 \quad (x, y \in X,\ \alpha > 0). \tag{1.5}$$

2 METRIC SUBREGULARITY AND LOCAL SUBDIFFERENTIABILITY

We recall that a set-valued function $H : X \rightrightarrows Y$ is *metrically subregular* at $\bar{x}$ for $\bar{y}$ if $\bar{y} \in H(\bar{x})$ and there exists a constant $\kappa > 0$ as well as neighbourhoods $U \ni \bar{x}$ and $V \ni \bar{y}$ such that

$$\operatorname{dist}(x, H^{-1}(\bar{y})) \le \kappa \operatorname{dist}(\bar{y}, H(x) \cap V) \quad (x \in U).$$

If the stronger inequality

$$\|x - \bar{x}\|_X \le \kappa \operatorname{dist}(\bar{y}, H(x) \cap V) \quad (x \in U)$$

holds, then we say that H is *strongly metrically subregular* at $\bar{x}$ for $\bar{y}$. The latter property can equivalently be stated as H being metrically subregular with $\bar{x}$ an isolated point of $H^{-1}(\bar{y})$.

We recall from [3, 4] the following characterisations of the metric subregularity and strong metric subregularity of convex subdifferentials.

Theorem 2.1 ([3, Theorem 3.3]). *Let $F : X \to \overline{\mathbb{R}}$ be a convex function on a Banach space X, $\bar{x} \in X$, and $\bar{x}^* \in \partial F(\bar{x})$. Then ∂F is metrically subregular at $\bar{x}$ for $\bar{x}^*$ if and only if there exists a neighbourhood U of $\bar{x}$ and a constant $\gamma > 0$ such that*

$$F(x) \ge F(\bar{x}) + \langle \bar{x}^* | x - \bar{x} \rangle + \gamma \operatorname{dist}^2(x, [\partial F]^{-1}(\bar{x}^*)) \quad (x \in U). \tag{2.1}$$

More precisely, (2.1) *implies metric subregularity with* $\kappa = 1/\gamma$ *while metric subregularity implies* (2.1) *for any* $0 < \gamma < 1/(4\kappa)$.

Theorem 2.2 ([3, Theorem 3.5]). *Let* $F : X \to \overline{\mathbb{R}}$ *be a convex function on a Banach space* X, $\bar{x} \in X$, *and* $\bar{x}^* \in \partial F(\bar{x})$. *Then* ∂F *is strongly metrically subregular at* $\bar{x}$ *for* $\bar{x}^*$ *if and only if there exists a neighbourhood* U *of* $\bar{x}$ *and a constant* $\gamma > 0$ *such that*

$$F(x) \geq F(\bar{x}) + \langle \bar{x}^* | x - \bar{x} \rangle + \gamma \|x - \bar{x}\|_X^2 \quad (x \in U). \tag{2.2}$$

More precisely, (2.2) *implies strong metric subregularity with* $\kappa = 1/\gamma$ *while strong metric subregularity implies* (2.2) *for any* $0 < \gamma < 1/(4\kappa)$.

We call the expressions (2.1) and (2.2) the *semi-strong* and *strong local subdifferentiability* of F at $\bar{x}$ for $\bar{x}^*$. Compared to standard strong subdifferentiability[2], both conditions localise the notion to a neighbourhood of $\bar{x}$. Moreover, (2.1) in a sense squeezes the set $[\partial F]^{-1}(\bar{x}^*)$ into a single point. Consider, for example, $F(x) = \max\{0, |x| - 1\}$ in $\mathbb{R}$. Let $\bar{x} \in [-1, 1]$ be arbitrary and $\bar{x}^* = 0$. Then $[\partial F]^{-1}(\bar{x}^*) = [-1, 1]$, so that (2.1) only requires F to grow once x leaves $[-1, 1] \ni \bar{x}$ instead of immediately as it leaves $\{\bar{x}\}$ as is required by (2.2). In the context of optimisation methods, (2.1) is useful for the study of convergence of iterates to an entire set of minimisers of F, without singling out one, while (2.2) is useful for studying the convergence to a known specific minimiser. In the following Section 3, we will likewise work both with a set of ground-truths $\hat{X}$ and a specific ground-truth $\hat{x}$.

Minding the locality of the definitions, using Clarke subdifferentials [13], it is not difficult to extend Theorems 2.1 and 2.2 to locally convex functions, i.e., non-convex functions that locally satisfy a second-order growth condition. However, to get useful regularisation results, we need to replace (non-strong) metric subregularity with an intermediate and slightly relaxed version, motivated by the notions of partial submonotonicity and subregularity introduced in [37]. We therefore handle non-convex F through direct definitions analogous to (2.1) and (2.2).

Namely, on Banach spaces X and Y, let $A \in \mathbb{L}(X; Y)$ and let $F : X \to \overline{\mathbb{R}}$ be a (possibly non-convex) function. Also pick $\mu > 0$. Then we say that f is (A, μ)*-strongly locally subdifferentiable* at $\bar{x} \in X$ for $\bar{x}^* \in X^*$ with respect to a set $\bar{X} \subset X$ if there exists a neighbourhood U of $\bar{x}$ and a constant $\gamma > 0$ such that

$$F(x) \geq F(\bar{x}) + \langle \bar{x}^* | x - \bar{x} \rangle + \gamma \|A(x - \bar{x})\|_Y^2 + \gamma\mu \operatorname{dist}^2(x, \bar{X}) \quad (x \in U). \tag{2.3}$$

Typically $\bar{X} \ni \bar{x}$. We do not assume $\bar{x}^*$ to be a subdifferential of F at $\bar{x}$ in any conventional sense. The idea is that for $F(x) = \frac{1}{2}\|Ax - b_\delta\|_Y^2 + \alpha_\delta R(x)$ as in (1.1), we will in $\operatorname{ran} A^*$ exploit the available growth away from $\bar{x}$, but otherwise only the growth away from the set $\bar{X}$.

We will in Section 4 provide examples of the different forms of strong local subdifferentiability and strong metric subregularity. Before this, in the next section, we will use them to derive regularisation theory.

3 LINEAR INVERSE PROBLEMS

We now derive subregularity-based regularisation theory for linear inverse problems. For all corruption parameters $\delta > 0$ and measured data $b_\delta \in Y$, we approximate solutions $\hat{x} \in X$ to the problem

$$A\hat{x} = \hat{b}, \tag{3.1}$$

with unknown data $\hat{b}$, through the regularised problems

$$\min_{x \in X} J_\delta(x) + \alpha_\delta R(x) \quad \text{where} \quad J_\delta(x) := \frac{1}{2}\|Ax - b_\delta\|_Y^2. \tag{3.2}$$

[2] Equivalent to strong convexity in Hilbert spaces; see, e.g., [14].

3.1 GENERAL ASSUMPTIONS AND CONCEPTS

We denote the set of R-minimising solutions to (3.1) by $\hat{X}$. Under the assumptions that we state next, these are characterised exactly through the satisfaction for some $\hat{w} \in Y$ of the *basic source condition*

$$A\hat{x} = \hat{b} \quad \text{and} \quad A^*\hat{w} + \partial R(\hat{x}) \ni 0. \tag{3.3}$$

Assumption 3.1 (Corruption level and solution accuracy). On a Banach space X and a Hilbert space Y, the regularisation functional $R : X \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous, and $A \in \mathbb{L}(X; Y)$. The parametrisation $\delta > 0$ of the corruption level is such that the corrupted measurements $b_\delta \in Y$ of the ground-truth $\hat{b} \in Y$ satisfy

$$\|b_\delta - \hat{b}\|_Y \le \delta \quad (\delta > 0). \tag{3.4}$$

Moreover, we assume that (3.2) is solved to a given accuracy $e_\delta \ge 0$ in the sense that

$$[J_\delta + \alpha_\delta R](x_\delta) - [J_\delta + \alpha_\delta R](\hat{x}) \le e_\delta \quad (\delta > 0) \tag{3.5}$$

for any given $\hat{x} \in \hat{X}$.

It does not matter which $\hat{x} \in \hat{X}$ we use in the accuracy condition (3.5) since

$$J_\delta(\hat{x}) + \alpha_\delta R(\hat{x}) = \frac{1}{2}\|\hat{b} - b_\delta\|^2 + \alpha_\delta R(\hat{x}) = J_\delta(\tilde{x}) + \alpha_\delta R(\tilde{x}) \quad (\hat{x}, \tilde{x} \in \hat{X}).$$

Also, since $\hat{x}$ is not in general a solution to (3.2) for $\alpha_\delta > 0$, even with $e_\delta = 0$, the problem (3.2) may not have to be solved to full accuracy to satisfy (3.5)

3.2 A BASIC OPTIMISATION-BASED ESTIMATE

To motivate our contributions, we start by deriving basic estimates based on Bregman divergences- These are similar to those in [10], however allow for the inexact solution of (3.2). Specifically, the following result shows that for convergence, as the corruption level $\delta \searrow 0$, we need $\alpha_\delta \searrow 0$, $\delta^2/\alpha_\delta \searrow 0$, and $e_\delta/\alpha_\delta \searrow 0$.

Theorem 3.2. *Suppose Assumption 3.1 holds and that $\hat{x} \in X$ satisfies for some $\hat{w} \in Y^*$ the basic source condition (3.3). Then*

$$0 \le B_R^{-A^*\hat{w}}(x_\delta, \hat{x}) \le \frac{e_\delta}{\alpha_\delta} + \frac{\delta^2}{\alpha_\delta} + \alpha_\delta \|\hat{w}\|_Y^2.$$

Proof. By Assumption 3.1 and expansion we have

$$\begin{aligned} e_\delta &\ge [J_\delta + \alpha_\delta R](x_\delta) - [J_\delta + \alpha_\delta R](\hat{x}) \\ &= \frac{1}{2}\|Ax_\delta - b_\delta\|_Y^2 - \frac{1}{2}\|A\hat{x} - b_\delta\|_Y^2 + \alpha_\delta[R(x_\delta) - R(\hat{x})]. \end{aligned}$$

Continuing with Pythagoras' three-point identity (1.4) we rearrange the squared norms to obtain

$$e_\delta \ge \langle A\hat{x} - b_\delta, A(x_\delta - \hat{x})\rangle + \frac{1}{2}\|A(x_\delta - \hat{x})\|_Y^2 + \alpha_\delta[R(x_\delta) - R(\hat{x})].$$

Since $A\hat{x} = \hat{b}$, using the definition (1.3) of the Bregman divergence, this further rerarranges as

$$e_\delta \ge \langle \hat{b} - b_\delta - \alpha_\delta \hat{w}, A(x_\delta - \hat{x})\rangle + \frac{1}{2}\|A(x_\delta - \hat{x})\|_Y^2 + \alpha_\delta B_R^{-A^*\hat{w}}(x_\delta, \hat{x}).$$

Now using Young's inequality (1.5) on the inner product term , we obtain

$$e_\delta \geq -\frac{1}{2}\|\hat{b} - b_\delta - \alpha_\delta \hat{w}\|_Y^2 + \alpha_\delta B_R^{-A^*\hat{w}}(x_\delta, \hat{x}). \tag{3.6}$$

Further estimating using Young's inequality that

$$\frac{1}{2\alpha_\delta}\|\hat{b} - b_\delta - \alpha_\delta \hat{w}\|_Y^2 \leq \frac{1}{\alpha_\delta}\|b_\delta - \hat{b}\|_Y^2 + \alpha_\delta\|\hat{w}\|_Y^2 \leq \frac{\delta^2}{\alpha_\delta} + \alpha_\delta\|\hat{w}\|_Y^2,$$

we therefore obtain the claim. □

Remark 3.3. In place of (3.6) we could alternatively estimate

$$\langle \hat{b} - b_\delta - \alpha_\delta \hat{w}, A(x_\delta - \hat{x})\rangle + \frac{1}{2}\|A(x_\delta - \hat{x})\|_Y^2 \geq -\frac{1}{2(1-\alpha_\delta)}\|\hat{b} - b_\delta - \alpha_\delta \hat{w}\|_Y^2 + \frac{\alpha_\delta}{2}\|A(x_\delta - \hat{x})\|_Y^2.$$

Akin to the approach of [23], we would then obtain

$$\frac{1}{2\alpha_\delta}\|A(x_\delta - \hat{x})\|_Y^2 + B_R^{-A^*\hat{w}}(x_\delta, \hat{x}) \leq \frac{e_\delta}{\alpha_\delta} + \frac{\alpha_\delta}{1-\alpha_\delta}\|\hat{w}\|_Y^2 + \frac{\delta^2}{\alpha_\delta(1-\alpha_\delta)}.$$

Thus $A(x_\delta - \hat{x}) \to 0$ significantly faster than $B_R^{-A^*\hat{w}}(x_\delta, \hat{x}) \to 0$ provided, as before, that $\alpha_\delta \to 0$, $\delta/\alpha_\delta \to 0$, and $e_\delta/\alpha_\delta \to 0$ as $\delta \searrow 0$. This motivates our next results, essentially combining, via strong metric subregularity, the two different convergences to yield norm convergence.

3.3 ESTIMATES BASED ON A STRONG SOURCE CONDITION

We start with the next lemma that we will be used to show that the approximate regularised solutions x_δ are close to $\hat{x}$ for small enough noise level, regularisation parameter, and accuracy parameter.

Lemma 3.4. *Suppose Assumption 3.1 holds at $\hat{x} \in \hat{X}$. Then*

$$\|A(x_\delta - \hat{x})\|_Y^2 \leq 4(e_\delta + \delta + \alpha_\delta R(\hat{x}))$$

and

$$R(x_\delta) \leq R(\hat{x}) + \frac{e_\delta + \delta^2}{\alpha_\delta}.$$

Proof. By Assumption 3.1, since $A\hat{x} = \hat{b}$, first using Young's inequality we have

$$\begin{aligned}\frac{1}{2}\|A(x_\delta - \hat{x})\|_Y^2 + 2\alpha_\delta R(x_\delta) &\leq \|Ax_\delta - b_\delta\|_Y^2 + 2\alpha_\delta R(x_\delta) + \|b_\delta - \hat{b}\|_Y^2 \\ &\leq 2e_\delta + 2\|b_\delta - \hat{b}\|_Y^2 + 2\alpha_\delta R(\hat{x}) \\ &\leq 2(e_\delta + \delta^2 + \alpha_\delta R(\hat{x})).\end{aligned}$$

This finishes the proof. □

We will need the following "strong source condition" based on strong metric subregularity.

