TECHNISCHE UNIVERSITÄT
CHEMNITZ

Fakultät für Mathematik

Habilitationsschrift

zur Erlangung des akademischen Grades
Doctor rerum naturalium habilitatus (Dr. rer. nat. habil.)

Quadratic Inverse Problems and Sparsity Promoting Regularization

Two subjects, some links between them, and an application in laser optics

vorgelegt von

Dr. rer. nat. Jens Flemming

geboren am 1. Oktober 1985 in Gera

Chemnitz, den 12. Januar 2018

vorgelegt am 18. Mai 2017, angenommen am 12. Oktober 2017

Gutachter: Prof. Dr. Bernd Hofmann (TU Chemnitz)
Prof. Dr. Dirk Lorenz (TU Braunschweig)
Prof. Dr. Otmar Scherzer (Universität Wien)

Flemming, Jens
Quadratic Inverse Problems and Sparsity Promoting Regularization
Two subjects, some links between them, and an application in laser optics
Habilitationsschrift, Fakultät für Mathematik
Technische Universität Chemnitz, Januar 2018

Contents

Preface

In this thesis I compile and slightly extend my research results found between 2011 and 2017, partly obtained in cooperation with other researchers. Compared to previously published articles the thesis includes additional explanations and all results are presented in consistent notation.

The two topics of the thesis, quadratic inverse problems and ℓ^1-regularization, seem to be quite different. The first is concerned with nonlinear mappings in a classical Hilbert space setting, whereas the second deals with linear mappings in non-reflexive Banach spaces. Both subjects met more or less by chance at my desk: work on quadratic problems was heavily influenced by a project on measurement techniques in laser optics between TU Chemnitz and Max Born Institute Berlin, work on ℓ^1-regularization was instigated by colleagues knowing my articles on convergence rate theory in Banach spaces and asking how to obtain rates for ℓ^1-regularization in case of non-sparse solutions.

At the second sight, both subjects have similar structures and their handling shows several parallels. Nevertheless, I decided to devide the thesis into two independent parts and to give hints on cross connections from time to time. The advantage of this decisison is that the reader may study both parts in arbitrary order.

Next to some auxiliary material the appendix contains an unpublished result on variational source conditions for convex regularization in Banach spaces.

Finishing this thesis would not have been possible without constant support and advice by Prof. Bernd Hofmann (TU Chemnitz). I thank him a lot for his efforts in several regards during all the years I have been working in his research group. I also want to thank my colleagues and coauthors, especially Steven Bürger and Daniel Gerth, for interesting and fruitful discussions. Last but not least I have to express my thanks to the Faculty of Mathematics at TU Chemnitz as a whole for the cordial and cooperative working atmosphere.

Chemnitz, April 2017
Jens Flemming

Part I.

Quadratic inverse problems

1. What are quadratic inverse problems?

Ill-posed inverse problems are frequently divided into two classes: linear and nonlinear ones. The reason for this distinction is that for linear inverse problems there is a huge and almost closed theory of regularization whereas for nonlinear ones there are only weak theoretical results and each concrete nonlinear inverse problem has to be handled in a different way. In the present part of the thesis we consider a subclass of nonlinear inverse problems and develop theoretical results and algorithms which can be applied to all inverse problems from this class. We call the subclass the class of *quadratic inverse problems*.

1.1. Definition and basic properties

Let X and Y be Banach spaces over $\mathbb{R}$ or $\mathbb{C}$ and let $F : X \to Y$ be a (nonlinear) mapping. The inverse problem under consideration is the equation

$$F(x) = y^\dagger \tag{1.1}$$

with given $y^\dagger$ in Y and sought-for x in X. We assume that there exists a solution, that is, $y^\dagger$ belongs to the range of F.

Definition 1.1. The mapping F is called *quadratic* if there is a continuous bilinear mapping $B : X \times X \to Y$ such that

$$F(x) = B(x, x)$$

holds for all x in X.

Note that by definition quadratic mappings are always continuous.

The quadratic structure implies several simple facts which will be used throughout the thesis. So let F be quadratic with underlying bilinear mapping B. Then we have

$$\begin{aligned} F(x+u) &= B(x+u, x+u) = B(x,x) + B(x,u) + B(u,x) + B(u,u) \\ &= F(x) + B(x,u) + B(u,x) + F(u) \end{aligned}$$

for x and u in X and also

$$\begin{aligned} F(t\,x) &= B(t\,x, t\,x) = t^2\,B(x,x) \\ &= t^2\,F(x) \end{aligned}$$

for scalars t. In particular, quadratic mappings cannot be injective because

$$F(x) = F(-x).$$

We also see
$$F(0) = 0.$$

Note that there might be many different bilinear mappings B, the diagonals of which produce the same quadratic mapping F. But restricting our attention to symmetric bilinear mappings, we enforce uniqueness.

Proposition 1.2. *Let F be quadratic. Then there is exactly one symmetric bilinear mapping $B_F : X \times X \to Y$ with*
$$F(x) = B_F(x, x)$$
for all x in X. The mapping B_F is given by
$$B_F(x, u) = F\left(\frac{x + u}{2}\right) - F\left(\frac{x - u}{2}\right) \tag{1.2}$$
for x and u in X.

Proof. Obviously, $B_F(x, x) = F(x)$ for all x and from
$$F\left(\frac{x - u}{2}\right) = F\left(-\frac{u - x}{2}\right) = F\left(\frac{u - x}{2}\right)$$
we see that B_F is symmetric. Assume that there is another symmetric bilinear mapping B with $B(x, x) = F(x)$. Then
$$\begin{aligned} B(x, u) &= \frac{1}{4}\big(B(x + u, x + u) - B(x - u, x - u)\big) \\ &= F\left(\frac{x + u}{2}\right) - F\left(\frac{x - u}{2}\right) = B_F(x, u) \end{aligned}$$
for all x and u in X. That is, $B_F = B$. □

We already mentioned above that quadratic mappings cannot be injective because $F(x) = F(-x)$. We want to distinguish between mappings which only have this simple non-injectivity and mappings which are really non-injective.

Definition 1.3. A quadratic mapping F is *injective up to sign* if for all x and u in X equality $F(x) = F(u)$ implies $x = u$ or $x = -u$.

Injectivity up to sign can be characterized by the behavior of F around zero. This is quite similar to linear mappings.

Proposition 1.4. *A quadratic mapping F is injective up to sign if and only if for all x and u in X equality $B_F(x, u) = 0$ implies $x = 0$ or $u = 0$.*

Proof. Injectivity up to sign is equivalent to
$$F(\tilde{x}) = F(\tilde{u}) \quad \Rightarrow \quad \tilde{x} = \tilde{u} \text{ or } \tilde{x} = -\tilde{u}.$$

Setting $x := \tilde{x} - \tilde{u}$ and $u := \tilde{x} + \tilde{u}$ this can be rewritten as

$$F\left(\frac{x+u}{2}\right) = F\left(\frac{x-u}{2}\right) \quad \Rightarrow \quad x = 0 \text{ or } u = 0$$

and simplifying the equality on the left we obtain

$$B_F(x,u) = 0 \quad \Rightarrow \quad x = 0 \text{ or } u = 0.$$

□

Like for linear mappings global properties of quadratic mappings can be deduced from their local behavior.

Proposition 1.5. *Let F be quadratic and let ε be positive. The mapping F is uniquely determined by its values on the closed ε-ball around any point x_0.*

Proof. For x in $X \setminus \{0\}$ write

$$F(x) = \frac{\|x\|^2}{\varepsilon^2}\left(\frac{1}{2}F\left(x_0 + \varepsilon\frac{x}{\|x\|}\right) + \frac{1}{2}F\left(x_0 - \varepsilon\frac{x}{\|x\|}\right) - F(x_0)\right)$$

to verify the assertion. □

Another similarity to linear mappings is that quadratic mappings constitute a normed vector space. Obviously, (pointwise) sums and scalar multiples of quadratic mappings are again quadratic mappings. It remains to define a norm on this vector space. But since the vector space of all continuous bilinear mappings B carries the norm

$$\|B\| := \sup_{\substack{x,u \in X \\ \|x\| \le 1, \|u\| \le 1}} \|B(x,u)\|$$

we may define

$$\|F\| := \|B_F\|.$$

This immediately provides the estimate

$$\|F(x)\| \le \|F\| \, \|x\|^2 \tag{1.3}$$

for all x in X. Together with Proposition 1.2 we obtain that the space of quadratic mappings is isometrically isomorphic to the space of continuous symmetric bilinear mappings.

For later use we introduce the notion of the adjoint of a bilinear mapping. Given a symmetric bounded bilinear mapping $B : X \times X \to Y$ and elements x from X and η from Y^* the functional

$$u \mapsto \langle \eta, B(x,u) \rangle$$

on X is obviously linear and bounded. Thus it can be represented by some element from X^* which we denote by $B^*(x,\eta)$. In this way we obtain a bounded sesquilinear mapping $B^* : X \times Y^* \to X^*$, to which we refer as the adjoint mapping of B. The mapping B is antilinear in its first component and linear in its second, that is,

$B^*(a\,x, b\,\eta) = \overline{a}\,b\,B^*(x,\eta)$ for complex numbers a and b. The same construction works for the functional $u \mapsto \langle \eta, B(u,x)\rangle$ and, since B is symmetric, yields the same sesquilinear mapping B^*. The adjoint B^* is uniquely determined by

$$\langle \eta, B(x,u)\rangle = \langle B^*(x,\eta), u\rangle \qquad \text{for all } u,\ x \text{ in } X \text{ and all } \eta \text{ in } Y^*. \tag{1.4}$$

A last basic property of quadratic mappings we want to mention in this introductory section is that quadratic mappings always are differentiable.