Assumption 3.5 (Strong source condition). Assume that $\hat{x} \in X$ satisfies for some $\hat{w} \in Y$ the basic source condition (3.3). Moreover, for all $\delta > 0$, for given $\alpha_\delta, \gamma_\delta > 0$, assume that $J_\delta + \alpha_\delta R$ is strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta A^*\hat{w}$ with respect to the norm

$$\|x\|_\delta := \sqrt{\|Ax\|_Y^2 + \gamma_\delta\|x\|_X^2} \quad (x \in X).$$

The factor $\gamma > 0$ of strong local subdifferentiability, as defined in (2.3), must be independent of $\delta > 0$ and, for some $\rho > 0$, we must have

$$U \supset U_\rho := \{x \in X \mid \|A(x - \hat{x})\| \le \rho,\ R(x) \le R(\hat{x}) + \rho\}. \tag{3.7}$$

Then we say that $\hat{x}$ satisfies for $\hat{w}$ the *strong source condition.*

As we recall from Theorem 2.2 due to [3, 4], the strong local subdifferentiability required in Assumption 3.5 is equivalent to the strong metric subregularity of $\partial[J_\delta + \alpha_\delta R]$, i.e., of $x \mapsto A^*(Ax - b^\delta) + \alpha_\delta \partial R(x)$ at $\hat{x}$ for $A^*(A\hat{x} - b_\delta) - \alpha_\delta A^*\hat{w} = A^*(\hat{b} - b_\delta - \alpha_\delta \hat{w})$.

Theorem 3.6. *Suppose Assumption 3.1 and the* strong *source condition of Assumption 3.5 hold at $\hat{x}$ for some $\hat{w}$. Suppose $(e_\delta + \delta^2)/\alpha_\delta \searrow 0$ and $\alpha_\delta \searrow 0$ as $\delta \searrow 0$. Then there exists $\bar{\delta} > 0$ such that if $\delta \in (0, \bar{\delta})$, we have*

$$\|x_\delta - \hat{x}\|_X^2 \le \frac{e_\delta}{\gamma\gamma_\delta} + \frac{\delta^2}{2\gamma^2\gamma_\delta} + \frac{\alpha_\delta^2}{2\gamma^2\gamma_\delta}\|\hat{w}\|_Y^2. \tag{3.8}$$

Proof. We have

$$J_\delta'(\hat{x}) - \alpha_\delta A^*\hat{w} = A^*(A\hat{x} - b_\delta - \alpha_\delta\hat{w}) = A^*(\hat{b} - b_\delta - \alpha_\delta\hat{w}).$$

By the assumption that $(e_\delta + \delta^2)/\alpha_\delta \searrow 0$ and $\alpha_\delta \searrow 0$ as $\delta \searrow 0$, Lemma 3.4, and (3.7) in Assumption 3.5, for suitably small $\delta > 0$, we have $x_\delta \in U$. Hence by Assumptions 3.1 and 3.5 followed by Young's inequality (1.5),

$$\begin{aligned} e_\delta &\ge [J_\delta + \alpha_\delta R](x_\delta) - [J_\delta + \alpha_\delta R](\hat{x}) \\ &\ge \langle J_\delta'(\hat{x}) - \alpha_\delta A^*\hat{w} | x_\delta - \hat{x}\rangle + \gamma\|x_\delta - \hat{x}\|_\delta^2 \\ &= \langle \hat{b} - b_\delta - \alpha_\delta\hat{w}, A(x_\delta - \hat{x})\rangle + \gamma\|A(x_\delta - \hat{x})\|_Y^2 + \gamma\gamma_\delta\|x_\delta - \hat{x}\|_X^2 \\ &\ge -\frac{1}{4\gamma}\|\hat{b} - b_\delta - \alpha_\delta\hat{w}\|_Y^2 + \gamma\gamma_\delta\|x_\delta - \hat{x}\|_X^2. \end{aligned} \tag{3.9}$$

Thus, again using Young's inequality and $\|\hat{b} - b_\delta\| \le \delta$ from Assumption 3.1, we obtain

$$\|x_\delta - \hat{x}\|_X^2 \le \frac{e_\delta}{\gamma\gamma_\delta} + \frac{1}{4\gamma^2\gamma_\delta}\|\hat{b} - b_\delta - \alpha_\delta\hat{w}\|_Y^2 \le \frac{e_\delta}{\gamma\gamma_\delta} + \frac{\delta^2}{2\gamma^2\gamma_\delta} + \frac{\alpha_\delta^2}{2\gamma^2\gamma_\delta}\|\hat{w}\|_Y^2. \tag{3.10}$$

This is the claim. □

If $\gamma_\delta \propto \alpha_\delta$, the following corollary shows norm convergence under similar parameter choices as in Theorem 3.2.

Corollary 3.7. *Suppose Assumption 3.1 and the* strong *source condition of Assumption 3.5 hold at $\hat{x}$. If*

$$\lim_{\delta \searrow 0} \frac{1}{\min\{\alpha_\delta, \gamma_\delta\}}(\alpha_\delta^2, \delta^2, e_\delta) = 0, \tag{3.11}$$

then

$$\lim_{\delta \searrow 0} \|x_\delta - \hat{x}\|_X = 0.$$

3.4 ESTIMATES BASED ON A SEMI-STRONG SOURCE CONDITION

We now replace the assumption of strong metric subregularity, i.e., strong local subdifferentiability, with mere (A, γ_δ)-strong local subdifferentiability.

Assumption 3.8 (Semi-strong source condition). We assume that $\hat{x} \in X$ satisfies for some $\hat{w} \in Y$ the basic source condition (3.3). Moreover, for all $\delta > 0$, for given $\alpha_\delta, \gamma_\delta > 0$, assume that $J_\delta + \alpha_\delta R$ is (A, γ_δ)-strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta A^* \hat{w}$ with respect to $\hat{X}$. The factor $\gamma > 0$ and neighbourhood $U = U^{\hat{x}}$ of (A, γ_δ)-strong local subdifferentiability must be independent of $\delta > 0$. Then we say that $\hat{x}$ satisfies for $\hat{w}$ the *semi-strong source condition.*

It is not that easy to very the semi-strong source condition with $U^{\hat{x}} \supset U_\rho$ as in Theorem 3.6 without assumptions that would verify strong source condition of Assumption 3.5. We therefore drop this assumption for the next lemma at the cost of weaker results in the ensuing theorem.

Lemma 3.9. *Suppose Assumption 3.1 and the* semi-strong *source condition of Assumption 3.8 hold at some $\hat{x} \in X$ for some $\hat{w}$. If $x_\delta \in U^{\hat{x}}$, then*

$$\operatorname{dist}^2(x_\delta, \hat{X}) \le \frac{e_\delta}{\gamma\gamma_\delta} + \frac{\delta^2}{2\gamma^2\gamma_\delta} + \frac{\alpha_\delta^2}{2\gamma^2\gamma_\delta}\|\hat{w}\|_Y^2.$$

Proof. As in (3.9), using the assumed (A, γ_δ)-strong local subdifferentiability of $J_\delta + \alpha_\delta R$, we estimate

$$\begin{aligned} e_\delta &\ge [J_\delta + \alpha_\delta R](x_\delta) - [J_\delta + \alpha_\delta R](\hat{x}) \\ &\ge \langle J_\delta'(\hat{x}) - \alpha_\delta A^* \hat{w} | x_\delta - \hat{x}\rangle + \gamma\|A(x_\delta - \hat{x})\|_Y^2 + \gamma\gamma_\delta \operatorname{dist}^2(x_\delta, \hat{X}) \\ &= \langle \hat{b} - b_\delta - \alpha_\delta \hat{w}, A(x_\delta - \hat{x})\rangle + \gamma\|A(x_\delta - \hat{x})\|_Y^2 + \gamma\gamma_\delta \operatorname{dist}^2(x_\delta, \hat{X}) \\ &\ge -\frac{1}{4\gamma}\|\hat{b} - b_\delta - \alpha_\delta \hat{w}\|_Y^2 + \gamma \operatorname{dist}^2(x_\delta, \hat{X}). \end{aligned}$$

Now estimating as in (3.10) yields the claim. □

Theorem 3.10. *Suppose Assumption 3.1 holds and that there exists a collection $\tilde{X} \subset \hat{X}$ of points satisfying the* semi-strong *source condition of Assumption 3.8 with $\bigcup_{\hat{x}\in\tilde{X}} U^{\hat{x}} \supset \hat{X}$. Also suppose that U_ρ defined in (3.7) is weakly or weakly-$*$ compact for some $\rho > 0$, and each $U^{\hat{x}}$ for all $\hat{x} \in \tilde{X}$ correspondingly weakly or weakly-$*$ open. If*

$$\lim_{\delta \searrow 0} \frac{1}{\min\{\alpha_\delta, \gamma_\delta\}}(\alpha_\delta^2, \delta^2, e_\delta) = 0, \tag{3.12}$$

then

$$\lim_{\delta \searrow 0} \operatorname{dist}(x_\delta, \hat{X}) = 0.$$

Proof. Suppose, to reach a contradiction, for a sequence $\delta_k \searrow 0$ that $\inf_k \operatorname{dist}(x_{\delta_k}, \hat{X}) > 0$. By Lemma 3.4 and (3.12), we have $x_{\delta_k} \in U_\rho$ for large enough k. Since U_ρ is weakly(-$*$) compact, we can extract a subsequence, unrelabelled, such that also $x_{\delta_k} \rightharpoonup \tilde{x}$ weakly(-$*$). Since $A\tilde{x} = \hat{b}$ for $\hat{x} \in \hat{X}$, we have using Assumption 3.1 and (3.12) that

$$\begin{aligned} \lim_{k\to\infty} \|Ax_{\delta_k} - b_{\delta_k}\|^2 &\le \lim_{k\to\infty} \left(\|Ax_{\delta_k} - b_{\delta_k}\|^2 + \alpha_k R(x_{\delta_k})\right) \\ &\le \lim_{k\to\infty} \left(\|A\hat{x} - b_{\delta_k}\|^2 + \alpha_k R(\hat{x}) + e_{\delta_k}\right) \\ &= \lim_{k\to\infty} \|\hat{b} - b_{\delta_k}\|^2 = \lim_{k\to\infty} \delta_k^2 = 0 \quad (\hat{x} \in \hat{X}). \end{aligned}$$

Using (3.4) and (3.12) yields $Ax_{\delta_k} \to \hat{b}$. Moreover, (3.4) and (3.5) in Assumption 3.1 give

$$R(x_{\delta_k}) \le \frac{\delta_k^2}{\alpha_{\delta_k}} + R(\hat{x}).$$

Since R is assumed convex and lower semicontinuous, it is weakly(-$*$) lower semicontinuous; see, e.g., [18, Corollary 2.2] or [14, Lemma 1.10]. Hence, using (3.12), we obtain $R(\tilde{x}) \leq R(\hat{x})$ and consequently $\tilde{x} \in \hat{X}$. By assumption, $\tilde{x} \in U^{\hat{x}}$ for some $\hat{x} \in \tilde{X}$ satisfying the semi-strong source condition of Assumption 3.8. By the weak(-$*$) openness of $U^{\hat{x}}$, it follows that $x_{\delta_k} \in U^{\hat{x}}$ for k large enough. Following the proof of Lemma 3.9 and using (3.12), we deduce that $\operatorname{dist}(x_{\delta_k}, \hat{X}) \to 0$. □

The following will be useful for verifying the strong metric subregularity or (A, γ_δ)-strong local subdifferentiability required in Assumptions 3.5 and 3.8. The proof of the first lemma follows from that of the second by taking $\hat{X} = \{\hat{x}\}$ and expanding $\|\cdot\|_\delta$.

Lemma 3.11. *Suppose A and R are as in Assumption 3.1. Then $J_\delta + \alpha_\delta R$ is strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta \hat{d}$ with respect to the norm $\|\cdot\|_\delta$ if, for corresponding neighbourhood $U \ni \hat{x}$ and $\gamma > 0$,*

$$\alpha_\delta[R(x) - R(\hat{x}) - \langle \hat{d} | x - \hat{x}\rangle] + \left(\frac{1}{2} - \gamma\right) \|A(x - \hat{x})\|_2^2 \geq \gamma\gamma_\delta \|x - \hat{x}\|^2 \quad (x \in U).$$

Lemma 3.12. *Suppose A and R are as in Assumption 3.1. Then $J_\delta + \alpha_\delta R$ is (A, γ_δ)-strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta \hat{d}$ with respect to $\hat{X}$ if, for corresponding neighbourhood $U \ni \hat{x}$ and $\gamma > 0$,*

$$\alpha_\delta[R(x) - R(\hat{x}) - \langle \hat{d} | x - \hat{x}\rangle] + \left(\frac{1}{2} - \gamma\right) \|A(x - \hat{x})\|_2^2 \geq \gamma\gamma_\delta \operatorname{dist}^2(x, \hat{X}) \quad (x \in U). \tag{3.13}$$

Proof. Expanding (2.3), we need to prove

$$\begin{aligned}\frac{1}{2}\|Ax - b_\delta\|_2^2 + \alpha_\delta R(x) - \frac{1}{2}\|A\hat{x} - b_\delta\|_2^2 - \alpha_\delta R(\hat{x}) \\ \geq \langle A^*(A\hat{x} - b_\delta) + \alpha_\delta \hat{d} | x - \hat{x}\rangle + \gamma\|A(x - \hat{x})\|^2 + \gamma\gamma_\delta \operatorname{dist}^2(x, \hat{X}) \quad (x \in U).\end{aligned}$$

Using the properties of the Hilbert space norm (Pythagoras' identity), this rearranges as (3.13). □

4 EXAMPLES

We now look at a few examples that demonstrate Theorem 3.6 and Corollary 3.7.