Proposition 1.6. *Each quadratic mapping F is Fréchet differentiable on X. The Fréchet derivative $F'[x] : X \to Y$ at x is given by*

$$F'[x]\,h = 2\,B_F(x,h) \qquad \text{for } h \text{ in } X.$$

Proof. For x and h in X we have

$$\frac{\|F(x+h) - F(x) - 2\,B_F(x,h)\|}{\|h\|} = \frac{\|F(h)\|}{\|h\|} \leq \|F\|\,\|h\|.$$

Taking the limit $\|h\| \to 0$ completes the proof. □

Quadratic mappings are also twice Fréchet differentiable. For F'' we have

$$F''[x](h_1,h_2) = 2\,B(h_1,h_2)$$

for all h_1, h_2 and x in X. Obviously, all higher derivatives are zero. Thus, the Taylor expansion of F at x_0 is

$$F(x) = F(x_0) + 2\,B(x_0, x - x_0) + B(x - x_0, x - x_0)$$

for all x in X.

1.2. Examples

We provide several examples of quadratic mappings and corresponding inverse problems. The focus in this thesis is on different types of autoconvolutions but the results apply to other quadratic mappings, too.

1.2.1. Autoconvolutions

Let $X = L^2(0,1)$ be the space of real-valued or complex-valued square integrable functions over the interval $(0,1)$. There are three common types of autoconvolution operations on this space.

Autoconvolution of functions with uniformly bounded support

We may interpret functions in $L^2(0,1)$ as functions defined on the whole real line, but with support contained in $(0,1)$. In other words, the functions have uniformly bounded support and without loss of generality we assume that the support is bounded by zero and one. Then the usual autoconvolution of real-valued or complex-valued functions over the real line reduces to

$$\big(F(x)\big)(s) := \begin{cases} \int\limits_0^s x(s-t)\,x(t)\,\mathrm{d}t, & \text{if } s \in (0,1), \\ \int\limits_{s-1}^1 x(s-t)\,x(t)\,\mathrm{d}t, & \text{if } s \in (1,2) \end{cases}$$

or, which is the same,

$$\big(F(x)\big)(s) = \int\limits_{\max\{0,\,s-1\}}^{\min\{s,\,1\}} x(s-t)\,x(t)\,\mathrm{d}t, \quad s \in (0,2). \tag{1.5}$$

Proposition 1.7. *The mapping F defined by* (1.5) *is a quadratic mapping from $L^2(0,1)$ into $L^2(0,2)$ with $\sqrt{\frac{2}{3}} \leq \|F\| \leq 1$. Further, F is injective up to sign.*

Proof. With $B_F : L^2(0,1) \times L^2(0,1) \to L^2(0,2)$ given by

$$\big(B_F(x,u)\big)(s) := \int\limits_{\max\{0,\,s-1\}}^{\min\{s,\,1\}} x(s-t)\,u(t)\,\mathrm{d}t, \quad s \in (0,2) \tag{1.6}$$

we have $F(x) = B_F(x,x)$. To prove that F is quadratic it remains to show that the symmetric bilinear mapping B_F is continuous. We write

$$\begin{aligned} \|B_F(x,u)\|^2 &= \int\limits_0^2 \big|\big(B_F(x,u)\big)(s)\big|^2\,\mathrm{d}s \\ &= \int\limits_0^1 \left|\int\limits_0^s x(s-t)\,u(t)\,\mathrm{d}t\right|^2 + \left|\int\limits_s^1 x(s+1-t)\,u(t)\,\mathrm{d}t\right|^2 \mathrm{d}s \end{aligned}$$

and apply the Cauchy–Schwarz inequality to bound the first inner integral by

$$\left|\int\limits_0^s x(s-t)\,u(t)\,\mathrm{d}t\right|^2 \leq \big\|x|_{(0,s)}\big\|^2\,\big\|u|_{(0,s)}\big\|^2$$

and the second by

$$\begin{aligned} \left|\int\limits_s^1 x(s+1-t)\,u(t)\,\mathrm{d}t\right|^2 &\leq \big\|x|_{(s,1)}\big\|^2\,\big\|u|_{(s,1)}\big\|^2 \\ &= \Big(\|x\|^2 - \big\|x|_{(0,s)}\big\|^2\Big)\,\Big(\|u\|^2 - \big\|u|_{(0,s)}\big\|^2\Big). \end{aligned}$$

Thus, we see

$$\begin{aligned}
&\|B_F(x,u)\|^2\\
&\leq \int_0^1 \|x\|^2\,\|u\|^2 - \left\|x|_{(0,s)}\right\|^2 \left(\|u\|^2 - \left\|u|_{(0,s)}\right\|^2\right) - \left\|u|_{(0,s)}\right\|^2 \left(\|x\|^2 - \left\|x|_{(0,s)}\right\|^2\right) \mathrm{d}s\\
&= \int_0^1 \|x\|^2\,\|u\|^2 - \left\|x|_{(0,s)}\right\|^2 \left\|u|_{(s,1)}\right\|^2 - \left\|u|_{(0,s)}\right\|^2 \left\|x|_{(s,1)}\right\|^2 \mathrm{d}s
\end{aligned}$$

and therefore

$$\|B_F(x,u)\|^2 \leq \int_0^1 \|x\|^2\,\|u\|^2\,\mathrm{d}s = \|x\|^2\,\|u\|^2.$$

Next to the continuity of F we also obtain from this estimate that $\|F\| \leq 1$.

Injectivity up to sign is a consequence of Titchmarsh's theorem (see [Tit26, Theorem VII]). It states that if $B_F(x,u) = 0$ then either $x = 0$ or $u = 0$ has to be true. This is exactly the characterization of injectivity up to sign provided by Proposition 1.4.

The lower bound for $\|F\|$ follows from $\|F(x)\| = \sqrt{\frac{2}{3}}$ if $x(t) = 1$ for all t. □

A proof of $\|F\| \leq 1$ has already been given in [Bür16, Corollary 24]. But one seems to be only an upper bound which is not sharp.

Conjecture 1.8. *For F defined by* (1.5) *we conjecture that*

$$\|F\| = \sqrt{\frac{2}{3}}.$$

This conjecture is based on two observations: For $x(t) = 1$, $t \in (0,1)$, we have $\|F(x)\| = \sqrt{\frac{2}{3}}$ and numerical approximation of $\|F\|$ suggests that $\|F(x)\|$ is maximized by $x \equiv 1$ under the constraint $\|x\| \leq 1$. Several attempts to prove this conjecture failed.

For real spaces in [FH96, Proposition 2.3] local ill-posedness (cf. Definition 1.13) of F at every point was shown if the domain of F is restricted to non-negative functions. Local ill-posedness in every point was shown in [BH15, Examples 3.1, 3.2] for the complex case. The mapping F is known to be weakly continuous and to not being compact, see [ABHS16, Proposition 1].

The type of autoconvolution discussed here appears for example if the density of two identically and independently distributed random variables with density supported in $[0,1]$ has to be computed. The corresponding inverse problem is to find the density of the underlying random variables from the density of their sum.

Truncated autoconvolution of functions with uniformly bounded support

The autoconvolution operation introduced above maps from $L^2(0,1)$ into $L^2(0,2)$. Restricting the images to the interval $(0,1)$ yields a mapping $F : L^2(0,1) \to L^2(0,1)$ given by

$$\big(F(x)\big)(s) := \int_0^s x(s-t)\,x(t)\;\mathrm{d}t, \quad s \in (0,1). \tag{1.7}$$

Proposition 1.9. *The mapping F defined by* (1.7) *is a quadratic mapping from $L^2(0,1)$ into $L^2(0,1)$ with* $0.6860 \leq \|F\| \leq 1$.

Proof. The proposition can be proven analogously to Proposition 1.7. The lower bound for $\|F\|$ follows from $\|F(x)\| = \sqrt{\frac{1}{6} + \frac{3}{\pi^2}} \geq 0.6860$ if $x(t) = \sqrt{2}\,\cos\bigl(\frac{\pi}{2}\,t\bigr)$ for all t. □

Obviously the mapping F is not injective up to sign because all functions with support contained in $(\frac{1}{2}, 1)$ are mapped to zero. Note that the lower bound for $\|F\|$ is greater than the conjectured norm of F in the untruncated case, see previous subsection.

In [GH94] several further properties of F in case of real spaces were proven. For instance, that F is weakly continuous on properly restricted domains, that $F(x)$ is a continuous function for all x, and that F is compact on certain subsets but not on the whole space. Weak continuity on the whole (real) space was shown in [Bür16, Proposition 8]. The mapping F between real spaces is locally ill-posed everywhere, see [BH15, Example 2.1]. The same is true if the L^2-spaces are replaced by spaces of continuous functions, see [Bür14, Example 2.2].

Truncated autoconvolution mappings as discussed here play an important role in spectroscopy. See, e. g. ,[Bau91] for an application in appearance potential spectroscopy.

Autoconvolution of periodic functions

We may interpret real-valued or complex-valued functions in $L^2(0,1)$ as periodic functions on the whole real line. Then the usual autoconvolution of periodic functions can be written as

$$\bigl(F(x)\bigr)(s) = \int\limits_0^s x(s-t)\,x(t)\,\mathrm{d}t + \int\limits_s^1 x(s+1-t)\,x(t)\,\mathrm{d}t, \quad s \in (0,1). \tag{1.8}$$

Proposition 1.10. *The mapping F defined by* (1.8) *is a quadratic mapping from $L^2(0,1)$ into $L^2(0,1)$ with* $\|F\| = 1$.

Proof. The underlying symmetric bilinear mapping B_F can be defined in the obvious way, cf. proof of Proposition 1.7. To show its boundedness we write

$$\|B_F(x,u)\|^2 = \int\limits_0^1 \left(\int\limits_0^s x(s-t)\,u(t)\,\mathrm{d}t + \int\limits_s^1 x(s+1-t)\,u(t)\,\mathrm{d}t \right)^2 \mathrm{d}s$$

and apply the Cauchy–Schwarz inequality to obtain

$$\|B_F(x,u)\|^2 \leq \int\limits_0^1 \Bigl(\bigl\|x|_{(0,s)}\bigr\|\,\bigl\|u|_{(0,s)}\bigr\| + \bigl\|x|_{(s,1)}\bigr\|\,\bigl\|u|_{(s,1)}\bigr\| \Bigr)^2 \mathrm{d}s.$$

Using again the Cauchy–Schwarz inequality, but now for vectors with two components, we see

$$\begin{aligned} &\bigl\|x|_{(0,s)}\bigr\|\,\bigl\|u|_{(0,s)}\bigr\| + \bigl\|x|_{(s,1)}\bigr\|\,\bigl\|u|_{(s,1)}\bigr\| \\ &\leq \sqrt{\bigl\|x|_{(0,s)}\bigr\|^2 + \bigl\|x|_{(s,1)}\bigr\|^2}\,\sqrt{\bigl\|u|_{(0,s)}\bigr\|^2 + \bigl\|u|_{(s,1)}\bigr\|^2} = \|x\|\,\|u\| \end{aligned}$$

and therefore

$$\|B_F(x,u)\|^2 \leq \|x\|^2 \, \|u\|^2.$$

Thus, B_F is bounded and $\|F\| \leq 1$. Equality $\|F\| = 1$ follows from $F(x) = x$ for $x(t) = 1$, $t \in (0,1)$. □

Note that F defined by (1.8) is not injective up to sign, because we find $x \neq 0$ and $u \neq 0$ with $B_F(x,u) \neq 0$ (cf. Proposition 1.4). Choose, e.g.,

$$x(t) = \sin(2\,\pi\,t) \quad \text{and} \quad u(t) = \sin(4\,\pi\,t), \quad t \in (0,1).$$

Exploiting the Fourier convolution theorem we see that $B_F(x,u) = 0$ if and only if the Fourier transforms of x und u have disjoint supports.

This type of autoconvolution is well-known in many branches of mathematics. In the sequel we will not consider the corresponding inverse problem, but we will use this type of autoconvolution for proving results for the before mentioned variants of autoconvolution (cf. Example 4.7).

1.2.2. Kernel-based autoconvolution in laser optics

The example of a quadratic mapping presented in this subsection originates from joint research work of TU Chemnitz (professorship ‘Inverse Problems’) and Max Born Institute for Nonlinear Optics and Short Pulse Spectroscopy in Berlin (research group ‘Solid State Light Sources’).

Ultra-short laser pulses

With today’s laser technology scientists are able to produce sequences of extremely short laser pulses. Up to 100 million pulses per second, each lasting only five to 100 femtoseconds, can be generated with so called femtosecond lasers. Such ultra-short pulses are the shortest events mankind can produce. To get a better idea of the duration of a femtosecond we note that a femtosecond is related to a second in the same way as a second is related to 31.7 million years. Also note that light travels only 0.3 micrometers per femtosecond and that the period of visible light is about two femtoseconds. Since the reaction time of the pigments in the human eye is at 200 femtoseconds we are not able to see ultra-short laser pulses.

Applications of ultra-short laser pulse are manifold. They allow, for example, to machine inside a material without affecting its surface or to drill and cut without measurable burr and without build-up heat. Also the observation of chemical reactions in realtime can be realized with the help of ultra-short laser pulses.

SD-SPIDER method

Characterizing ultra-short laser pulses is a major challenge due to the fact that no direct measurements can be obtained. Each pulse lasts only few optical cycles and thus even light is to slow to obtain full information about the variation of the electric field during one pulse.

Existing approaches for characterizing ultra-short laser pulses are autocorrelation techniques, frequency-resolved optical gating (FROG) and spectral phase interferometry for direct electric field reconstruction (SPIDER). In 2010 the research group 'Solid State Light Sources' at Max Born Institute for Nonlinear Optics and Short Pulse Spectroscopy in Berlin presented the self-diffraction SPIDER method (SD-SPIDER), which allows for fully characterizing ultra-short pulses but requires the solution of an autoconvolution problem.

Figure 1.1 shows the experimental setup at Max Born Institute Berlin and Figure 1.2 provides a scheme of the SD-SPIDER method.

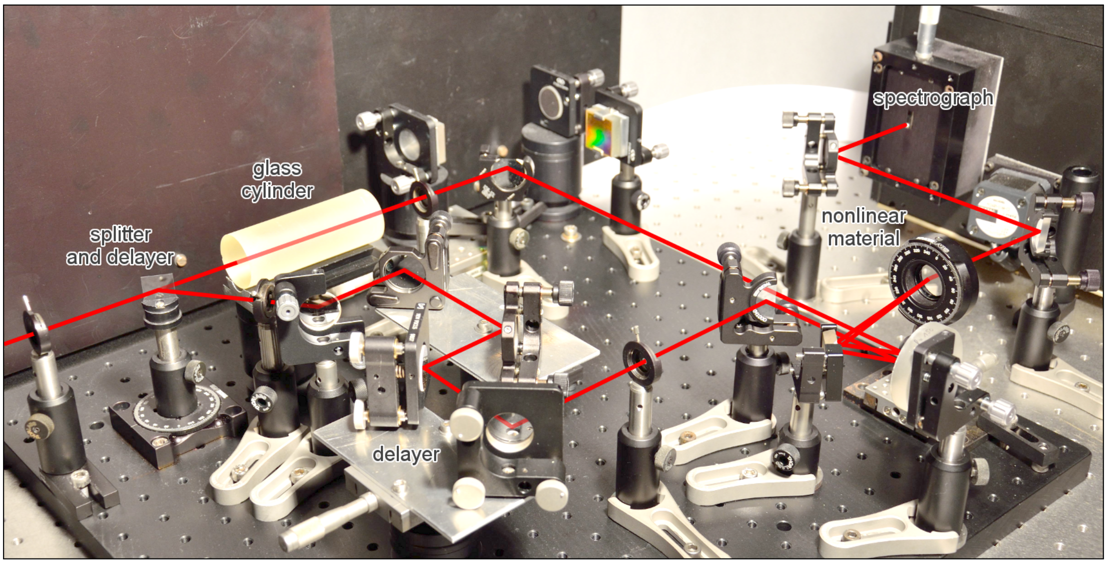

Figure 1.1.: Experimental setup for SD-SPIDER method. The read lines indicate the path of the laser pulses (photo courtesy by S. Birkholz from Max Born Institute Berlin, red lines and text added by the author).

At first the beam of ultra-short pulses generated by the laser is split into two beams. One passes through a long glass cylinder which results in chirped pulses, that is, due to frequency-dependent speed of light in glass frequencies are segregated and the pulse is stretched. The other beam is split again. After delaying one part slightly both parts are rejoined, resulting in a doubling of each pulse. In practice the doubling is realized by the first splitter since both sides of the splitter reflect the beam with a slight delay between both reflected beams.

The two beams are rejoined inside a so called nonlinear optical material with such a delay that a chirped pulse meets a doubled pulse, resulting in a beam of spectrally sheared delayed pairs of pulses, which then enters a spectrograph. With the help of the spectrograph the interferences in the frequency domain are recorded. The nonlinear material used at Max Born Institute Berlin is barium fluoride.

The inverse problem

We do not go into the details of the full path from the measured data to absolute value and phase of the original laser pulse. Figure 1.3 shows a data set from the experiments

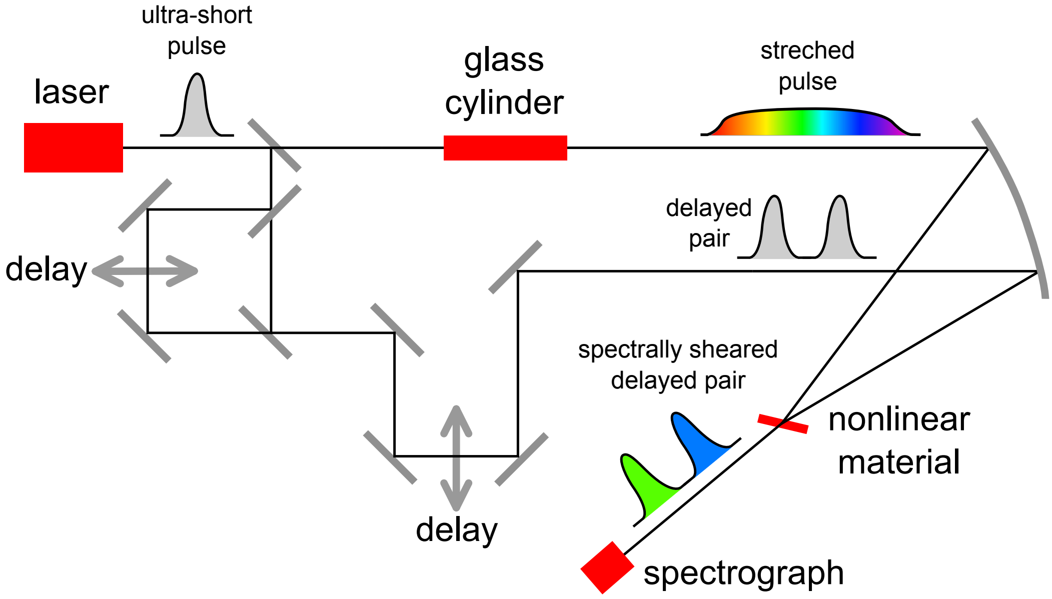

Figure 1.2.: Scheme of experimental setup for SD-SPIDER method.

in Berlin and Figure 1.4 shows the data which results from preprocessing steps (mainly the Takeda algorithm) and which is the relevant information for the inverse problem part of the reconstruction process.