4.1 BASIC EXAMPLES

Taking $R(x) = \frac{1}{2}\|x\|_X^2$ and $e_\delta = 0$, we recover from Corollary 3.7 classical results on Tikhonov regularisation [19]:

Theorem 4.1 (Norm-squared regularisation). *Let X and Y be Hilbert spaces, $A \in \mathbb{L}(X; Y)$. Suppose $A\hat{x} = \hat{b}$ and $\hat{x} \in \operatorname{ran} A^*$, i.e. $\hat{x} = -A^*\hat{w}$ for some $\hat{w}$. For all $\delta > 0$, let $x_\delta \in X$ solve*

$$(A^*A + \alpha_\delta)x_\delta + A^*b_\delta = 0.$$

Then $\|x_\delta - \hat{x}\|_X^2 \leq \frac{\delta^2}{2\alpha_\delta} + \frac{\alpha_\delta}{2}\|\hat{w}\|_Y^2$. In particular $\|x_\delta - \hat{x}\| \searrow 0$ as $\delta \searrow 0$ provided we choose the regularisation parameter $\alpha_\delta \searrow 0$ such that $\delta^2/\alpha_\delta \searrow 0$.

Proof. Since $\partial R(x) = \{x\}$, the basic source condition (3.3) holds due to $\hat{x} \in \operatorname{ran} A^*$. Since R is strongly convex with parameter $\gamma = 1$, the strong strong local subdifferentiability required by Assumption 3.5 holds with $\gamma_\delta = \alpha_\delta$, $\gamma = 1$, and $U = X$. Consequently, taking accuracy $e_\delta \equiv 0$, Theorem 3.6 and Corollary 3.7 yield the claims. □

We can also add constraints. Denoting by $N_{[0,\infty)^\Omega}(\hat{x})$ the normal cone to $[0,\infty)^\Omega$ in $L^2(\Omega)$, in the next result we take $R(x) = \frac{1}{2}\|x\|_X^2 + \delta_{[0,\infty)^\Omega}(x)$ in $X = L^2(\Omega)$.

Theorem 4.2 (Norm-squared regularisation with non-negativity constraints). *Let $X = L^2(\Omega)$ for some $\Omega \subset \mathbb{R}^d$, and let Y be a Hilbert space, $A \in \mathbb{L}(X;Y)$. Suppose $A\hat{x} = \hat{b}$ and $\hat{x}$ satisfies the source condition $\hat{x} \in \operatorname{ran} A^* - N_{[0,\infty)^\Omega}(\hat{x})$, i.e. $\hat{x} \in -A^*\hat{w} - N_{[0,\infty)^\Omega}(\hat{x})$ for some $\hat{w}$. For all $\delta > 0$, let*

$$x_\delta \in \operatorname*{arg\,min}_{0 \le x \in L^2(\Omega)} \frac{1}{2}\|Ax - b_\delta\|_Y^2 + \frac{\alpha_\delta}{2}\|x\|_X^2.$$

Then $\|x_\delta - \hat{x}\|_X^2 \le \frac{\delta^2}{2\alpha_\delta} + \frac{\alpha_\delta}{2}\|\hat{w}\|_Y^2$ In particular $\|x_\delta - \hat{x}\| \searrow 0$ as $\delta \searrow 0$ provided we choose the regularisation parameter $\alpha_\delta \searrow 0$ such that $\delta^2/\alpha_\delta \searrow 0$.

Proof. Since $\partial R(x) = x + N_{[0,\infty)^\Omega}(\hat{x})$, the basic source condition (3.3) holds due to $\hat{x} \in -A^*\hat{w} - N_{[0,\infty)^\Omega}(\hat{x})$. Since R is strongly convex with parameter $\gamma = 1$, the strong strong local subdifferentiability required by Assumption 3.5 holds with $\gamma_\delta = \alpha_\delta$, $\gamma = 1$, and $U = X$. Consequently, taking accuracy $e_\delta \equiv 0$, Theorem 3.6 and Corollary 3.7 yield the claims. □

We next look at nonsmooth regularisation, first in finite dimensions and then in infinite dimensions.

4.2 ℓ^1-REGULARISED REGRESSION

We now take $R(x) = \|x\|_1$ in $\mathbb{R}^n$ with $A \in \mathbb{R}^{m\times n}$. The basic source condition (3.3) then holds if there exist $\hat{x}, \hat{d} \in \mathbb{R}^n$ such that

$$A\hat{x} = \hat{b} \quad \text{and} \quad \hat{d} \in \operatorname{ran} A^* \cap \operatorname{Sign} \hat{x}, \tag{4.1}$$

where we recall for $x \in \mathbb{R}^n$, with Π denoting a cartesian product of sets, that

$$\partial\|\cdot\|_1(x) = \operatorname{Sign} x := \prod_{k=1}^n \begin{cases} \{-1\}, & x_k < 0, \\ \{1\}, & x_k > 0, \\ [-1,1], & x_k = 0. \end{cases}$$

We also write $\operatorname{sign} x$ for the (scalar) sign of $x \neq 0$. As we recall, (4.1) just means that $\hat{x}$ solves $\min_{Ax=\hat{b}} \|x\|_1$. Then $\hat{X}$ is the set of these minimisers. Otherwise said, (4.1) holds if $\hat{x}$ is an admissible ground-truth ($A\hat{x} = \hat{b}$) and $\operatorname{ran} A^* \cap \operatorname{Sign} \hat{x} \neq \emptyset$.

With $\hat{d} \in \operatorname{Sign} \hat{x}$, we write

$$Z(\hat{x}, \hat{d}) := \{k \in \{1,\dots,n\} \mid \hat{x}_k = 0,\ |\hat{d}_k| < 1\}$$

and call $(\hat{x}, \hat{d})$ *strictly complementary* if $Z(\hat{x}, \hat{d}) = \{k \in \{1,\dots,n\} \mid \hat{x}_k = 0\}$. That is, $(\hat{x}, \hat{d})$ are strictly complementary if $\hat{x}_k = 0$ implies that $\hat{d}_k$ is not on the boundary of $[-1,1]$.

The next result is extracted from [40, proof of Theorem 2].

Lemma 4.3. *Suppose $\hat{x}$ and $\hat{d}$ satisfy the source condition (4.1). Then there exists $\bar{d} \in \mathbb{R}^n$ such that $(\bar{x}, \bar{d})$ is strictly complementary and satisfies (4.1).*

Proof. Let $r \in \mathbb{R}^n$ be defined by

$$r_k := \begin{cases} -1, & \hat{x}_k = 0,\ \hat{d}_k = 1, \\ 1, & \hat{x}_k = 0,\ \hat{d}_k = -1, \\ 0, & \text{otherwise}. \end{cases}$$

We try to replace $(\hat{x}, \hat{d})$ by a strictly complementary solution by considering for $\alpha > 0$, $t := \|\hat{x}\|_1$ and $B_1 \subset \mathbb{R}^n$ the 1-norm unit ball the problem

$$\min_{x \in \mathbb{R}^n} J(x) = \big(\alpha\langle r, x\rangle + \delta_{tB_1}(x)\big) + \delta_{\{\hat{b}\}}(Ax).$$

Clearly $x = \hat{x}$ is a solution to this problem and $J(\hat{x}) = 0$. Moreover $A\hat{x} = \hat{b} \in \operatorname{ri}\{\hat{b}\} = \{\hat{b}\}$, where ri denotes the relative interior. Therefore, the Fenchel–Rockafellar dual problem (see, e.g., [18, 14]) is

$$\min_{w \in \mathbb{R}^m} Q(w) := t\|A^* w + \alpha r\|_\infty + \langle \hat{b}, w\rangle,$$

and a minimiser $\bar{w}$ satisfies $Q(\bar{w}) = -J(\hat{x}) = 0$. In other words $t\|A^*\bar{w} + \alpha r\|_\infty = -\langle \hat{b}, \bar{w}\rangle$.

If $A^*\bar{w} = -\alpha r$, then for small enough $\alpha > 0$ we have

$$\bar{d} := \hat{d} - A^*\bar{w} = \hat{d} + \alpha r \in \operatorname{Sign}\hat{x} \cap \operatorname{ran} A^*$$

and $(\hat{x}, \bar{d})$ is strictly complementary. Thus it fulfills our claims.

If $A^*\bar{w} \neq -\alpha r$, let $\tilde{w} := -\bar{w}/s$ for $s := \|A^*\bar{w} + \alpha r\|_\infty = -\langle \hat{b}, \bar{w}\rangle/t$. Then

$$\langle \hat{x}, A^*\tilde{w}\rangle = \langle A\hat{x}, \tilde{w}\rangle = \langle \hat{b}, \tilde{w}\rangle = -s^{-1}\langle \hat{b}, \bar{w}\rangle = t = \|\hat{x}\|_1 \tag{4.2}$$

and

$$\|A^*\tilde{w} - \frac{\alpha}{s} r\|_\infty \leq 1. \tag{4.3}$$

Let $k \in \{1, \dots, n\}$. If $\hat{x}_k \neq 0$ or $k \in Z(\hat{x}, \hat{d})$, (4.3) implies by the definition of r_k that $[A^*\hat{w}]_k \in [-1, 1]$. Due to (4.2), we must therefore have that $[A^*\hat{w}]_k = \hat{d}_k = \operatorname{sign}\hat{x}_k$ when $\hat{x}_k \neq 0$. If, on the other hand $\hat{x}_k = 0$ and $k \notin Z(\hat{x}, \hat{d})$, then (4.3) shows that

$$-1 - \frac{\alpha}{s}\operatorname{sign}\hat{d}_k \leq A^*\tilde{w} \leq 1 - \frac{\alpha}{s}\operatorname{sign}\hat{d}_k.$$

If $\hat{d}_k > 0$, this guarantees $-2 < [A^*\tilde{w}]_k < 1$. If $\hat{d}_k < 0$, this guarantess $-1 < [A^*\tilde{w}]_k < 2$. Consequently, for small enough $\alpha > 0$,

$$\bar{d} := \frac{1}{2}(\hat{d} + A^*\tilde{w}) \in \operatorname{Sign}\hat{x} \cap \operatorname{ran} A^*,$$

and $(\hat{x}, \bar{d})$ is strictly complementary. Indeed, if $\hat{d}_k \in (-1, 1)$, then still $\bar{d}_k \in (-1, 1)$ due to $[A^*\tilde{w}]_k \in [-1, 1]$. On the other hand, if $\hat{d}_k \in \{-1, 1\}$ and $\hat{x} \neq 0$, we have just proved that $\hat{d}_k + [A^*\tilde{w}]_k \in (-1, 1)$. Finally, if $\hat{x}_k \neq 0$, we have proved above that $[A^*\hat{w}]_k = \hat{d}_k \in \operatorname{sign}\hat{x}_k$. Thus $(\hat{x}, \bar{d})$ it fulfills our claims. □

The next lemma shows that local subdifferentiability is requisite for the basic source condition (4.1) to hold at a ground-truth $\hat{x}$ admitting a strictly complementary dual variable $\hat{d}$..

Lemma 4.4. *Suppose $(\hat{x}, \hat{d})$ is strictly complementary and satisfies the basic source condition (4.1). Let $F_\delta(x) := \frac{1}{2}\|Ax - b_\delta\|_2^2 + \alpha_\delta\|x\|_1$ on $\mathbb{R}^n$ for some $\alpha_\delta \in (0, 1/2)$ and $b_\delta \in \mathbb{R}^m$. Then ∂F_δ is (A, α_δ)-strongly locally subdifferentiable at $\hat{x}$ for $\hat{x}^* := A^*(A\hat{x} - b_\delta) + \alpha_\delta\hat{d} \in \partial F_\delta(\hat{x})$ in some open neighbourhood $U = U^{\hat{x}}$ of $\hat{x}$. The factor γ of (A, α_δ)-strong local subdifferentiability is independent of δ, as is U.*

Proof. By Lemma 3.12 we need to prove

$$\alpha_\delta[\|x\|_1 - \|\hat{x}\|_1 - \langle \hat{d}, x - \hat{x}\rangle] + \left(\frac{1}{2} - \gamma\right)\|A(x - \hat{x})\|_2^2 \geq \gamma\gamma_\delta \operatorname{dist}^2(x, \hat{X}) \quad (x \in U). \tag{4.4}$$

Let

$$M := A^*A + \sum_{k \in Z(\hat{x},\hat{d})} \mathbb{1}_k \mathbb{1}_k^\top, \tag{4.5}$$

where $\mathbb{1} := (1, \dots, 1)$. Suppose $z \in \ker M$. Then $z \in \ker A$ and $z_k = 0$ for all $k \in Z(\hat{x}, \hat{d})$. We will show that also $\sum_{k \notin Z(\hat{x},\hat{d})} z_k = 0$ and $\hat{x}+z \in \hat{X}$ if $\|z\|$ is small enough. Indeed, suppose, to reach a contradiction, that this does not hold. Let $\varepsilon > 0$. We may assume that $\|z\| \le \varepsilon$. We have $A(\hat{x} + z) = A\hat{x} = \hat{b}$ and, for small enough $\varepsilon > 0$,

$$\|\hat{x} + z\|_1 = \sum_{k \notin Z(\hat{x},\hat{d})} |\hat{x}_k + z_k| = \sum_{k:\hat{x}_k \neq 0} |\hat{x}_k + z_k| = \sum_{k:\hat{x}_k \neq 0} (|\hat{x}_k| + z_k) = \|\hat{x}\|_1 + \sum_{k \notin Z(\hat{x},\hat{d})} z_k.$$

If now $\sum_{k \notin Z(\hat{x},\hat{d})} z_k < 0$, this contradicts $\hat{x}$ being an $\|\cdot\|_1$-minimising solution to $A\hat{x} = \hat{b}$. Since also $-z \in \ker A$, we must therefore have $\sum_{k \notin Z(\hat{x},\hat{d})} z_k = 0$. Moreover, $\hat{x} + z \in \hat{X}$.