The inverse problem consists in solving a kernel-based autoconvolution equation. Denote by $L^2_{\mathbb{C}}(0,1)$ the space of complex-valued square integrable functions over the interval $(0,1)$ and let $k : \mathcal{D}(k) \to \mathbb{C}$ be a bounded and continuous function with domain

$$\mathcal{D}(k) = \big\{(s,t) : 0 \le s \le 2,\ \max\{0,\ s-1\} \le t \le \min\{s,\ 1\}\big\}.$$

Then the mapping $F : L^2_{\mathbb{C}}(0,1) \to L^2_{\mathbb{C}}(0,2)$ in (1.1) for the SD-SPIDER reconstruction problem is given by

$$\big(F(x)\big)(s) = \int\limits_{\max\{0,\,s-1\}}^{\min\{s,\,1\}} k(s,t)\,x(s-t)\,x(t)\,\mathrm{d}t, \quad s \in (0,2), \tag{1.9}$$

which is the same as (1.5) if the kernel k is one everywhere.

The function x represents the Fourier transform of the desired electric field of the ultra-short laser pulse over time and the function $F(x)$ is some intermediate step between the original electric field and the recorded SD-interferogram. In practice the absolute value of x can be measured, but the absolute value of $F(x)$ is only accessible with very low accuracy. These facts can be incorporated into the reconstruction process and we refer to [ABHS16, Bür16] for such approaches.

The kernel function depends on properties of the experimental setup. Essential influence have the thickness of the nonlinear material and the angle between the two beams arriving at it. Both parameters have to be measured as accurately as possible. The

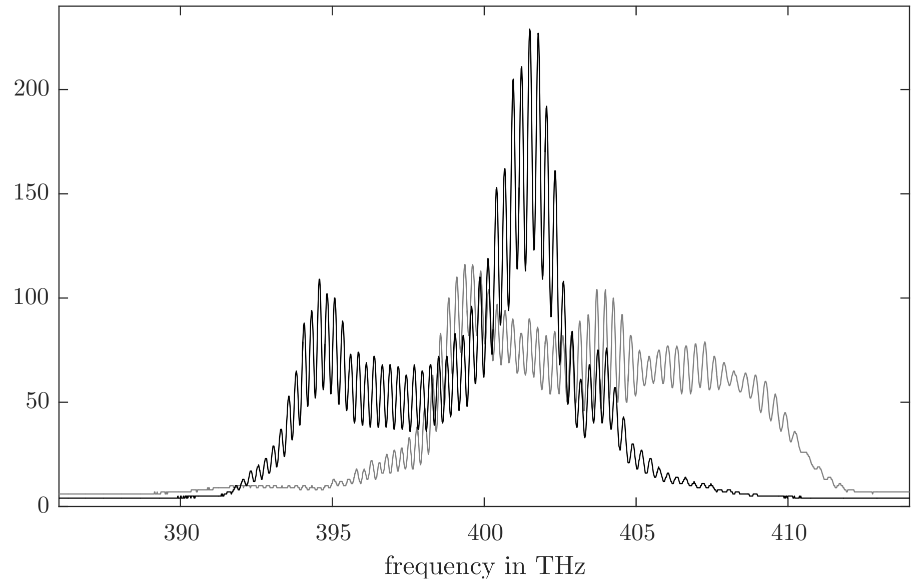

Figure 1.3.: SD-Interferograms with (black) and without (gray) nonlinear optical material recorded at Max Born Institute Berlin. The interferogram without nonlinear material is used for calculating the delay between the two pulses of a pair.

kernel always is bounded away from zero and it is bounded above. A typical kernel function is plotted in Figure 1.5. Details on the kernel and an explicit formula can be found in [Ger11a, Bür16].

Proposition 1.11. *Let $|k(s,t)| \leq c$ for all (s,t) in $\mathcal{D}(k)$. Then the mapping F defined by* (1.9) *is a quadratic mapping from $L^2_{\mathbb{C}}(0,1)$ into $L^2_{\mathbb{C}}(0,2)$ with $\|F\| \leq c$.*

Proof. The symmetric bilinear mapping B_F from Proposition 1.2 is

$$\big(B_F(x,u)\big)(s) = \int\limits_{\max\{0,\,s-1\}}^{\min\{s,\,1\}} k(s,t)\,\frac{x(s-t)\,u(t) + u(s-t)\,x(t)}{2}\,\mathrm{d}t, \quad s \in (0,2)$$

and can be rewritten as

$$\big(B_F(x,u)\big)(s) = \int\limits_{\max\{0,\,s-1\}}^{\min\{s,\,1\}} \frac{k(s,t) + k(s,s-t)}{2}\,x(s-t)\,u(t)\,\mathrm{d}t, \quad s \in (0,2).$$

Thus we have

$$\|B_F(x,u)\|^2 \leq c^2 \int\limits_0^2 \left(\int\limits_{\max\{0,\,s-1\}}^{\min\{s,\,1\}} |x(s-t)\,u(t)|\,\mathrm{d}t \right)^2 \mathrm{d}s,$$

which yields $\|B_F(x,u)\| \leq c\,\|x\|\,\|u\|$ as in the proof of Proposition 1.7. □

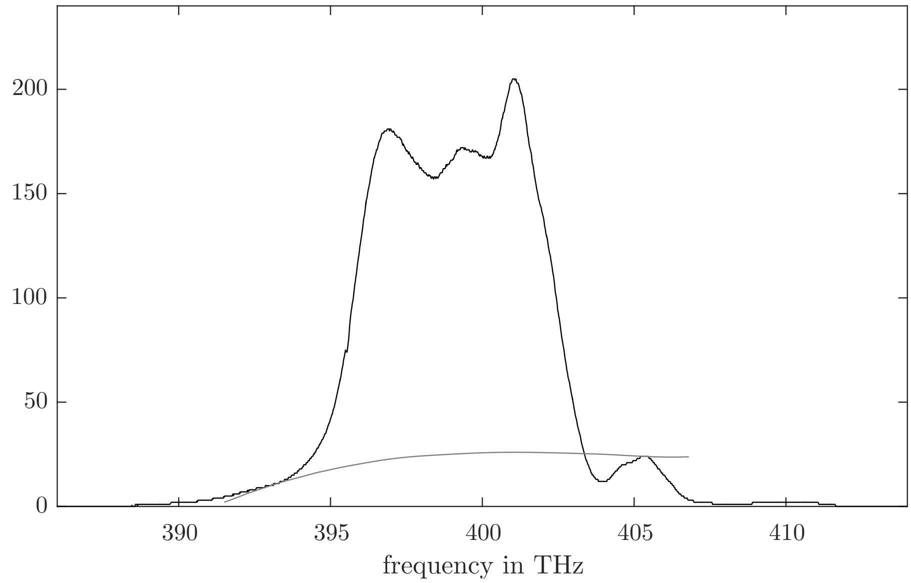

Figure 1.4.: Absolute value (black) and phase (gray) of right-hand side of the inverse problem (1.1). The phase is only sufficiently accurate where the absolute value is not close to zero. Thus only frequencies between 392 THz and 407 THz are of interest. This interval is rescaled to $(0, 1)$ for simplicity. Implementation of the preprocessing steps (Takeda algorithm) was done by Steven Bürger (TU Chemnitz).

In general the inverse problem with F defined in (1.9) is locally ill-posed everywhere, because for $k \equiv 1$ we obtain the kernelless autoconvolution problem discussed above. Fixing the amplitude and allowing only perturbations in the phase, locally well-posed situations can be observed, see [BH15, Proposition 3.3].

Analogously to the case $k \equiv 1$ the mapping F is weakly continuous and not compact, see [ABHS16, Propositions 1, 2].

1.2.3. Schlieren tomography

Ultrasound transducers are integral parts in medical diagnostics and treatment as well as in material testing and many other fields. Next to the wanted effects of the produced acoustic pressure also unwanted effects can occur (e.g. cavitation in medical applications). Therefore the shape and the intensity of the pressure distribution should be known as accurately as possible. But direct measurements at sufficiently many points inside the sonicated medium are too expensive and too time-consuming. Schlieren tomography is a widely applied and not too complex alternative to direct measurements, which allows to visualize the pressure distribution in fluids with high resolution and accuracy.

The principal construction of a Schlieren system is shown in Figure 1.6. A cylindrical tank filled with water is illuminated by parallel light. The ultrasound transducer, the pressure distribution of which shall be examined, is mounted at the top of the cylinder and at the bottom a sound absorbing material avoids reflections. Due to variations in

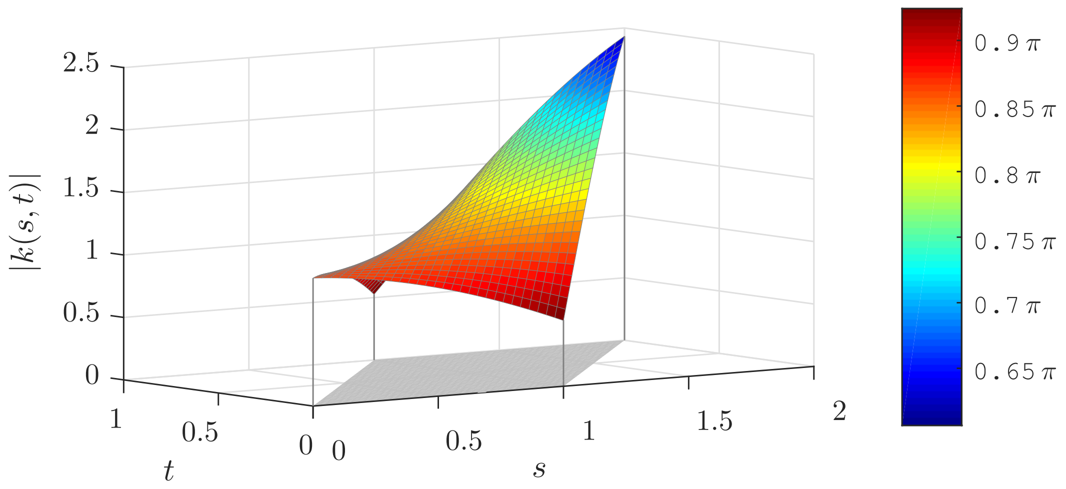

Figure 1.5.: A kernel function for (1.9) as it occurs in the SD-SPIDER method. The absolute value is given on the vertical axis and the phase is indicated by the color.

the water's density light is diffracted more or less. A lens focuses the diffracted light onto a screen and the undiffracted light is filtered out. The results are dark and light areas on the screen corresponding to negative and positive pressure regions along the light's path.

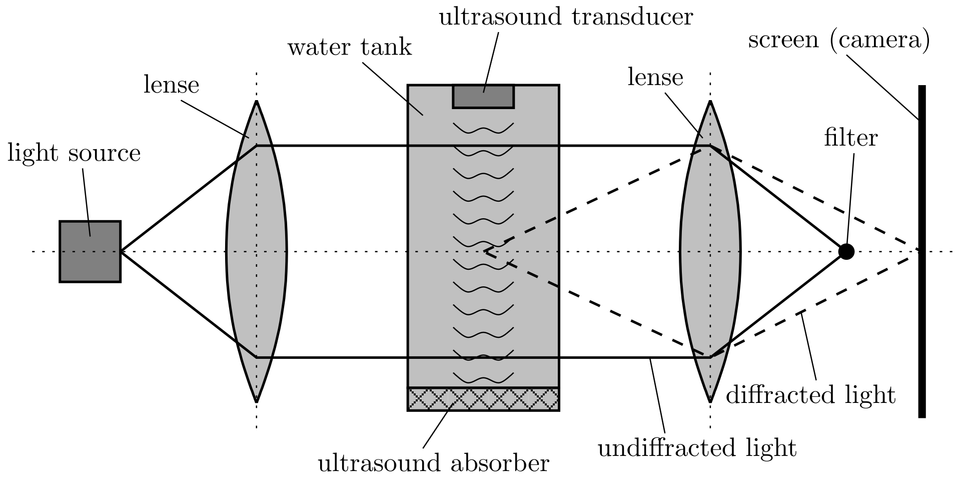

Figure 1.6.: The principal setup of a Schlieren imaging system.

Up to negligible side effects the intensity of light arriving at a point of the screen is proportional to the square of the integral over the pressure distribution along the corresponding ray. Thus, up to the square, the relation between pressure and observed image is the same as in X-ray imaging between density of the body and observed image. Taking images from many sides of the cylinder, tomographic reconstruction of the pressure distribution in space is possible. The corresponding mapping F in (1.1) is the pointwise square of the Radon transform.

To make things more precise we look at a horizontal slice

$$\Omega := \{(v,w) \in \mathbb{R}^2 : v^2 + w^2 < 1\}$$

of the cylinder and denote by

$$\mathbb{S}^1 := \{\sigma \in \mathbb{R}^2 : \sigma_1^2 + \sigma_2^2 = 1\}$$

the set of all directions for which Schlieren images are taken. Here, the screen is assumed to be in parallel to the chosen direction (σ and $-\sigma$ yield the same image, but for simplicity we do not exclude this doubling). By $\sigma^\perp := (-\sigma_2, \sigma_1)$ we denote the direction which is orthogonal to a given direction σ. Figure 1.7 shows a sketch of the setting.

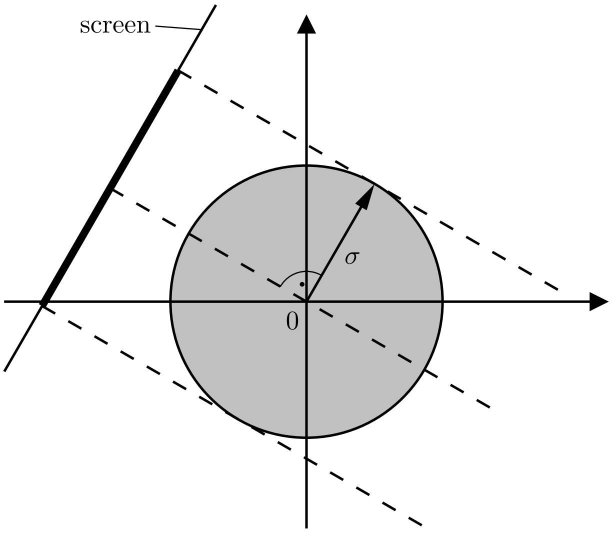

Figure 1.7.: Parametrization of two-dimensional slice through the cylinder.

With this notation at hand the Schlieren mapping is given by

$$\big(F(x)\big)(s,\sigma) := \left(\int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} x(s\,\sigma + t\,\sigma^\perp)\,\mathrm{d}t \right)^2, \quad s \in (-1,1),\ \sigma \in \mathbb{S}^1, \tag{1.10}$$

where x is real-valued and defined on Ω.

Proposition 1.12. *The mapping $F : L^2_{\mathbb{R}}(\Omega) \to L^1_{\mathbb{R}}\big((0,1) \times \mathbb{S}^1\big)$ defined by* (1.10) *is a quadratic mapping with $\|F\| \le 4\,\pi$.*

Proof. The underlying symmetric bilinear mapping B_F is given by

$$\big(B(x,u)\big)(s,\sigma) := \left(\int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} x(s\,\sigma + t\,\sigma^\perp)\,\mathrm{d}t \right) \left(\int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} u(s\,\sigma + t\,\sigma^\perp)\,\mathrm{d}t \right)$$

and we have

$$\|B_F(x,u)\| = \int\limits_{(-1,1)\times\mathbb{S}^1} \left| \int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} x(s\,\sigma + t\,\sigma^\perp)\,\mathrm{d}t \right| \left| \int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} u(s\,\sigma + t\,\sigma^\perp)\,\mathrm{d}t \right| \mathrm{d}(s,\sigma).$$

The Cauchy–Schwarz inequality applied to both inner integrals and then to the outer integral in combination with the estimate $\sqrt{1-s^2} \leq 1$ yields

$$\begin{aligned}
&\|B_F(x,u)\| \\
&\leq 2 \int\limits_{(-1,1)\times\mathbb{S}^1} \sqrt{\int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} x(s\,\sigma + t\,\sigma^\perp)^2\,\mathrm{d}t} \sqrt{\int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} u(s\,\sigma + t\,\sigma^\perp)^2\,\mathrm{d}t\,\mathrm{d}(s,\sigma)} \\
&\leq 2 \sqrt{\int\limits_{(-1,1)\times\mathbb{S}^1} \int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} x(s\,\sigma + t\,\sigma^\perp)^2\,\mathrm{d}t\,\mathrm{d}(s,\sigma) \int\limits_{(-1,1)\times\mathbb{S}^1} \int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} u(s\,\sigma + t\,\sigma^\perp)^2\,\mathrm{d}t\,\mathrm{d}(s,\sigma)}.
\end{aligned}$$

The first double integral reduces to

$$\begin{aligned}
\int\limits_{(-1,1)\times\mathbb{S}^1} \int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} x(s\,\sigma + t\,\sigma^\perp)^2\,\mathrm{d}t\,\mathrm{d}(s,\sigma) &= \int\limits_{\mathbb{S}^1} \int\limits_{-1}^{1} \int\limits_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} x(s\,\sigma + t\,\sigma^\perp)^2\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}\sigma \\
&= \int\limits_{\mathbb{S}^1} \|x\|^2\,\mathrm{d}\sigma = 2\,\pi\,\|x\|^2
\end{aligned}$$

and the second analogously to $2\,\pi\,\|u\|^2$. Thus, $\|B_F(x,u)\| \leq 4\,\pi$. □

1.3. Local versus global ill-posedness

For nonlinear mappings there is no generally acknowledged definition of the terms *well-posed* and *ill-posed*. The classical Hadamard definition requires injectivity, surjectivity and continuity of the inverse mapping. In modern regularization theory for nonlinear mappings existence of solutions is assumed, uniqueness is sometimes enforced by restriction to norm minimizing solutions or uniqueness is not required at all. The latter is the case here, because quadratic equations always have at least two (norm minimizing) solutions. The most often used definition of well-posedness seems to be the one in [HS98, Definition 1.1]. But that definition implies that solutions are isolated from each other. We prefer a definition which includes no assumptions on the shape of the solution set.

Definition 1.13. A mapping $F : X \to Y$ is *locally well-posed* at x_0 if for each convergent sequence $(y_k)_{k\in\mathbb{N}}$ in the range $\mathcal{R}(F)$ with limit $F(x_0)$ each sequence $(x_k)_{k\in\mathbb{N}}$ of preimages x_k from $F^{-1}(y_k)$ has a convergent subsequence and the corresponding limit belongs to $F^{-1}(F(x_0))$. Otherwise F is *locally ill-posed* at x_0.

For our purposes this definition is the right one, because we consider norm minimizing solutions, implying boundedness of the solution set. For the sake of completeness we mention that in case of unbounded solution sets the following definition is more suitable, since it allows to approximate the solution set with sequences which do not converge.

Definition 1.14. A mapping $F : X \to Y$ is *locally well-posed* at x_0 if for each convergent sequence $(y_k)_{k\in\mathbb{N}}$ in $\mathcal{R}(F)$ with limit $F(x_0)$ each sequence $(x_k)_{k\in\mathbb{N}}$ of preimages x_k from $F^{-1}(y_k)$ satisfies $\operatorname{dist}(x_n, F^{-1}(F(x_0))) \to 0$. Otherwise F is *locally ill-posed* at x_0.