For now, let $U = U^{\hat{x}}$ be an arbitrary bounded neighbourhood of $\hat{x}$ and $\rho = \sup_{x \in U} \|x_k\|_\infty$. Then with $\beta_0 := \rho^{-1} \inf_{k \in Z(\hat{x},\hat{d})} (1 - |\hat{d}_k|) > 0$ we have

$$\begin{aligned} \|x\|_1 - \|\hat{x}\|_1 - \langle \hat{d}, x - \hat{x} \rangle &\ge \sum_{k \in Z(\hat{x},\hat{d})} (|x| - |\hat{x}| - \hat{d}_k(x_k - \hat{x}_k)) = \sum_{k \in Z(\hat{x},\hat{d})} (|x| - \hat{d}_k(x_k)) \\ &\ge \sum_{k \in Z(\hat{x},\hat{d})} (1 - |\hat{d}|)|x_k| \ge \rho\beta_0 \sum_{k \in Z(\hat{x},\hat{d})} |x_k| \ge \beta_0 \sum_{k \in Z(\hat{x},\hat{d})} |x_k|^2 \quad (x \in U). \end{aligned} \tag{4.6}$$

Since $\alpha_\delta \in (0, 1/2)$, we may find $\gamma \in (0, 1/2)$ with $\alpha_\delta \le \frac{1}{2} - \gamma$. By (4.6), for some $\beta := \min\{1, \beta_0\}$ we have

$$\begin{aligned} \alpha_\delta[\|x\|_1 - \|\hat{x}\|_1 - \langle \hat{d}, x - \hat{x} \rangle] + \left(\frac{1}{2} - \gamma\right) \|A(x - \hat{x})\|_2^2 \\ \ge \alpha_\delta\beta \sum_{k \in Z(\hat{x},\hat{d})} |x_k|^2 + \left(\frac{1}{2} - \gamma\right) \|A(x - \hat{x})\|_2^2 \ge \alpha_\delta\beta\langle M(x - \hat{x}), x - \hat{x} \rangle. \end{aligned} \tag{4.7}$$

With $x \in U$, write $x - \hat{x} = z + w$ where $z \in \ker M$ and $w \perp \ker M$. That is, $x - \bar{x} = w$ for $\bar{x} = \hat{x} + z$. By the previous paragraph, $\bar{x} \in \hat{X}$ if $\|z\|$ is small enough, i.e., the neighbourhood U of $\hat{x}$ is small enough and $x \in U$. Thus, for $\lambda_{\min} > 0$ the minimal non-zero eigenvalue of M,

$$\alpha_\delta\beta\langle M(x - \hat{x}), x - \hat{x} \rangle = \alpha_\delta\beta\langle Mw, w \rangle \ge \alpha_\delta\beta\lambda_{\min}\|w\|_X^2 = \alpha_\delta\beta\lambda_{\min}\|x - \bar{x}\|_X^2.$$

This with (4.7) proves (4.4), hence the claim for $\gamma_\delta = \alpha_\delta$ and any $0 < \gamma < \min\{1/2, \beta\lambda_{\min}\}$. □

Finally, we may show that asymptotically accurate and sufficiently regularised solutions to ℓ^1-regularised regression otherwise unconditionally converge to the *set* of 1-norm-minimising ground-truths. Based on the practical implementation of the procedure, ℓ^1-regularised regression is also known in the literature as "Lasso" for "least absolute shrinkage and selection operator".

Theorem 4.5 (Lasso). *Let $R(x) = \|x\|_1$ and and $A \in \mathbb{R}^{m \times n}$ with $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. Suppose Assumption 3.1 holds and that the accuracy e^δ and regularisation parameter $\alpha_\delta > 0$, moreover, satisfy*

$$\lim_{\delta \searrow 0} \left(\alpha_\delta, \frac{\delta^2}{\alpha_\delta}, \frac{e_\delta}{\alpha_\delta}\right) = 0.$$

Then $\operatorname{dist}(x_\delta, \hat{X}) \to 0$ *as* $\delta \searrow 0$.

Proof. We verify the conditions Theorem 3.10. The function $J(x) := \delta_{\{x|Ax=\hat{b}\}}(x) + \|x\|_1$ is coercive, lower semicontinuous, and bounded from below. Therefore the problem $\min J$ has a solution $\hat{x} \in \hat{X}$. As we have already discussed, these solutions are characterised by the source condition (4.1) for some $\hat{d}$. Thus by Lemma 4.3, for every $\hat{x} \in \hat{X}$ there exists $\hat{d}$ such that $(\hat{x}, \hat{d})$ is strictly complementary and satisfies (4.1). Lemma 4.4 shows for all $\hat{x} \in \hat{X}$ and all $\delta > 0$ such that $\alpha_\delta \in (0, 1/2)$ that $\partial[J_\delta + \alpha_\delta R]$ is (A, α_δ)-strongly locally subdifferentiable at $\hat{x}$ for $\hat{x}^*_\delta := A^*(A\hat{x} - b_\delta) + \alpha_\delta \hat{d}$ with the parameter $\gamma > 0$ independent of $\delta > 0$. Consequently the semi-strong source condition of Assumption 3.8 holds for all $\hat{x} \in \hat{X}$ in some neighbourhood $U^{\hat{x}} \ni \hat{x}$. Clearly $\bigcup_{\hat{x}\in\hat{X}} U^{\hat{x}} \supset \hat{X}$. Moreover, U_ρ is compact by the lower semicontinuity and coercivity of R and the continuity of A, and $U^{\hat{x}}$ is open. By the finite-dimensionality of X the respective weak compactness and weak openness assumptions of Theorem 3.10 hold. The rest follows from Theorem 3.10 and $\gamma_\delta = \alpha_\delta$. □

Remark 4.6. If M defined in (4.5) is positive definite, then the strong source condition of Assumption 3.5 holds at $\hat{x}$. In that case we may apply Theorem 3.6 to obtain $\|x_\delta - \hat{x}\| \to 0$ and the estimates (3.8).

4.3 TOTAL VARIATION REGULARISED IMAGE RECONSTRUCTION

Suppose, as before, that $\Omega \subset \mathbb{R}^n$. On the space $X = L^2(\Omega)$, define the *total variation* regulariser

$$R(x) = \begin{cases} \|Dx\|_{\mathcal{M}(\Omega)}, & \mathrm{BV}(\Omega) \cap L^2(\Omega), \\ \infty, & \text{otherwise.} \end{cases}$$

Here $D \in \mathbb{L}(\mathrm{BV}(\Omega); \mathcal{M}(\Omega))$ is the distributional differential, mapping functions of bounded variation on a domain $\Omega \subset \mathbb{R}^m$ to Radon measures. The Radon norm $\|Dx\|_{\mathcal{M}(\Omega)} := \int_\Omega d|Dx|$, where $|Dx|$ is the total variation measure of Dx. For details on functions of bounded variation we refer to [1]. For brevity we write $\|\cdot\|_2 = \|\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{\mathcal{M}} = \|\cdot\|_{\mathcal{M}(\Omega;\mathbb{R}^m)}$.

Any $\hat{d} \in L^2(\Omega) \cap \partial\|D\cdot\|_{\mathcal{M}(\Omega;\mathbb{R}^m)}(\hat{x})$ satisfies by [2] or [30, Proposition 5 & Lemma 3 & Definition 11] for some $\hat{\varphi} \in L^\infty(\Omega; \mathbb{R}^m)$ that

$$\hat{d} = -\operatorname{div}\hat{\varphi}, \quad \|\hat{\varphi}\|_\infty \le 1, \quad \hat{\varphi} D\hat{x} = |D\hat{x}|. \tag{4.8a}$$

Hence the basic source condition (3.3) reduces to (4.8a) with

$$A\hat{x} = \hat{b} \quad \text{and} \quad \hat{d} \in \operatorname{ran} A^*. \tag{4.8b}$$

The set $\hat{X}$ is given by $\hat{x}$ satisfying (4.8) for some $\hat{d}$.

Remark 4.7. It may often seem from the literature that *all* $x^* \in \partial\|D\cdot\|_{\mathcal{M}}(x)$ would have the form $x^* = -\operatorname{div}\varphi$ from some $\varphi \in L^\infty(\Omega; \mathbb{R}^m)$. This is not the case. Consider, for example, the step function $\hat{x} = \chi_{[0,\infty)}$ on $\mathbb{R}$. Then $D\hat{x} = \delta_{\{0\}}$. Take $\hat{x}^* \in \mathrm{BV}(\mathbb{R})^*$ given by $\hat{x}^*(x) = D^j x(\mathbb{R})$ for all $x \in \mathrm{BV}(\mathbb{R})$, i.e., the measure of the jump part of the differential of x. Then $\hat{x}^*(\hat{x}) = \|D\hat{x}\|_{\mathcal{M}}$ and $\|Dx\|_{\mathcal{M}} \ge \hat{x}^*(x)$ for all $x \in \mathrm{BV}(\mathbb{R})$. Thus

$$\|Dx\|_{\mathcal{M}} - \|D\hat{x}\|_{\mathcal{M}} \ge \hat{x}^*(x - \hat{x})$$

meaning that $\hat{x}^* \in \partial\|D\cdot\|_{\mathcal{M}}(\hat{x})$. However, if we had $\hat{x}^* = -\operatorname{div}\varphi$, then we would have $\hat{x}^*(x) = Dx(\varphi)$ contradicting the definition of $\hat{x}^*$ for suitable $x \in \mathrm{BV}(\mathbb{R})$. It is thus important in our overall example that *we are actually working in* $L^2(\Omega)$: $\operatorname{ran} A^* \subset L^2(\Omega)$. This allows us to limit our attention to subderivatives $x^* \in L^2(\Omega)$.

Let $\mathcal{L}$ denote the Lebesgue measure on $\mathbb{R}^m$. For a collection $\mathcal{O}$ of disjoint measurable $\Theta \subset \Omega$ with $\mathcal{L}(\Theta) > 0$, we define the centring operator $K_{\mathcal{O}} : L^2(\Omega) \to L^2(\bigcup_{\Theta\in\mathcal{O}} \Theta)$ by

$$(K_{\mathcal{O}} x)|\Theta := x|\Theta - \frac{1}{\mathcal{L}(\Theta)} \int_\Theta x \, d\mathcal{L} \quad (\Theta \in \mathcal{O}).$$

Here $x|\Theta$ denotes the restriction of x on the subdomain Θ. For brevity, we call the possibly empty collection $\mathcal{O}$ of disjoint subsets $\Theta \subset \Omega$ a *collection of flat areas* for x" if each $\Theta \in \mathcal{O}$ has a Lipschitz boundary and is such that $|Dx|(\Theta) = 0$ (i.e., x is a.e. constant in Θ). If, moreover, (4.8a) holds, and $\sup_{\xi\in\Theta} |\hat{\varphi}(\xi)| < 1$ for all $\Theta \in \mathcal{O}$,

we call $\mathcal{O}$ a *collection of strictly flat areas* for $\hat{x}$ at $\hat{\varphi}$.

Note that $K_{\mathcal{O}}x = 0$ if $\mathcal{O}$ is a collection of flat areas for x.

Remark 4.8 (Stair-casing). Image reconstructions based on Tikhonov-style total variation regularisation commonly exhibit large flat areas or "stair-casing" [31, 7]. The *strictness* of the collection of flat areas $\mathcal{O}$ can be related to strict complementarity conditions in optimisation.

We start with a technical lemma.

Lemma 4.9. *Let $\hat{x}, \hat{d} \in L^2(\Omega)$ and $\hat{\varphi} \in L^\infty(\Omega; \mathbb{R}^m)$ satisfy (4.8a). Let $\mathcal{O}$ be a finite collection of strictly flat areas for $\hat{x}$ at $\hat{\varphi}$ and suppose the neighbourhood $U \ni \hat{x}$ satisfies for some $C > 0$ that*

$$U \subset \{x \in L^2(\Omega) \mid \|Dx\|_{\mathcal{M}} \le C\}. \tag{4.9}$$

Then, for some constant $\varepsilon = \varepsilon(\mathcal{O}, \hat{d}) > 0$,

$$\|Dx\|_{\mathcal{M}} - \|D\hat{x}\|_{\mathcal{M}} - \langle \hat{d}|x - \hat{x}\rangle \ge \frac{\varepsilon}{C}\|K_{\mathcal{O}}(x - \hat{x})\|^2_{L^2(\bigcup_{\Theta\in\mathcal{O}}\Theta)} \quad (x \in U).$$

Proof. Observing (4.8a), we have

$$\begin{aligned} \|Dx\|_{\mathcal{M}} - \|D\hat{x}\|_{\mathcal{M}} - \langle \hat{d}|x - \hat{x}\rangle &= \int_\Omega d|Dx| - \int_\Omega d|D\hat{x}| - \int_\Omega \hat{\varphi} d(Dx - D\hat{x}) \\ &\ge \int_\Omega 1 - |\hat{\varphi}|\, d|Dx|. \end{aligned} \tag{4.10}$$

We recall that Poincaré's inequality [1, Remark 3.50] establishes for some constants $C_\Theta > 0$ that

$$\|K_{\{\Theta\}}x\|_{L^2(\Theta)} \le C_\Theta \|Dx\|_{\mathcal{M}(\Theta)} \quad (\Theta \in \mathcal{O},\ x \in L^2(\Omega)).$$

Since, by assumption, $|D\hat{x}|(\Theta) = 0$ and $1 - |\hat{\varphi}| \ge \varepsilon_{\Theta,\hat{\varphi}} > 0$ for some $\varepsilon_{\Theta,\hat{\varphi}} > 0$ for every $\Theta \in \mathcal{O}$, we therefore obtain

$$\int_\Omega 1 - |\hat{\varphi}|\, d|Dx| \ge \sum_{\Theta\in\mathcal{O}} \varepsilon_{\Theta,\hat{\varphi}} \int_\Theta d|Dx| \ge \sum_{\Theta\in\mathcal{O}} C_\Theta^{-1}\varepsilon_{\Theta,\hat{\varphi}}\|K_{\{\Theta\}}x\|_{L^2(\Theta)}. \tag{4.11}$$

(In particular, if $m = 2$ and $\Theta = B(\xi, \rho)$, we have $C_\Theta = C$ independent of ξ and ρ.) On the other hand, due to (4.9), Poincaré's inequality, and $K_{\{\Theta\}}\hat{x} = 0$, we also have

$$C \ge \|Dx\|_{\mathcal{M}(\Omega;\mathbb{R}^m)} \ge \|Dx\|_{\mathcal{M}(\Theta);\mathbb{R}^m)} \ge C_\Theta^{-1}\|K_{\{\Theta\}}x\|_{L^2(\Theta)} = C_\Theta^{-1}\|K_{\{\Theta\}}(x - \hat{x})\|_{L^2(\Theta)}.$$

Combining this estimate with (4.10) and (4.11) yields the claimed estimate with $\varepsilon = \inf_{\Theta\in\mathcal{O}} \varepsilon_{\Theta,\hat{\varphi}}/C_\Theta^2$. Since we assumed $\mathcal{O}$ to be finite, we have $\varepsilon > 0$. □

Remark 4.10. If the constant C_Θ from Poincaré's inequality is bounded over all $\Theta \in \mathcal{O}$, and we have $\sup_{\Theta\in\mathcal{O}\,\xi\in\Theta} |\hat{\varphi}(\xi)| < 1$, then the finiteness assumption on $\mathcal{O}$ can be dropped.

A discretised version of the next lemma on total variation can be found in [26, Appendix A]. It says that with regard to strong metric subregularity, the lack of positivity of A^*A can be compensated for by the strictly flat areas.

Lemma 4.11. *Let $F_\delta(x) := \frac{1}{2}\|Ax - b_\delta\|_2^2 + \alpha_\delta\|Dx\|_{\mathcal{M}}$ on $X = L^2(\Omega)$ with $\Omega \subset \mathbb{R}^n$ and $\alpha_\delta \in (0, 1/2)$. Suppose:*

(i) $\hat{x} \in L^2(\Omega) \cap \mathrm{BV}(\Omega)$, and $\hat{d} \in L^2(\Omega)$, and $\hat{\varphi} \in L^\infty(\Omega; \mathbb{R}^m)$ satisfy (4.8a).

(ii) There exists a a finite collection $\mathcal{O}$ of strictly flat areas for $\hat{x}$ such that

$$K_{\mathcal{O}}^* K_{\mathcal{O}} + A^* A \geq \tilde{\varepsilon}\,\mathrm{Id} \text{ for some } \tilde{\varepsilon} > 0. \tag{4.12}$$

(iii) $U \subset L^2(\Omega)$ satisfies (4.9) for some $C > 0$.

Then F is strongly locally subdifferentiable in U at $\hat{x}$ for $\hat{x}^* := A^*(A\hat{x} - b_\delta) + \alpha_\delta\hat{d} \in \partial F_\delta(\hat{x})$ with respect to the norm $\|\cdot\|_\delta$ with $\gamma_\delta = \alpha_\delta$. The factor κ (or γ) of strong metric subregularity is independent of δ.

Proof. By Lemma 3.11 we need to prove

$$L := \alpha_\delta\left(\|Dx\|_{\mathcal{M}} - \|D\hat{x}\|_{\mathcal{M}} - \langle\hat{d}|x - \hat{x}\rangle\right) + \left(\frac{1}{2} - \gamma\right)\|A(x - \hat{x})\|_2^2 \geq \gamma\gamma_\delta\|x - \hat{x}\|_X^2 \quad (x \in U). \tag{4.13}$$

Lemma 4.9 provides for some $\varepsilon = \varepsilon(\Theta, \hat{d})$ the estimate

$$\|Dx\|_{\mathcal{M}} - \|D\hat{x}\|_{\mathcal{M}} - \langle\hat{d}|x - \hat{x}\rangle \geq \frac{\varepsilon}{C}\|K_{\mathcal{O}}(x - \hat{x})\|_{L^2(\Theta)}^2.$$

This and (4.12) yield

$$L \geq \min\left\{\frac{\varepsilon\alpha_\delta}{C}, \frac{1}{2} - \gamma\right\}\tilde{\varepsilon}\|x - \hat{x}\|_2^2.$$

Since we assume that $\alpha_\delta \in (0, 1/2)$, taking $\gamma_\delta = \alpha_\delta$ and $\alpha_\delta \leq \frac{1}{2} - \gamma$ with $\gamma \in (0, \frac{1}{2})$, we now prove (4.13) for small enough $\gamma > 0$ independent of $\delta > 0$. □

Provided the strictly flat areas in the ground-truth compensate for the kernel of the forward operator, we can now show the convergence of total variation regularised approximate solutions:

Theorem 4.12 (Total variation regularised image reconstruction). *Let $R(x) = \|Dx\|_{\mathcal{M}(\Omega;\mathbb{R}^m)}$ in $X = L^2(\Omega)$ and suppose Assumption 3.1 holds. Also suppose $\hat{x}, \hat{d} \in L^2(\Omega)$ satisfy (4.8) and that there exists a corresponding collection $\mathcal{O}$ of strictly flat areas satisfying (4.12). Also suppose that the accuracy and regularisation parameters satisfy*

$$\lim_{\delta \searrow 0}\left(\alpha_\delta, \frac{\delta^2}{\alpha_\delta}, \frac{e_\delta}{\alpha_\delta}\right) = 0.$$

Then $\|x_\delta - \hat{x}\|_2 \searrow 0$. Moreover, (3.8) holds for small enough $\delta > 0$.

Proof. For some $\rho > 0$, let

$$U := U_\rho = \{x \in L^2(\Omega) \mid \|A(x - \hat{x})\| \leq \rho,\ \|Dx\|_{\mathcal{M}} \leq \|D\hat{x}\|_{\mathcal{M}} + \rho\}$$

Then (3.7) holds as does (4.9) with $C := \|D\hat{x}\|_{\mathcal{M}} + \rho$. Lemma 4.11 now shows that $J_\delta + \alpha_\delta R$ is strongly locally subdifferentiable at $\hat{x}$ for $\hat{x}_\delta^* := A^*(A\hat{x} - b_\delta) + \alpha_\delta\hat{d}$ with respect to the norm $\|\cdot\|_\delta$ with $\gamma_\delta = \alpha_\delta$ and with the parameter $\gamma > 0$ independent of $\delta > 0$. Since (4.8) verifies the basic source condition (3.3), this verifies the strong source condition of Assumption 3.5. Hence we may apply Theorem 3.6 and Corollary 3.7 to deduce the claim. □

5 NONLINEAR INVERSE PROBLEMS AND GENERAL DISCREPANCIES

We now consider the nonlinear inverse problem $A(\hat{x}) = \hat{b}$, which for corrupted data b_δ we solve via

$$\min_{x \in X} J_\delta(x) + \alpha_\delta R(x) \quad \text{where} \quad J_\delta(x) := E(A(x) - b_\delta)$$

for some convex, proper, lower semicontinuous regularisation functional $R : X \to \overline{\mathbb{R}}$ and a data fidelity $E : X \to \mathbb{R}$. We assume the forward operator $A \in C^1(X; Y)$ on the Banach spaces X and Y. We will shortly impose assumptions on E.

5.1 A BASIC SOURCE CONDITION AND GENERAL ASSUMPTIONS

We again take as our starting point for admissible ground-truths $\hat{x}$ those that minimise R, i.e., solve the problem

$$\min_{A(x)=\hat{b}} R(x). \tag{5.1}$$

We use the theory of Clarke subdifferentials on extended real-valued functions [13, Section 2.9] to write the necessary optimality conditions for $\hat{x}$ to solve this problem. Indeed, by the Fermat principle, $0 \in J(\hat{x})$ for $J(x) := R(x) + \delta_{\{x:A(x)=\hat{b}\}}$. Since $A \in C^1(X; Y)$ is regular in the sense of Clarke's theory, as is R as a convex function, the sum and composition rules [13, Theorems 2.9.8 & 2.9.9] hold as equalities for J. Thus the Fermat principle writes out as $\hat{x}$ having to satisfy for some $\hat{w} \in Y^*$ the *basic source condition*

$$A'(\hat{x}) = \hat{b} \quad \text{and} \quad A'(\hat{x})^* \hat{w} + \partial R(\hat{x}) \ni 0. \tag{5.2}$$

We write $\hat{X}$ for the set of $\hat{x}$ satisfying (5.2). This set may be larger than the set of minimisers of (5.1).

We now state our main assumption regarding accuracy and how E relates to the noise parameter $\delta > 0$. Essentially, the magnitude of E' has to be compatible with the parameterisation δ of the corruption level, and the convergence below zero of E has to be Hölderian as the corruption in the data vanishes.

Assumption 5.1 (Corruption level and solution accuracy). On Banach spaces X and Y, $A \in C^1(X; Y)$, and $R : X \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous, and $E : Y \to \mathbb{R}$ is convex and Fréchet differentiable. For given accuracy parameters $e_\delta \geq 0$ and all $\hat{x} \in \hat{X}$ we have

$$[J_\delta + \alpha_\delta R](x_\delta) - [J_\delta + \alpha_\delta R](\hat{x}) \leq e_\delta \quad (\delta > 0). \tag{5.3}$$

Moreover, the parametrisation $\delta > 0$ of the noise or corruption level is such that

$$\|E'(b_\delta - \hat{b})\|_{Y^*} \leq \delta.$$

and, for some $C', q > 0$,

$$E(\hat{b} - b_\delta) \leq C' \delta^q$$

Finally, for some $C, p > 0$ the function E satisfies the pseudo-Hölder estimate

$$C^{-1} E(z) \leq E(w) + \|E'(z - w)\|_Y^p \quad (z, w \in Y). \tag{5.4}$$

Example 5.2. Let $E(z) := \frac{1}{2}\|z\|_Y^2$ on a Hilbert space Y. Then $E'(z - w) = z - w$ so by Assumption 5.1 the noise level has to satisfy

$$\|b_\delta - \hat{b}\|_Y \leq \delta$$

and with this we can take $q = 2$ and $C' = 1/2$. Also (5.4) holds with $C = 3$ and $p = 2$. Indeed

$$\begin{aligned}\frac{1}{2C}\|z\|_Y^2 &= \frac{1}{2C}\|w - (z - w)\|_Y^2 \\ &\le \left(\frac{1}{2C} + \frac{C-1}{2C}\right)\|z\|_Y^2 + \left(\frac{1}{2C} + \frac{C}{2(C-1)}\right)\|z - w\|_Y^2 \le \frac{1}{2}\|w\|_Y^2 + \|z - w\|_Y^2.\end{aligned}$$

5.2 AN ESTIMATE BASED ON A STRONG SOURCE CONDITION

We modify the strong source condition of Assumption 3.5 for nonlinear A.

Assumption 5.3 (Strong source condition; nonlinear case). Assume that $\hat{x} \in X$ satisfies for some $\hat{w} \in Y$ the basic source condition (5.2). Moreover, for all $\delta > 0$, for given $\alpha_\delta, \gamma_\delta > 0$, assume that $J_\delta + \alpha_\delta R$ is strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta A'(\hat{x})^* \hat{w}$ with respect to the norm

$$\|x\|_\delta := \sqrt{\|A'(\hat{x})x\|_Y^2 + \gamma_\delta \|x\|_X^2} \quad (x \in X).$$

The factor $\gamma > 0$ of strong local subdifferentiability must be independent of $\delta > 0$ and, for some $\rho > 0$, we must have

$$U \supset U_\rho := \{x \in X \mid \|A(x) - A(\hat{x})\| \le \rho,\ R(x) \le R(\hat{x}) + \rho\}. \tag{5.5}$$

Then we say that $\hat{x}$ satisfies for $\hat{w}$ the *strong source condition.*

We again start with a simple bound:

Lemma 5.4. *Suppose Assumption 5.1 holds. Then*

$$E(A(x_\delta) - A(\hat{x})) \le C(e_\delta + C'\delta^q + \delta^p + \alpha_\delta R(\hat{x}))$$

and

$$R(x_\delta) \le R(\hat{x}) + \frac{e_\delta + C'\delta^q + \delta^p}{\alpha_\delta}.$$

Proof. By Assumption 5.1 and (5.3), since $A(\hat{x}) = \hat{b}$, we have

$$\begin{aligned}C^{-1}E(A(x_\delta) - A(\hat{x})) + \alpha_\delta R(x_\delta) &\le E(A(x_\delta) - b_\delta) + \alpha_\delta R(x_\delta) + \|E'(b_\delta - \hat{b})\|_Y^p \\ &\le E(\hat{b} - b_\delta) + \alpha_\delta R(\hat{x}) + e_\delta + \|E'(b_\delta - \hat{b})\|_Y^p \\ &\le e_\delta + C'\delta^q + \delta^p + \alpha_\delta R(\hat{x}).\end{aligned}$$

This finishes the proof. □

The next result generalises Theorem 3.6 to approximately linear operators A. The claim only differs by the factor η, which becomes the unit if A is linear.