In [LF12] different ill-posedness definitions for nonlinear mappings and their implications are discussed.

Quadratic mappings are nonlinear and ill-posedness properties may vary from point to point. On the other hand Proposition 1.5 shows, that information about local ill-posedness or well-posedness at each point is contained in an arbitrarily small ball around zero (or any other point) and thus should have some structure. As a result in this direction the following proposition states that ill-posedness or well-posedness of quadratic mappings does not vary on rays.

Proposition 1.15. *If a quadratic mapping F is locally well-posed (or ill-posed) at x_0 then it is locally well-posed (or ill-posed) at $t\,x_0$ for all t in $\mathbb{R} \setminus \{0\}$.*

Proof. Let F be locally well-posed at x_0. Denote by $(y_k)_{k\in\mathbb{N}}$ a sequence in $\mathcal{R}(F)$ converging to $F(t\,x_0)$ and let $(x_k)_{k\in\mathbb{N}}$ be a sequence in X with $F(x_k) = y_k$ for all k. Then

$$F\left(\frac{1}{t}\,x_k\right) = \frac{1}{t^2}\,y_k \to \frac{1}{t^2}\,F(t\,x_0) = F(x_0).$$

Local well-posedness at x_0 implies existence of a convergent subsequence of $(\frac{1}{t}\,x_k)_{k\in\mathbb{N}}$ and the corresponding limit x satisfies $F(x) = F(x_0)$. Denoting the subsequence again by $(\frac{1}{t}\,x_k)_{k\in\mathbb{N}}$ we obtain

$$F(x_k) = t^2\,F\left(\frac{1}{t}\,x_k\right) \to t^2\,F(x_0) = F(t\,x_0).$$

This shows local well-posedness at $t\,x_0$.

If F is locally ill-posed at x_0 but not locally ill-posed at some point $t\,x_0$, then the proof's first part would imply local well-posedness at x_0 (use $\tilde{x}_0 := t\,x_0$). Thus, the proof is complete. □

There are quadratic mappings which are everywhere locally well-posed, for instance strong quadratic isometries, see Section 3.1. There are also quadratic mappings which are everywhere locally ill-posed. This is for instance the case for autoconvolution of functions with uniformly bounded support presented in Subsection 1.2.1.

The problem, whether there exist quadratic mappings which are locally ill-posed at some point but locally well-posed at another point, remains unsolved. The author tends to the conjecture that such mappings do not exist. This would imply that ill-posedness is not a local but a global property of quadratic mappings.

1.4. Geometric properties of quadratic mappings' ranges

Here we collect some observations on the geometric structure of the range of a quadratic mapping. These observation will not be used in subsequent chapters, but may help to get some intuition about the behavior of quadratic mappings. As we will see, ellipses play a central role. Therefore we start with some remarks on the definition of ellipses.

In the plane $\mathbb{R}^2$ an *ellipse* with half-axis lengths a and b in standard form is the set of all points (x_1, x_2) such that

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1.$$

A *diameter* of an ellipse is a line segment between two points of the ellipse through the ellipse's center. Two diameters are called *conjugate diameters* if the one is in parallel with the tangents to the ellipse at the other's intersection points with the ellipse. Analogously *conjugate radii* can be defined.

Now let E be an ellipse in the plane $\mathbb{R}^2$ with center (x_1, x_2) and with conjugate radii represented by the vectors $[r_1, r_2]^{\mathrm{T}}$ and $[s_1, s_2]^{\mathrm{T}}$. Then

$$E = \{x + \alpha\, r + \beta\, s : \, \alpha^2 + \beta^2 = 1\}.$$

If, on the other hand, a set E can be represented in this form, then E is an ellipse with corresponding center and conjugate radii. In analogy to convex combinations, we call a linear combination, the squared coefficients of which add to one, *elliptic combination.*

The same way ellipses can be described as elliptic combinations of two conjugate radii, n-dimensional ellipsoids are determined by there center and n conjugate radii (tangents in the definition of conjugate radii have to be replaced by tangent planes). With this technique finite-dimensional ellipsoids can be defined in infinite-dimensional spaces, too. Given three elements x, r, s in a Hilbert space X, by

$$\operatorname{ell}(x, r, s) := \{x + \alpha\, r + \beta\, s : \, \alpha^2 + \beta^2 = 1\}$$

we denote the ellipse with center x and conjugate radii r, s.

In the following X and Y are real Hilbert spaces and $F : X \to Y$ is a quadratic mapping.

Proposition 1.16. *Each quadratic mapping maps two-dimensional subspaces to elliptic cones. More precisely,*

$$F\big(\operatorname{span}\{x, u\}\big) = \bigcup_{t \geq 0} \left(t \operatorname{ell}\left(\frac{F(x) + F(u)}{2}, \frac{F(x) - F(u)}{2}, B_F(x, u) \right) \right).$$

Proof. For coefficients a, b with $a^2 + b^2 = 1$ we have

$$\begin{aligned} F(a\, x + b\, u) &= a^2\, F(x) + 2\, a\, b\, B_F(x, u) + b^2\, F(u) \\ &= \frac{a^2 + b^2}{2}\, (F(x) + F(u)) + \frac{a^2 - b^2}{2}\, (F(x) - F(u)) + 2\, a\, b\, B_F(x, u) \\ &= \frac{1}{2}\, (F(x) + F(u)) + \frac{2\, a^2 - 1}{2}\, (F(x) - F(u)) + 2\, a\, b\, B_F(x, u). \end{aligned}$$

Because

$$(2\,a^2-1)^2+(2\,a\,b)^2=4\,a^4-4\,a^2+1+4\,a^2\,(1-a^2)=1,$$

we see that $\mathrm{ell}(0,x,u)$ is mapped to $\mathrm{ell}\bigl(\frac{F(x)+F(u)}{2},\frac{F(x)-F(u)}{2},B_F(x,u)\bigr)$. The observation

$$\operatorname{span}\{x,u\}=\bigcup_{t\geq 0}\bigl(t\;\mathrm{ell}(0,x,u)\bigr)$$

completes the proof. □

From the proof we immediately see that ellipses centered at the origin are mapped to ellipses centered somewhere else. In particular, intersections of the unit sphere and two-dimensional subspaces are mapped to ellipses. This observation instigates the idea that intersections of the unit sphere and n-dimensional subspaces are mapped to ellipsoids. But this is not the case if $n>2$. Figure 1.8 shows the image of the unit sphere in three-dimensional space under the quadratic mapping $F:\mathbb{R}^3\to\mathbb{R}^3$ defined by

$$F(x):=\bigl(x_1^2+\sqrt{2}\,x_2\,x_3,\;x_2^2+\sqrt{2}\,x_1\,x_3,\;x_3^2+\sqrt{2}\,x_1\,x_2\bigr),\quad x\in\mathbb{R}^3.$$

This is not an ellipsoid.

Nevertheless, the structure of the unit sphere's image under quadratic mappings can be described with the help of ellipses. Note, that the image of the whole space then is the cone spanned by this image set.

Proposition 1.17. *Let $(e_k)_{k\in\mathbb{N}}$ be an orthonormal basis in X and denote by S_{n-1} the intersection of the unit sphere in X with $\operatorname{span}\{e_1,\ldots,e_n\}$. Then $F(S_1)$ is an ellipse and*

$$F(S_{n-1})=\bigcup_{x\in S_{n-1}}\mathrm{ell}\left(\frac{F(e_n)+F(x)}{2},\frac{F(e_n)-F(x)}{2},B_F(e_n,x)\right)$$

for $n>1$.

Proof. That $F(S_1)$ is an ellipse follows from the proof of Proposition 1.16. For each element in S_{n-1} there is some x in S_{n-2} such that the element is contained in $\mathrm{ell}(0,e_n,x)$. Thus,

$$F(S_{n-1})=\bigcup_{x\in S_{n-2}}F\bigl(\mathrm{ell}(0,e_n,x)\bigr)$$

and the assertion follows as in the proof of Proposition 1.16. □

1.5. Literature on quadratic mappings

There is only very few literature on quadratic mappings, especially on quadratic mappings in infinite-dimensional spaces. Most of the publications focus on autoconvolutions. In this section we provide a brief overview of the literature relevant for the first part of this thesis.