Theorem 5.5. *Suppose Assumption 5.1 holds along with the strong source condition of Assumption 5.3 at some $\hat{x}$ for $\hat{w}$. Suppose $(e_\delta + \delta^q + \delta^p)/\alpha_\delta \searrow 0$ and $\alpha_\delta \searrow 0$ as $\delta \searrow 0$. Then there exists $\delta > 0$ such that if $\delta \in (0, \bar{\delta})$, we have*

$$\frac{1}{2}\|x_\delta - \hat{x}\|_X^2 \le \frac{e_\delta}{\gamma\gamma_\delta} + \frac{\delta^2}{2\gamma^2\gamma_\delta} + \frac{\alpha_\delta^2}{2\gamma^2\gamma_\delta}\|\hat{w}\|_Y^2. \tag{5.6}$$

Proof. Since $A(\hat{x}) = \hat{b}$, we have

$$J_\delta'(\hat{x}) - \alpha_\delta A'(\hat{x})^*\hat{w} = A'(\hat{x})^* E'(A(\hat{x}) - b_\delta) - \alpha_\delta A'(\hat{x})^*\hat{w} = A'(\hat{x})^*(E'(\hat{b} - b_\delta) - \alpha_\delta \hat{w}).$$

Let $\rho > 0$ be as in Assumption 5.1. By the assumption that $(e_\delta + \delta^q + \delta^p)/\alpha_\delta \searrow 0$ and $\alpha_\delta \searrow 0$ as $\delta \searrow 0$, Lemma 5.4, and (5.5), for suitably small $\delta > 0$, we have $x_\delta \in U_\rho \subset U$. Hence by the accuracy estimate in Assumption 5.1 and the strong local subdifferentiability included in Assumption 5.3,

$$\begin{aligned} e_\delta &\geq [J_\delta + \alpha_\delta R](x_\delta) - [J_\delta + \alpha_\delta R](\hat{x}) \\ &\geq \langle J_\delta'(\hat{x}) - \alpha_\delta A'(\hat{x})^* \hat{w} | x_\delta - \hat{x} \rangle + \gamma \|x_\delta - \hat{x}\|_\delta^2 \\ &= \langle E'(\hat{b} - b_\delta) - \alpha_\delta \hat{w}, A'(\hat{x})(x_\delta - \hat{x}) \rangle + \gamma \|A'(\hat{x})(x_\delta - \hat{x})\|_Y^2 + \gamma\gamma_\delta \|x_\delta - \hat{x}\|_X^2. \end{aligned} \tag{5.7}$$

Now, for any $\gamma' \in (0, \gamma)$, using Young's inequality twice, we obtain

$$e_\delta \geq -\frac{1}{4\gamma'} \|E'(\hat{b} - b_\delta)\|_Y^2 - \frac{\alpha_\delta^2}{4(\gamma - \gamma')} \|\hat{w}\|^2 + \gamma\gamma_\delta \|x_\delta - \hat{x}\|_X^2. \tag{5.8}$$

Thus

$$\|x_\delta - \hat{x}\|_X^2 \leq \frac{e_\delta}{\gamma\gamma_\delta} + \frac{\delta^2}{4\gamma'\gamma\gamma_\delta} + \frac{\alpha_\delta^2}{4(\gamma - \gamma')\gamma\gamma_\delta} \|\hat{w}\|_Y^2.$$

Taking $\gamma' = \frac{1}{2}\gamma$, this yields the claim. □

Corollary 5.6. *Suppose Assumption 5.1 holds along with the strong source condition of Assumption 5.3 at $\hat{x}$. If*

$$\lim_{\delta \searrow 0} \frac{1}{\min\{\alpha_\delta, \gamma_\delta\}} (\alpha_\delta^2, \delta^{\min\{2,q,p\}}, e_\delta) = 0$$

then

$$\lim_{\delta \searrow 0} \|x_\delta - \hat{x}\|_X = 0.$$

5.3 AN ESTIMATE BASED ON A SEMI-STRONG SOURCE CONDITION

The following assumption and theorem adapt Assumption 3.8 and Lemma 3.9 to non-linear A.

Assumption 5.7 (Semi-strong source condition; nonlinear case). Assume that $\hat{x} \in X$ satisfies for some $\hat{w} \in Y$ the basic source condition of (5.2). Moreover, for all $\delta > 0$, for given $\alpha_\delta, \gamma_\delta > 0$, assume that $J_\delta + \alpha_\delta R$ is $(A'(\hat{x}), \gamma_\delta)$-strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta A'(\hat{x})^* \hat{w}$ with respect to the set $\hat{X}$. The factor $\gamma > 0$ and neighbourhood $U = U^{\hat{x}}$ of $(A'(\hat{x}), \gamma_\delta)$-strong local subdifferentiability must be independent of $\delta > 0$. Then we say that $\hat{x}$ satisfies for $\hat{w}$ the *semi-strong source condition.*

Lemma 5.8. *Suppose Assumption 5.1 and the* semi-strong *source condition of Assumption 5.7 hold at some $\hat{x} \in X$ for some $\hat{w}$ with the neighbourhood of strong local subdifferentiability $U^{\hat{x}} \supset U_\rho$ for some $\rho > 0$. Suppose $(e_\delta + \delta^q + \delta^p)/\alpha_\delta \searrow 0$ and $\alpha_\delta \searrow 0$ as $\delta \searrow 0$. Then there exists $\bar{\delta} > 0$ and $\gamma = \gamma(\kappa)$ such that if $\delta \in (0, \bar{\delta})$ and $x_\delta \in U$, we have*

$$\operatorname{dist}^2(x_\delta, \hat{X}) \leq \frac{e_\delta}{\gamma\gamma_\delta} + \frac{\delta^2}{4\gamma'\gamma\gamma_\delta} + \frac{\eta^2 \alpha_\delta^2}{4(\gamma - \gamma')\gamma\gamma_\delta} \|\hat{w}\|_Y^2.$$

Proof. By Lemma 3.4 and the assumptions that $(e_\delta + \delta^q + \delta^p)/\alpha_\delta \searrow 0$ and $\alpha_\delta \searrow 0$ as $\delta \searrow 0$ and $U^{\hat{x}} \supset U_\rho$, for suitably small $\delta > 0$, we have $x_\delta \in U^{\hat{x}}$. As in (3.9), using the assumed (A, γ_δ)-strong local subdifferentiability of $J_\delta + \alpha_\delta R$, analogously to (5.7) we have

$$e_\delta \geq \langle E'(\hat{b} - b_\delta) - \alpha_\delta \hat{w}, A'(\hat{x})(x_\delta - \hat{x}) \rangle + \gamma \|A'(\hat{x})(x_\delta - \hat{x})\|_Y^2 + \gamma\gamma_\delta \operatorname{dist}^2(x_\delta, \hat{X}).$$

From here we proceed as in the proof of Theorem 5.5. □

The next result is proved exactly as Theorem 3.10 using Lemma 5.8 in place of Lemma 3.9.

Theorem 5.9. *Suppose Assumption 5.1 that there exists a collection $\tilde{X} \subset \hat{X}$ of points satisfying the* semi-strong *source condition of Assumption 5.7 such that $\bigcup_{\hat{x}\in\tilde{X}} U^{\hat{x}} \supset \hat{X}$. Also suppose that U_ρ, as defined in (5.5), is weakly or weakly-$*$ compact for some $\rho > 0$, and each $U^{\hat{x}}$ for all $\hat{x} \in \tilde{X}$ correspondingly weakly or weakly-$*$ open. If*

$$\lim_{\delta\searrow 0} \frac{1}{\min\{\alpha_\delta, \gamma_\delta\}} (\alpha_\delta^2, \delta^{\min\{2,q,p\}}, e_\delta) = 0$$

then

$$\lim_{\delta\searrow 0} \operatorname{dist}(x_\delta, \hat{X}) = 0.$$

5.4 EXAMPLES

The following Lemmas 5.10 and 5.11 are the counterparts of Lemmas 3.11 and 3.12 for nonlinear A. We concentrate for simplicity on $E = \frac{1}{2}\|\cdot\|_Y^2$. We need the *approximate linearity* condition

$$\frac{1}{2}\|A(x) - A(\hat{x})\|^2 + \langle A(\hat{x}) - b_\delta, A(x) - A(\hat{x}) - A'(\hat{x})(x - \hat{x})\rangle \geq \eta\|A'(\hat{x})(x - \hat{x})\|^2 \quad (x \in U_A) \tag{5.9}$$

for some $\eta > 0$ and a neighbourhood U_A of $\hat{x} \in \hat{X}$. By Pythagoras' three-point identity, this holds with $\eta = \frac{1}{2}$ and $U_A = X$ if A is linear. Given that Example 5.2 establishes $\|A(\hat{x}) - b_\delta\| = \|\hat{b} - b_\delta\| \leq \delta$ for $E = \frac{1}{2}\|\cdot\|_Y^2$, (5.9) follows from

$$\frac{1}{2}\|A(x) - A(\hat{x})\|^2 \geq \delta\|A(x) - A(\hat{x}) - A'(\hat{x})(x - \hat{x})\| + \eta\|A'(\hat{x})(x - \hat{x})\|^2 \quad (x \in U_A).$$

Lemma 5.10. *Let $E = \frac{1}{2}\|\cdot\|_Y^2$ on a Hilbert space Y and suppose A and R are as in Assumption 5.1. Suppose (5.9) holds at $\hat{x}$ for a given $\delta > 0$. Then $J_\delta + \alpha_\delta R$ is strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta\hat{d}$ with respect to the norm $\|\cdot\|_\delta$ if, for corresponding neighbourhood $U \ni \hat{x}$, $U \subset U_A$ and factor $\gamma > 0$,*

$$\alpha_\delta[R(x) - R(\hat{x}) - \langle\hat{d}|x - \hat{x}\rangle] + (\eta - \gamma)\,\|A'(\hat{x})(x - \hat{x})\|_2^2 \geq \gamma\gamma_\delta\|x - \hat{x}\|^2 \quad (x \in U). \tag{5.10}$$

The proof follows from the proof of the next lemma after expanding $\|\cdot\|_\delta$ and taking $\hat{X} = \{\hat{x}\}$.

Lemma 5.11. *Let $E = \frac{1}{2}\|\cdot\|_Y^2$ on a Hilbert space Y and suppose A and R are as in Assumption 5.1. Suppose (5.9) holds at $\hat{x}$ for a given $\delta > 0$. Then $J_\delta + \alpha_\delta R$ is (A, γ_δ)-strongly locally subdifferentiable at $\hat{x}$ for $J_\delta'(\hat{x}) - \alpha_\delta\hat{d}$ with respect to $\hat{X}$ if, for corresponding neighbourhood $U \ni \hat{x}$, $U \subset U_A$ and factor $\gamma > 0$,*

$$\alpha_\delta[R(x) - R(\hat{x}) - \langle\hat{d}|x - \hat{x}\rangle] + (\eta - \gamma)\,\|A'(\hat{x})(x - \hat{x})\|_2^2 \geq \gamma\gamma_\delta \operatorname{dist}^2(x, \hat{X}) \quad (x \in U). \tag{5.11}$$

Proof. Minding the definition of J_δ, we need to show that

$$\begin{aligned} &E(A(x) - b_\delta) - E(A(\hat{x}) - b_\delta) - \langle E'(A(\hat{x}) - b_\delta), A'(\hat{x})(x - \hat{x})\rangle \\ &\quad + \alpha_\delta[R(x) - R(\hat{x}) - \langle A'(\hat{x})^*\hat{w}|x - \hat{x}\rangle] \geq \gamma\|A'(\hat{x})(x - \hat{x})\|_Y^2 + \gamma\gamma_\delta \operatorname{dist}(x, \hat{X})^2 \quad (x \in U). \end{aligned} \tag{5.12}$$

For $E = \frac{1}{2}\|\cdot\|_Y^2$, using Pythagoras' identity (1.4) and the approximate linearity condition (5.9), we have

$$\begin{aligned} &E(A(x) - b_\delta) - E(A(\hat{x}) - b_\delta) - \langle E'(A(\hat{x}) - b_\delta), A'(\hat{x})(x - \hat{x})\rangle \\ &\quad = \frac{1}{2}\|A(x) - b_\delta\|^2 - \frac{1}{2}\|A(\hat{x}) - b_\delta\|^2 - \langle A(\hat{x}) - b_\delta, A'(\hat{x})(x - \hat{x})\rangle \\ &\quad = \frac{1}{2}\|A(x) - A(\hat{x})\|_Y^2 + \langle A(\hat{x}) - b_\delta, A(x) - A(\hat{x}) - A'(\hat{x})(x - \hat{x})\rangle \\ &\quad \geq \eta\|A'(\hat{x})(x - \hat{x})\|^2. \end{aligned}$$

Applying this in (5.11) proves (5.12). □

Corollary 5.12 (Nonlinear total variation regularised image reconstruction). *Let $E = \frac{1}{2}\|\cdot\|_Y^2$, $R(x) = \|Dx\|_{\mathcal{M}(\Omega;\mathbb{R}^m)}$, and $A \in C^1(X;Y)$ in $X = L^2(\Omega)$ and a Hilbert space Y. Suppose for some $\bar{\delta}, \rho > 0$ that the approximate linearity condition (5.9) holds at $\hat{x}$ for all $\delta \in (0, \bar{\delta})$ with $U_A \supset U_\rho$. Also suppose (5.3) holds, $\hat{x}, \hat{d} \in L^2(\Omega)$ satisfy (4.8), and there exists a corresponding collection $\mathcal{O}$ of strictly flat areas satisfying*

$$K_{\mathcal{O}}^* K_{\mathcal{O}} + A'(\hat{x})^* A'(\hat{x}) \geq \varepsilon \,\mathrm{Id} \quad \textit{for some } \varepsilon > 0. \tag{5.13}$$

If the accuracy and regularisation parameters satisfy

$$\lim_{\delta \searrow 0} \left(\alpha_\delta, \frac{\delta^{\min\{2,p,q\}}}{\alpha_\delta}, \frac{e_\delta}{\alpha_\delta} \right) = 0, \tag{5.14}$$

then $\|x_\delta - \hat{x}\|_2 \searrow 0$. Moreover, (5.6) holds for small enough $\delta > 0$.

Proof. Due to Example 5.2 and (5.3), Assumption 5.1 holds. By the assumption $U_A \supset U_\rho$, the approximate linearity condition (5.9) is valid for $\delta > 0$ small enough that $x_\delta \in U_\rho$ due to Lemma 5.4 and (5.14). Therefore, in the proofs of Theorem 4.12 and Lemma 4.11, where $\gamma_\delta = \alpha_\delta$, we simply replace Theorem 3.6 by Theorem 5.5, and Lemma 3.11 by Lemma 5.10. □

The unconditional Lasso example of Theorem 4.5 does not extend as readily to nonlinear A. However, if we assume that $A'(\hat{x})^* A'(\hat{x}) + \sum_{k \in Z(\hat{x},\hat{d})} \mathbb{1}_k \mathbb{1}_k^\top \geq \varepsilon \,\mathrm{Id}$ (compare (4.5) and (5.13)), then it is possible to produce convergence to specific $\hat{x}$ as in Corollary 5.12.

6 REGULARISATION COMPLEXITY OF OPTIMISATION METHODS IN HILBERT SPACES

We now briefly discuss how we can use some popular nonsmooth optimisation methods to construct x_δ satisfying the accuracy estimate (3.5) and the parameter convergence conditions (3.11). We start with forward-backward splitting, mainly applicable to the ℓ^1-regularised regression of Theorem 4.5, in which case it is also known as iterative soft-thresholding [11, 15, 39]. We then look at the more widely applicable primal-dual proximal splitting (PDPS), also known as the Chambolle–Pock method. Besides the original references below, the methods and their convergence properties are discussed, for example, in [14, 35]. Due to the necessities of effective first-order methods, we need to restrict our attention to Hilbert spaces.

6.1 FORWARD-BACKWARD SPLITTING

The forward-backward splitting method of [29] applies to problems of the form

$$\min_{x \in X} F(x) + G(x),$$

on Hilbert spaces X where $F : X \to \overline{\mathbb{R}}$ and $G : X \to \mathbb{R}$ are convex, proper, and lower semicontinuous, and G has an L-Lipschitz gradient. Take a step length parameter $\tau > 0$ such that $\tau L < 1$ and an initial iterate $x^0 \in X$. If F has a simple closed-form proximal operator $\mathrm{prox}_{\tau F}(x) := \min_{\tilde{x}} \frac{1}{2}\|x - \tilde{x}\| + \tau F(\tilde{x})$, the method iterates

$$x^{k+1} := \mathrm{prox}_{\tau F}(x^k - \tau \nabla G(x^k)).$$

Taking $F = J_\delta$ and $G = \alpha_\delta R$, we now apply the method to (3.2).

Theorem 6.1. *Suppose that Assumption 3.1 and the strong source condition of Assumption 3.5 hold. For each $\delta > 0$, take N_δ iterations of forward-backward splitting, starting from the same initial iterate $x^0 \in X$ with the same step length parameter $\tau > 0$ satisfying $\tau L < 1$. Denote the iterates by $\{x_\delta^k\}_{k \in \mathbb{N}}$. If*

$$\lim_{\delta \searrow 0} \frac{1}{\min\{\alpha_\delta, \gamma_\delta\}} (\alpha_\delta^2, \delta^2, N_\delta^{-1}) = 0, \tag{6.1}$$

then

$$\lim_{\delta \searrow 0} \|x_\delta^{N_\delta} - \hat{x}\|_X = 0.$$

Proof. The iterates of the forward-backward splitting method are monotone ($[F+G](x^{k+1}) \leq [F+G](x^k)$) and satisfy for any $\bar{x} \in X$ the estimate (see, e.g., [6, 14])

$$[F+G](x^N) - [F+G](\bar{x}) \leq \frac{1}{2\tau N}\|x^0 - \bar{x}\|^2 \quad (N \in \mathbb{N}).$$

Therefore, with $F(x) = \frac{1}{2}\|Ax - b_\delta\|_Y^2$ and $G(x) := \alpha_\delta R(x)$, the accuracy estimate (3.5) is satisfied for

$$e_\delta = \frac{1}{2\tau N_\delta}\|x^0 - \hat{x}\|^2$$

after taking N_δ iterations from the fixed initial iterate x^0. Thus the condition (3.11) of Corollary 3.7 is satisfied by choosing $\alpha_\delta > 0$ and N_δ such that (6.1) holds. □

In particular, if $\alpha_\delta = \gamma_\delta$ as in the Lasso of Theorem 4.5, it suffices to take $\alpha_\delta \searrow 0$ and $N_\delta \nearrow \infty$ such that $\delta^2/\alpha_\delta \searrow 0$, and $\alpha_\delta N_\delta \nearrow \infty$ as $\delta \searrow 0$.

6.2 PRIMAL-DUAL PROXIMAL SPLITTING

Primal-dual methods, for example the primal-dual proximal splitting (PDPS) of Chambolle and Pock [12], do not directly provide an accuracy estimate of the type (3.5). They provide estimates on a gap functional. To be more specific, consider the general problem

$$\min_{x \in X} F(x) + G(Kx), \tag{6.2}$$

for convex, proper, lower semicontinuous $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$, and $K \in \mathbb{L}(X;Y)$ on Hilbert spaces X and Y. Writing G^* for Fenchel conjugate of G, taking step length parameters $\tau, \sigma > 0$ with $\tau\sigma\|K\|^2 < 1$ and an initial iterate $(x^0, y^0) \in X\times$, the PDPS then iterates

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau F}(x^k - \tau K^* y^k), \\ y^{k+1} := \operatorname{prox}_{\sigma G^*}(y^k + \sigma K(2x^{k+1} - x^k)). \end{cases} \tag{6.3}$$

Define the Lagrangian gap functional

$$\mathcal{G}(x, y; \bar{x}, \bar{y}) := (F(x) + \langle Kx, \bar{y}\rangle - G^*(\bar{y})) - (F(\bar{x}) + \langle K\bar{x}, y\rangle - G^*(y)).$$

The iterates of the PDPS satisfy for all comparison points $(\tilde{x}, \tilde{y}) \in X \times Y$, for some constant $C > 0$ that [35, 14, 22]

$$\frac{1}{2}\|(x^N, y^N) - (\bar{x}, \bar{y})\|_M^2 + \sum_{k=0}^{N-1} \mathcal{G}(x^k, y^k; \tilde{x}, \tilde{y}) \leq \frac{1}{2}\|(x^0, y^0) - (\tilde{x}, \tilde{y})\|_M^2 \quad (N \in \mathbb{N}), \tag{6.4}$$

where

$$\|u\|_M := \sqrt{\langle Mu, u\rangle} \quad \text{and} \quad M := \begin{pmatrix} \tau^{-1}\operatorname{Id} & -K^* \\ -K & \sigma^{-1}\operatorname{Id} \end{pmatrix}.$$

We want to develop (6.4) into a function value estimate to use the regularisation theory of Section 3. For the next lemma, we need to know that by the Fenchel–Rockafellar theorem, minimisers $\bar{x} \in X$ of (6.2) are characterised by the existence of a *primal-dual solution pair* $(\bar{x}, \bar{y}) \in X \times Y$ such that

$$-K^*\bar{y} \in \partial F(\bar{x}) \quad \text{and} \quad K^*\bar{x} \in \partial G^*(\bar{y}).$$

Lemma 6.2. *Let $\{(x^k, y^k)\}_{k=1}^{\infty}$ be generated by the PDPS for the problem (6.2) and initial iterates (x^0, y^0). Suppose the step length parameters satisfy $\tau\sigma\|K\|^2 < 1$. Let $(\bar{x}, \bar{y})$ be a primal-dual solution pair. Define*

$$C := \|\bar{x}\|_X + \frac{\sqrt{\tau}\|(x^0, y^0) - (\bar{x}, \bar{y})\|_M}{\sqrt{1 - \tau\sigma\|K\|^2}}. \tag{6.5}$$

For all $N \in \mathbb{N}$, define the ergodic variables $\tilde{x}^N := \frac{1}{N}\sum_{k=0}^{N-1} x^N$ and $\tilde{y}^N := \frac{1}{N}\sum_{k=0}^{N-1} y^N$. Suppose there exists a bounded set $B_Y \subset Y$ such that

$$\sup_{\tilde{y}\in B_Y} (\langle Kx, \tilde{y}\rangle - G^*(\tilde{y})) = G(Kx) \quad (\|x\| \le C). \tag{6.6}$$

Then, for any $\tilde{x} \in X$,

$$F(\tilde{x}^N) + G(K\tilde{x}^N) \le F(\tilde{x}) + G(K\tilde{x}) + \sup_{\tilde{y}\in B_Y} \frac{\|(x^0, y^0) - (\tilde{x}, \tilde{y})\|_M^2}{2N}.$$

Proof. Since $(\bar{x}, \bar{y})$ is a primal-dual solution pair, we have $\mathcal{G}(\,\cdot\,; \bar{x}, \bar{y}) \ge 0$ as a consequence of the Fenchel–Rockafellar theorem. Since $\tau\sigma\|K\| < 1$, Young's inequality shows that $\langle Mu, u\rangle \ge \tau^{-1}(1 - \tau\sigma\|K\|^2)\|x\|^2$. By (6.4) we therefore have for all $k \ge \mathbb{N}$ that

$$\tau^{-1}(1 - \tau\sigma\|K\|^2)\|x^k - \bar{x}\|_X^2 \le \|(x^0, y^0) - (\bar{x}, \bar{y})\|_M^2.$$

In other words $\|x^k\| \le C$, consequently $\|\tilde{x}^N\| \le C$ for all $N \in \mathbb{N}$.

As in [38], we marginalise the gap with respect to the dual variable:

$$\inf_{y\in Y} \mathcal{G}(x, y; \tilde{x}, \tilde{y}) = (F(x) + \langle Kx, \tilde{y}\rangle - G^*(\tilde{y})) - (F(\tilde{x}) + G(K\tilde{x})) \quad (x, \tilde{x} \in X,\ \tilde{y} \in Y).$$

Using (6.6), it follows that

$$\sup_{\tilde{y}\in B_Y} \inf_{y\in Y} \mathcal{G}(x, y; \tilde{x}, \tilde{y}) \ge (F(x) + G(Kx)) - (F(\tilde{x}) + G(K\tilde{x})) \quad (\|x\| \le C).$$

Jensen's inequality and (6.4) give the ergodic gap estimate

$$\mathcal{G}(\tilde{x}^N, \tilde{y}^N; \tilde{x}, \tilde{y}) \le \frac{1}{2N}\|(x^0, y^0) - (\tilde{x}, \tilde{y})\|_M^2 \quad (N \in \mathbb{N}).$$

Using that $\|\tilde{x}^N\| \le C$ and combining these two inequalities, we obtain the claim. □

We now return to the problem (3.2) with $R = R_0 \circ Q$ for convex, proper and lower semicontinuous $G : Z \to \overline{\mathbb{R}}$ and $Q \in \mathbb{L}(X; Y)$. We assume that all X, Y, and Z are Hilbert spaces. For each $\delta > 0$, we define

$$F_\delta(x) := 0, \quad G_\delta(y, z) := \frac{1}{2}\|y - b_\delta\|^2 + \alpha_\delta R_0(z), \quad \text{and} \quad Kx := (Ax, Qx). \tag{6.7}$$

Then

$$J_\delta + \alpha_\delta R = F_\delta + G_\delta \circ K.$$

Theorem 6.3. *Suppose that (3.4) and the strong source condition of Assumption 3.5 hold at $\hat{x}$ with $R = R_0 \circ Q$ for convex, proper and lower semicontinuous $R_0 : Z \to \overline{\mathbb{R}}$ and $Q \in \mathbb{L}(X; Y)$. For each $\delta > 0$, take N_δ iterations of the PDPS for the problem $\min_x F_\delta(x) + G_\delta(Kx)$, starting for each $\delta > 0$ from the same initial iterate $(x^0, y_*^0, z_*^0) \in X \times Y \times Z$ with the same step length parameters $\tau, \sigma > 0$ satisfying $\tau\sigma(\|A\|^2 + \|Q\|^2) < 1$. Suppose $\operatorname{ran} \partial R_0 := \bigcup_{z\in Z} \partial R_0(z)$ is bounded and that $G_\delta \circ K$ is coercive for all*

$\delta > 0$. Denote the primal iterates by $\{x_\delta^k\}_{k\in\mathbb{N}}$, and the corresponding ergodic iterates by $\tilde{x}_\delta^N := \frac{1}{N}\sum_{k=0}^{N-1} x_\delta^k$. If

$$\lim_{\delta\searrow 0} \frac{1}{\min\{\alpha_\delta, \gamma_\delta\}}(\alpha_\delta^2, \delta^2, N_\delta^{-1}) = 0, \tag{6.8}$$

then

$$\lim_{\delta\searrow 0} \|\tilde{x}_\delta^{N_\delta} - \hat{x}\|_X = 0.$$

Proof. We use Corollary 3.7, for which we need to verify (3.11) for some e_δ satisfying (3.5). Assumption 3.5 and (3.4) of Assumption 3.1 we have assumed. We do this via Lemma 6.2 applied to $F = F_\delta$, $G = G_\delta$, and K, but need $C = C_\delta$ defined in (6.5) to be bounded over $\delta > 0$. We also need to construct B_Y satisfying (6.6) for $G = G_\delta$ for all small enough $\delta > 0$.

We have

$$\partial G_\delta(y, z) \subset \{y - b_\delta\} \times \alpha_\delta \partial R_0(z). \tag{6.9}$$

Since (3.4) implies that $b_\delta \to b$ as $\delta\searrow 0$, (6.9) and the assumption that $\operatorname{ran}\partial R_0$ is bounded prove the exists of a bounded set B_Y such that

$$\|x\| \le C \implies \partial G_\delta(Kx) \subset B_Y. \tag{6.10}$$

By the Fenchel–Young theorem (see, e.g., [18, 14]) we have

$$G_\delta(y, z) + G_\delta^*(y_*, y_*) = \langle y, y_*\rangle + \langle z, y_*\rangle \quad \text{when} \quad (y_*, z_*) \in \partial G_\delta(y, z).$$

Thus (6.10) proves (6.6).