The core material this thesis builds up on are publications on de-autoconvolution as an inverse problem. There was a first accumulation in the 1990s [GH94, FH96, Jan97], followed by the articles [Jan00, Ram03, DL08]. A second accumulation started in the

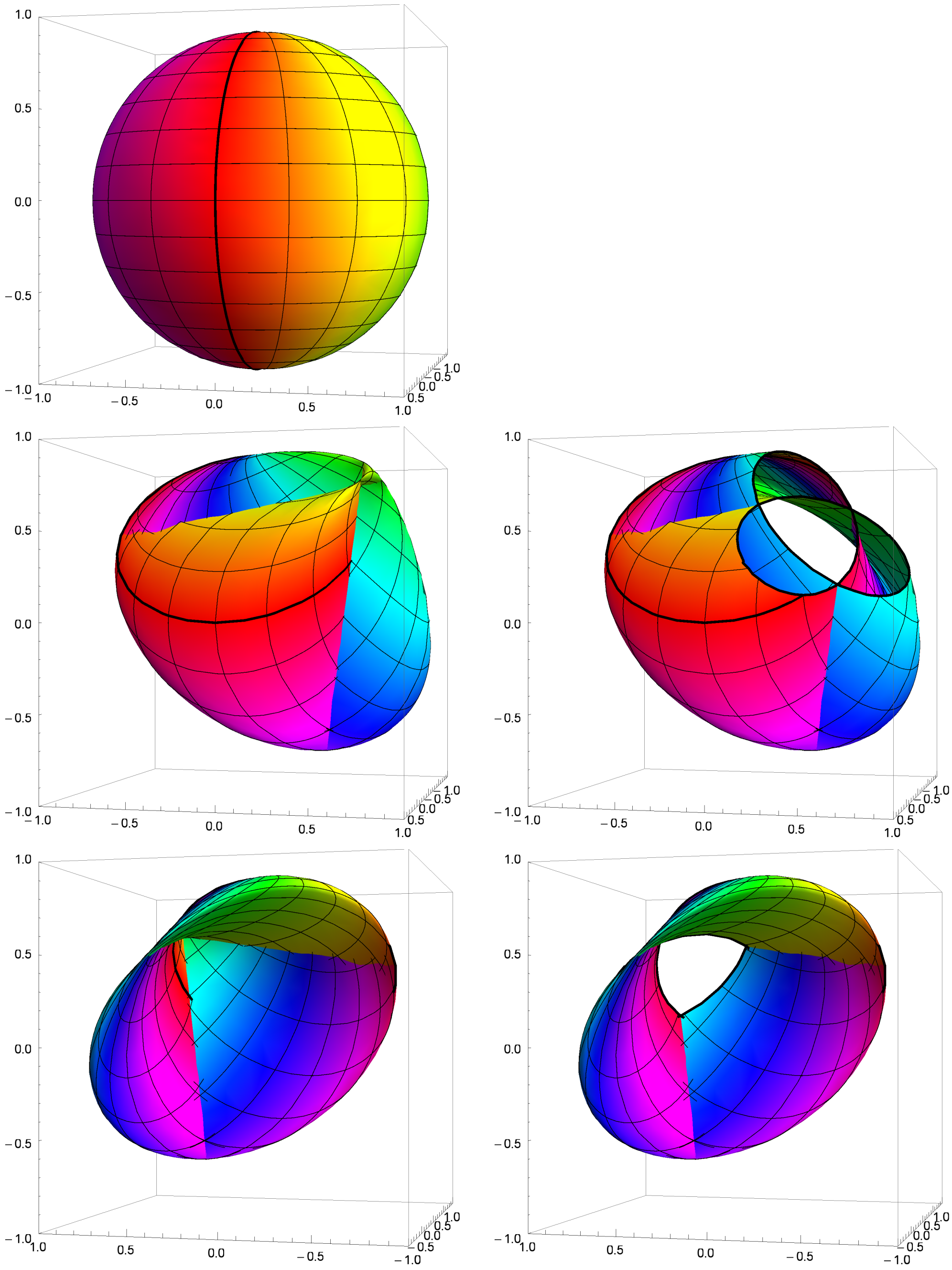

Figure 1.8.: Image of the unit sphere in $\mathbb{R}^3$ under the quadratic mapping $(x_1, x_2, x_3) \mapsto (x_1^2 + \sqrt{2}\,x_2\,x_3,\ x_2^2 + \sqrt{2}\,x_1\,x_3,\ x_3^2 + \sqrt{2}\,x_1\,x_2)$. Upper left: unit sphere, middle left: full image, lower left: same as middle left but seen from the opposite direction, middle right: same as middle left but with the cap cut off, lower right: same as lower left but with the cap cut off.

past five years with [Ger11a, GHB$^+$14, Fle14, Bür14, BSK$^+$15, BF15, BH15, ABHS16, Bür16, BFH16, BM16] and this thesis.

In the engineering literature several practical methods for de-autoconvolution in finite-dimensional settings are discussed. Some articles of this type can be found in the reference list of [Bau91]. There are also relatively old works on convexity properties of the range of quadratic mappings in finite dimensions, see [Toe18, Hau19, Din41]. For extensions of those results see [She13, Xia14] and references therein.

In principle a quadratic mapping can be regarded as a special case of tensors. Thus, results about tensors apply to some extend also to quadratic mappings. We mention [KB09, GER11b] here, where singular value decompositions for tensors in finite dimensions are discussed. Although such concepts could be useful for solving quadratic inverse problems, closer inspection and numerical tests lead to the decision to not follow this path.

2. Tikhonov regularization

A simple but effective regularization method to approximate solutions of ill-posed nonlinear equations is the method of Tikhonov. We restrict our attention to separable Hilbert spaces X and Y. Then Tikhonov's method for quadratic equations (1.1) consists in minimizing the Tikhonov functional

$$T_\alpha(x, y) := \|F(x) - y\|^2 + \alpha \, \|x - \bar{x}\|^2, \quad x \in X, \; y \in Y$$

with respect to x. The element y is an approximation to the exact right-hand side $y^\dagger$ of (1.1), the reference element $\bar{x}$ in X deals as initial guess of the exact solution and the positive regularization parameter α controls the trade-off between data fitting and stabilization.

Classical results in [EKN89] on existence, stability and convergence of Tikhonov minimizers are based an the assumption that the mapping F is sequentially weak-to-weak continuous. In general, quadratic mappings do not have this property. An example is the mapping $F : X \to \mathbb{R}$ defined by $F(x) = \|x\|^2$. For an orthonormal basis $(e_k)_{k\in\mathbb{N}}$ we have $F(e_k) = 1$ although the sequence (e_k) converges weakly to zero.

Verification of weak-to-weak continuity reduces to its verification at zero.

Proposition 2.1. *A quadratic mapping is sequentially weak-to-weak continuous on X if and only if it is sequentially weak-to-weak continuous at zero.*

Proof. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in X converging weakly to some x in X. The mapping F is weak-to-weak continuous if for each such sequence we have $\langle \eta, F(x_k) - F(x)\rangle \to 0$ for all η in Y. Rewriting $F(x_k) - F(x) = F(x_k - x) + 2\, B_F(x_k - x, x)$ and using the notion of adjoint bilinear mappings defined by (1.4) we see

$$\begin{aligned}\langle \eta, F(x_k) - F(x)\rangle &= \langle \eta, F(x_k - x)\rangle + 2\,\langle \eta, B_F(x_k - x, x)\rangle \\ &= \langle \eta, F(x_k - x)\rangle + 2\,\langle B_F^*(x, \eta), x_k - x)\rangle.\end{aligned}$$

Thus, $\langle \eta, F(x_k) - F(x)\rangle \to 0$ if and only if $\langle \eta, F(x_k - x)\rangle \to 0$. In other words, F is weak-to-weak continuous if and only if

$$x_k - x \rightharpoonup 0 \quad \Rightarrow \quad F(x_k - x) \rightharpoonup 0,$$

which by $F(0) = 0$ is simply the definition of weak-to-weak continuity at zero. □

All examples of quadratic mappings provided in Section 1.2 are weak-to-weak continuous (see provided references there) and hence Tikhonov regularization is a stable and convergent approximation method for the solutions of (1.1).

Remark 2.2. The problem of non-injectivity of quadratic mappings carries over to the Tikhonov functional. If the reference element $\bar{x}$ is zero, we obviously have that $T_\alpha(x, y) = T_\alpha(-x, y)$ for all x and y. This issue cannot be solved by choosing $\bar{x} \neq 0$, but only slightly tempered. For all x with $\langle x, \bar{x} \rangle = 0$ we still have $T_\alpha(x, y) = T_\alpha(-x, y)$. Thus, we have to expect multiple Tikhonov minimizers. Analogously, there might by multiple norm minimizing solutions to (1.1) (i.e., solutions with minimal distance to $\bar{x}$), even if $\bar{x} \neq 0$.

For $\bar{x} = 0$ we want to mention the useful observation that the range of sensible regularization parameters is bounded above. This behavior is typical only for Tikhonov regularization with quadratic mappings and for sparsity promoting regularization with linear mappings (cf. Proposition 6.7).

Proposition 2.3. *Let $y \in Y$ and set*

$$\alpha_{\max} := 2 \sup_{\substack{x \in X \\ \|x\| \leq 1}} \operatorname{Re} \langle F(x), y \rangle.$$

If $\alpha \geq \alpha_{\max}$, then

$$0 \in \operatorname*{argmin}_{x \in X} T_\alpha(x, y).$$

If, in addition, $\alpha > \alpha_{\max}$ or F is injective up to sign, then

$$\operatorname*{argmin}_{x \in X} T_\alpha(x, y) = \{0\}.$$

Proof. If $x \neq 0$ we have

$$\begin{aligned} T_\alpha(x, y) &= \|F(x)\|^2 - 2 \operatorname{Re} \langle F(x), y \rangle + \|y\|^2 + \alpha \|x\|^2 \\ &\geq \|F(x)\|^2 - 2 \operatorname{Re} \langle F(x), y \rangle + \|y\|^2 + 2 \operatorname{Re} \left\langle F\left(\frac{x}{\|x\|}\right), y \right\rangle \|x\|^2 \\ &= \|F(x)\|^2 + \|y\|^2 \geq \|y\|^2 = T_\alpha(0, y), \end{aligned}$$

proving the first assertion. If $\alpha > \alpha_{\max}$, the first inequality sign is strict. If F is injective up to sign, $x \neq 0$ implies $F(x) \neq 0$, making the second inequality sign a strict one. □

Remark 2.4. If $\alpha_{\max}$ is chosen greater than in the proposition, the proposition remains true. An easy to calculate replacement is $2 \|F\| \|y\|$.

Theory for Tikhonov regularization with nonlinear mappings is well developed. But in practice this method suffers from the need for global minimizers. For quadratic mappings the Tikhonov functional is not convex, which makes numerical minimization challenging. The only available tailor-made algorithm for our concrete minimization problem is the TIGRA method proposed in [Ram03].

3. Regularization by decomposition

In this chapter we propose a regularization method which splits the ill-posed quadratic problem (1.1) in real or complex separable Hilbert spaces X and Y into an ill-posed linear problem and a well-posed quadratic one. The technique is based on the notion of quadratic isometries, which we introduce in the first section of this chapter. Then we present the decomposition approach and its application in a regularization method. Numerical tests complete the chapter.