For all $\delta > 0$, let $\bar{x}_\delta$ be a minimiser of $F_\delta + G_\delta \circ K$. Such a point exists because we assume $G_\delta \circ K$ to be proper, coercive, and lower semicontinuous. By the Fenchel–Rockafellar theorem (see, e.g., [18, 14]), there also exists a dual solution $(\bar{y}_\delta^*, \bar{z}_\delta^*)$, i.e., minimiser of $F_\delta^*(-K^*\,\cdot\,) + G_\delta^*$. now $x_\delta = \bar{x}_\delta$ satisfies (3.5) with $e_\delta = 0$. On the other hand, (6.8) verifies (3.11) for $e_\delta = 0$. Theorem 3.6 consequently shows for given $\varepsilon > 0$ and $\delta \in (0, \bar{\delta})$ for small enough $\bar{\delta} > 0$ that $\|\bar{x}_\delta\|_X \le \|\hat{x}\|_X + \varepsilon$. By the Fenchel–Rockafellar theorem, we have $(\bar{y}_\delta^*, \bar{z}_\delta^*) \in \partial G_\delta(K\bar{x}_\delta)$. Thus (6.10) bounds $\|\bar{y}_\delta^*\|$ and $\|\bar{z}_\delta^*\|$ uniformly over $\delta \in (0, \bar{\delta})0$. Consequently $C = C_\delta$ defined in (6.5) is bounded over $\delta \in (0, \bar{\delta})$.

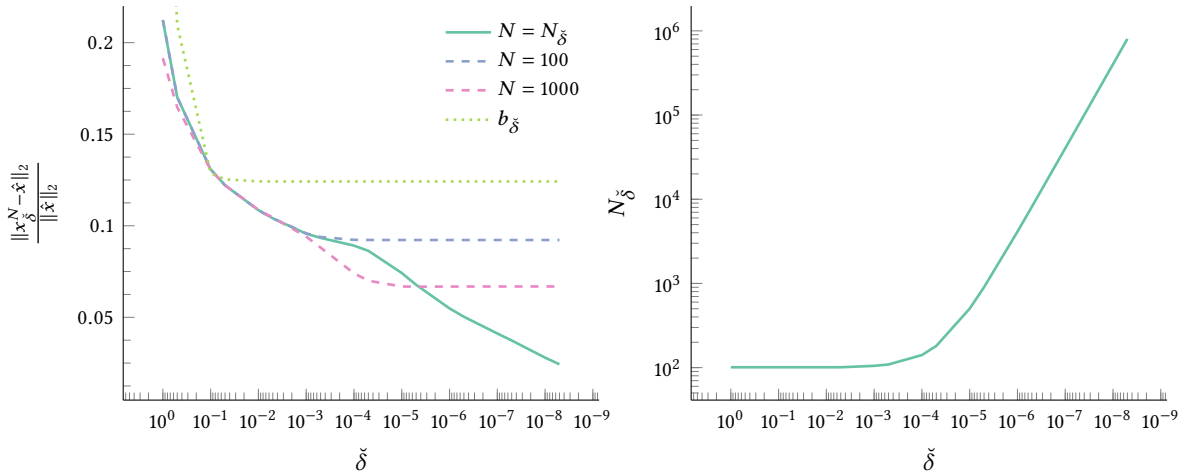
Now Lemma 6.2 proves (3.5) for all $\delta \in (0, \bar{\delta})$ for $x_\delta = \tilde{x}_\delta^{N_\delta}$ and

$$e_\delta := \sup_{\bar{y}\in B_Y} \frac{\|(x^0, y^0) - (\hat{x}, \bar{y})\|_M^2}{2N_\delta}.$$

Thus (6.8) implies (3.11). It remains to apply Corollary 3.7. □

In particular, if $\alpha_\delta = \gamma_\delta$, it suffices to take $\alpha_\delta\searrow 0$ and $N_\delta\nearrow\infty$ such that $\delta^2/\alpha_\delta\searrow 0$, and $\alpha_\delta N_\delta\nearrow\infty$ as $\delta\searrow 0$. However, we cannot directly apply Theorem 6.3 to the total variation Theorem 4.12 as it would require Banach spaces that the proximal steps in the PDPS cannot handle[3]. Nevertheless, Theorem 6.3 can be applied to discretised problems, as we next numerically demonstrate.

[3]It is, however, possible, to develop versions based on Bregman divergences; see, e.g., [34].



(a) Normalised distance to ground-truth (b) Iteration count $N_{\check\delta}$

Figure 1: Illustration of regularisation complexity of PDPS (Theorem 6.3) for total variation deblurring. In (a) we display the reconstruction quality in terms of the normalised distance to the ground truth after N iterations for fixed N and a choice $N_{\check\delta}$ satisfying the conditions of the theorem. We also display the quality of the corrupted data $b_{\check\delta}$. In (b) we plot the chosen iteration count $N_{\check\delta}$ against the pixelwise noise level $\check\delta$.

6.3 NUMERICAL ILLUSTRATION

We finish by numerically illustrating Theorem 6.3 for total variation deblurring. We take the 768×512 pixel "lighthouse" test image from the free Kodak image suite [20], converted to greyscale values in the range $[0, 1]$. This is the ground-truth $\hat x = \hat b$. Instead of the overall noise level $\delta = \delta(\check\delta)$, we use the pixelwise noise level $\check\delta$ as our main parameter.

To generate the data $b_{\check\delta}$, we apply pixelwise Gaussian noise of varying standard deviation $\check\delta$ to $\hat b$ and apply to the result our forward operator A, a convolution with a Gaussian kernel of standard deviation 2 in a window of 7×7 pixels. To employ isotropic total variation regularisation, we take as Q a forward-differences discretisation of the image gradient operator (cell width $h = 1$), and $R_0 = \|\cdot\|_{2,1}$ as the sum of two-norms of the gradient vectors over each image pixel. For each $\check\delta$ we take $\alpha_{\check\delta} = \check\delta/2$ based on rough visual inspection. To ensure that $N_{\check\delta}\alpha_{\check\delta} \nearrow \infty$ as $\check\delta \searrow 0$ with $N_{\check\delta} \nearrow \infty$ not too fast for numerical computation to become infeasible, and to always take at least 100 iterations, somewhat arbitrarily we choose $N_{\check\delta} = 100 + \alpha_{\check\delta}^{-1}(t \mapsto \log(1+t))^{1000}(\check\delta^{-1})$ (1000-fold composition of the logarithmic map). We plot this in Figure 1b.

We apply the PDPS (6.3) to the functions (6.7) for $\check\delta \in \{f \cdot 10^{-p} \mid f \in \{1, 0.5\},\ p \in \{0, \dots, 8\}\}$. We use zero initialisation and take as the step length parameters $\tau = 5/L$ and $\sigma = 0.99/(5L)$ for L a numerically computed estimate on the norm of K. We report in Figure 1a the normalised distances to the ground truth for $N = N_{\check\delta}$ and fixed $N = 100$ and $N = 1000$ iterations. The figure illustrates within numerical limits the convergence of the iterate $x_{\check\delta}^N$ for $N = N_{\check\delta}$ to the ground-truth as $\check\delta \searrow 0$, whereas with a fixed iteration count no convergence is observed. For further details, our Julia implementation of the experiments and algorithm is available on Zenodo [36].

REFERENCES

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

[2] G. Anzellotti, Pairings between measures and bounded functions and compensated compactness, *Annali di Matematica Pura ed Applicata* 135 (1983), 293–318, doi:10.1007/bf01781073.

[3] F. J. Aragón Artacho and M. H. Geoffroy, Characterization of metric regularity of subdifferentials, *Journal of Convex Analysis* 15 (2008), 365–380.

[4] F. J. Aragón Artacho and M. H. Geoffroy, Metric subregularity of the convex subdifferential in Banach spaces, *J. Nonlinear Convex Anal.* 15 (2014), 35–47.

[5] M. Bachmayr and M. Burger, Iterative total variation schemes for nonlinear inverse problems, *Inverse Problems* 25 (2009), doi:10.1088/0266-5611/25/10/105004.

[6] A. Beck, *First-Order Methods in Optimization*, SIAM, 2017, doi:10.1137/1.9781611974997.

[7] K. Bredies and M. Carioni, Sparsity of solutions for variational inverse problems with finite-dimensional data, *Calculus of Variations and Partial Differential Equations* 59 (2019), 14, doi:10.1007/s00526-019-1658-1.

[8] K. Bredies and M. Holler, Regularization of linear inverse problems with total generalized variation, *Journal of Inverse and Ill-posed Problems* 22 (2014), 871–913, doi:10.1515/jip-2013-0068.

[9] K. Bredies, K. Kunisch, and T. Pock, Total generalized variation, *SIAM Journal on Imaging Sciences* 3 (2011), 492–526, doi:10.1137/090769521.

[10] M. Burger and S. Osher, Convergence rates of convex variational regularization, *Inverse Problems* 20 (2004), 1411, doi:10.1088/0266-5611/20/5/005.

[11] A. Chambolle, R. A. DeVore, N. y. Lee, and B. J. Lucier, Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage, *IEEE Transactions on Image Processing* 7 (1998), 319–335, doi:10.1109/83.661182.

[12] 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.

[13] F. Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, 1990, doi:10.1137/1.9781611971309.

[14] C. Clason and T. Valkonen, Introduction to Nonsmooth Analysis and Optimization, 2020, arXiv:2001.00216. Work in progress.

[15] I. Daubechies, M. Defrise, and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communications on Pure and Applied Mathematics* 57 (2004), 1413–1457, doi:10.1002/cpa.20042.

[16] T. B. T. Do, *Discrete regularization for parameter identification problems*, PhD thesis, 2019, doi:10.17185/duepublico/70265.

[17] A. L. Dontchev and R. T. Rockafellar, Regularity and conditioning of solution mappings in variational analysis, *Set-valued and Variational Analysis* 12 (2004), 79–109, doi:10.1023/b:svan.0000023394.19482.30.

[18] I. Ekeland and R. Temam, *Convex analysis and variational problems*, SIAM, 1999.

[19] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Mathematics and Its Applications, Springer, 2000.

[20] 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/.

[21] M. Gaydu, Stability properties of the Tikhonov regularization for nonmonotone inclusions, *Journal of Global Optimization* 52 (2011), 843–853, doi:10.1007/s10898-011-9715-0.

[22] 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.

[23] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer, A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators, *Inverse Problems* 23 (2007), 987, doi:10.1088/0266-5611/23/3/009.

[24] A. Ioffe, *Variational Analysis of Regular Mappings: Theory and Applications*, Springer Monographs in Mathematics, Springer, 2017, doi:10.1007/978-3-319-64277-2.

[25] A. D. Ioffe, Regular points of Lipschitz functions, *Transactions of the American Mathematical Society* 251 (1979), 61–69, doi:10.1090/s0002-9947-1979-0531969-6.

[26] J. Jauhiainen, P. Kuusela, A. Seppänen, and T. Valkonen, Relaxed Gauss–Newton methods with applications to electrical impedance tomography, *SIAM Journal on Imaging Sciences* 13 (2020), 1415–1445, doi:10.1137/20m1321711, arXiv:2002.08044.

[27] B. Kaltenbacher, Minimization Based Formulations of Inverse Problems and Their Regularization, *SIAM Journal on Optimization* 28 (2018), 620–645, doi:10.1137/17m1124036.

[28] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, number 6 in Radon Series on Computational and Applied Mathematics, De Gruyter, 2008.

[29] P. Lions and B. Mercier, Splitting algorithms for the sum of two nonlinear operators, *SIAM Journal on Numerical Analysis* 16 (1979), 964–979, doi:10.1137/0716071.

[30] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations*, American Mathematical Society, 2001.

[31] W. Ring, Structural Properties of Solutions to Total Variation Regularization Problems, *ESAIM: Mathematical Modelling and Numerical Analysis* 34 (2000), 799–810, doi:10.1051/m2an:2000104.

[32] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer, 1998, doi:10.1007/978-3-642-02431-3.

[33] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. Kazimierski, *Regularization Methods in Banach Spaces*, Radon Series on Computational and Applied Mathematics, De Gruyter, 2012.

[34] T. Valkonen, First-order primal-dual methods for nonsmooth nonconvex optimisation, in *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*, K. Chen, C. B. Schönlieb, X. C. Tai, and L. Younes (eds.), Springer, 2020, arXiv:1910.00115. accepted.

[35] T. Valkonen, Testing and non-linear preconditioning of the proximal point method, *Applied Mathematics and Optimization* 82 (2020), doi:10.1007/s00245-018-9541-6, arXiv:1703.05705.

[36] T. Valkonen, Codes for "Regularisation, optimisation, subregularity", Software on Zenodo, 2021, doi:10.5281/zenodo.4432968.

[37] T. Valkonen, Preconditioned proximal point methods and notions of partial subregularity, *Journal of Convex Analysis* 28 (2021), 251–278, arXiv:1711.05123.

[38] T. Valkonen, Predictive online optimisation with applications to optical flow, *Journal of Mathematical Imaging and Vision* (2021), doi:10.1007/s10851-020-01000-4, arXiv:2002.03053, https://tuomov.iki.fi/m/predict.pdf.

[39] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, Sparse reconstruction by separable approximation, *IEEE Transactions on Signal Processing* 57 (2009), 2479–2493, doi:10.1109/tsp.2009.2016892.

[40] H. Zhang, W. Yin, and L. Cheng, Necessary and sufficient conditions of solution uniqueness in 1-norm minimization, *Journal of Optimization Theory and Applications* 164 (2015), 109–122, doi:10.1007/s10957-014-0581-z.