The results presented in this chapter were published in [Fle14, BF15], but only real Hilbert spaces were considered there.

3.1. Quadratic isometries

It is well known that a linear operator preserves inner products if and only if it preserves norms. In the quadratic case there are (strong) isometries, which preserve both inner products and norms, and there are (weak) isometries, which only preserve norms.

Definition 3.1. A quadratic mapping F is a *strong isometry* if

$$\langle F(x), F(u)\rangle = \langle x, u\rangle^2$$

for all x and u from X and a *weak isometry* if

$$\|F(x)\| = \|x\|^2$$

for all x.

Obviously, each strong isometry is also weak. The following example shows that there are weak quadratic isometries which are not strong.

Example 3.2. Define $F : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$F(x) := \begin{bmatrix} x_1^2 - x_2^2 \\ 2\,x_1\,x_2 \end{bmatrix}.$$

Then $\|F(x)\|^2 = (x_1^2 - x_2^2)^2 + 4\,x_1^2\,x_2^2 = \|x\|^2$ for all x, but

$$\left\langle F\left(\begin{bmatrix}1\\0\end{bmatrix}\right), F\left(\begin{bmatrix}0\\1\end{bmatrix}\right)\right\rangle = -1 \neq 0 = \left\langle \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}\right\rangle^2.$$

Thus, F is a weak isometry but not a strong one. □

An example of a strong quadratic isometry in infinite-dimensional spaces will be given in Section 3.2. Figure 3.1 visualizes a strong quadratic isometry mapping between $\mathbb{R}^2$ and $\mathbb{R}^3$.

For checking isometric properties of quadratic mappings we provide the following criterion.

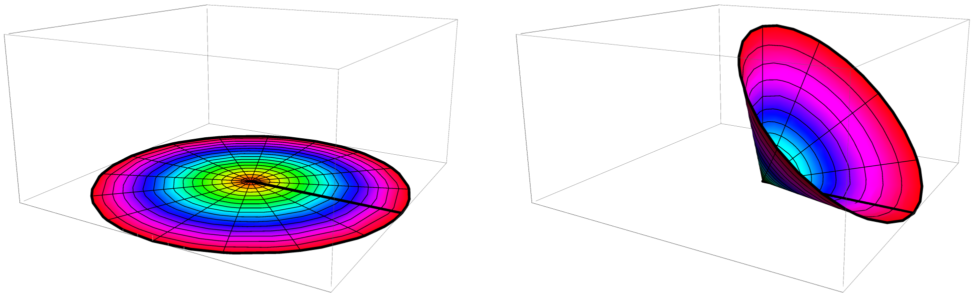

Figure 3.1.: Example of a strong quadratic isometry acting between $\mathbb{R}^2$ and $\mathbb{R}^3$. The element (x_1, x_2) is mapped to $(x_1^2, \sqrt{2}\, x_1\, x_2, x_2^2)$. Left-hand side: unit disc in $\mathbb{R}^2$. Right-hand side: image set of unit disc.

Proposition 3.3. *Let $(e_i)_{i\in\mathbb{N}}$ be an orthonormal basis in X. A quadratic mapping $F : X \to Y$ is a strong isometry if and only if the following two conditions hold:*

(i) $\|B_F(e_i, e_j)\| = \begin{cases} 1, & \textit{if } j = i, \\ \frac{1}{\sqrt{2}}, & j < i. \end{cases}$

(ii) The set $\{B_F(e_i, e_j) : i \in \mathbb{N},\, j \le i\}$ is an orthogonal system.

Proof. Necessity follows from calculation of $\langle B_F(e_i, e_j), B_F(e_k, e_l)\rangle$. With (1.2) we obtain

$$\begin{aligned}\langle B_F(e_i, e_j), B_F(e_k, e_l)\rangle &= \frac{1}{2}\,\langle e_i, e_k\rangle\,\langle e_j, e_l\rangle + \frac{1}{2}\,\langle e_i, e_l\rangle\,\langle e_j, e_k\rangle \\ &= \begin{cases} 1, & \text{if } i = j = k = l, \\ \frac{1}{2}, & \text{if } i = k \neq l = j \text{ or } i = l \neq k = j, \\ 0, & \text{else,} \end{cases}\end{aligned}$$

which directly yields the two conditions in the proposition.

For sufficiency we observe

$$\begin{aligned}&\left\langle F\left(\sum_{i=1}^{\infty} x_i\, e_i\right), F\left(\sum_{k=1}^{\infty} u_k\, e_k\right)\right\rangle \\ &\quad = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} x_i\, x_j\, \overline{u}_k\, \overline{u}_l\, \langle B_F(e_i, e_j), B_F(e_k, e_l)\rangle \\ &\quad = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i\, x_j\, \overline{u}_i\, \overline{u}_j = \left(\sum_{i=1}^{\infty} x_i\, \overline{u}_i\right)^2 = \left\langle \sum_{i=1}^{\infty} x_i\, e_i, \sum_{k=1}^{\infty} u_k\, e_k\right\rangle^2 .\end{aligned}$$

□

As one might expect from an isometry, each strong quadratic isometry is continuously invertible. Remember that quadratic mappings cannot be injective because $F(x) = F(-x)$ for all x. Thus, we have to use a slightly generalized notion of continuous invertibility. In view of Definition 1.13 strong quadratic isometries always are locally well-posed at each point.

Proposition 3.4. *Let F be a strong quadratic isometry and denote by $F^{-1}(y)$ the full preimage of F at some point y. If a sequence $(y_k)_{k\in\mathbb{N}}$ in Y converges to some y in Y and if $(x_k)_{k\in\mathbb{N}}$ is a sequence of corresponding preimages x_k from $F^{-1}(y_k)$ and x is a preimage of y, then $(x_k)_{k\in\mathbb{N}}$ converges to x or $-x$ or it decomposes into two subsequences $(x_k^+)_{k\in\mathbb{N}}$ and $(x_k^-)_{k\in\mathbb{N}}$ such that $x_k^+ \to x$ and $x_k^- \to -x$.*

Proof. Define index sets

$$I^+ := \{k \in \mathbb{N} : \operatorname{Re}\langle x_k, x\rangle \geq 0\} \quad \text{and} \quad I^- := \{k \in \mathbb{N} : \operatorname{Re}\langle x_k, x\rangle < 0\}.$$

If (x_k) does not converge to x or $-x$ both sets have infinitely many elements. Then $(x_k^+)_{k\in\mathbb{N}}$ is the subsequence $(x_k)_{k\in I^+}$ and $(x_k^-)_{k\in\mathbb{N}}$ is the subsequence $(x_k)_{k\in I^-}$.

Since F is a strong isometry we have

$$\|x_k^+ - x\|^2 = \|x_k^+\|^2 - 2\operatorname{Re}\langle x_k^+, x\rangle + \|x\|^2 = \|y_k\| - 2\left|\operatorname{Re}\sqrt{\langle y_k, y\rangle}\right| + \|y\|$$

where $\left|\operatorname{Re}\sqrt{\langle y_k, y\rangle}\right|$ yields the same value for both complex roots. The first summand converges to $\|y\|$ and the second to $-2\|y\|$. Thus,

$$\|x_k^+ - x\|^2 \to 0.$$

Analogously, we obtain

$$\|x_k^- - (-x)\|^2 = \|x_k^-\|^2 + 2\operatorname{Re}\langle x_k^-, x\rangle + \|x\|^2 = \|y_k\| - 2\left|\operatorname{Re}\sqrt{\langle y_k, y\rangle}\right| + \|y\| \to 0.$$

The second equality follows from $\langle x_k^-, x\rangle^2 = \langle y_k, y\rangle$ and $\operatorname{Re}\langle x_k^-, x\rangle < 0$. □

Remark 3.5. From the proposition we immediately see that strong isometries are injective up to sign. If $F(u) = F(x)$, choose $y = F(x)$, $y_k = F(u)$ and $x_k = u$ in the proposition to obtain $u = x$ or $u = -x$.

For linear mappings continuous invertibility is equivalent to closedness of the range. Since this in general is not true for nonlinear mappings we now prove the following result.

Proposition 3.6. *Let F be a weak quadratic isometry which is weak-to-weak continuous. Then its range $\mathcal{R}(F)$ is closed.*

Proof. Let $(F(x_k))_{k\in\mathbb{N}}$ be a sequence in $\mathcal{R}(F)$ which converges to some y in Y. Then

$$\|x_k\|^2 = \|F(x_k)\| \to \|y\|.$$

Thus, there is a weakly convergent subsequence $(x_{k_l})_{l\in\mathbb{N}}$ with limit x and weak-to-weak continuity of F implies

$$F(x_{k_l}) \rightharpoonup F(x).$$

Together with $F(x_{k_l}) \rightharpoonup y$ we obtain $y = F(x)$. □

If we look at strong quadratic isometries F and replace the image space Y by the closed subspace $\overline{\operatorname{span}\mathcal{R}(F)}$, then F is always weak-to-weak continuous and, thus, has closed range.

Proposition 3.7. *Each strong quadratic isometry $F : X \to Y$ is weak-to-weak continuous as a mapping between the Hilbert spaces X and $\overline{\operatorname{span}\mathcal{R}(F)}$.*

Proof. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in X converging weakly to zero. Then for all x in X we have

$$\langle F(x_k), F(x)\rangle = \langle x_k, x\rangle^2 \to 0,$$

that is, $\langle F(x_k), y\rangle \to 0$ for all y in $\mathcal{R}(F)$. Consequently this also holds for all y in $\operatorname{span}\mathcal{R}(F)$.

Now let y be an element from $\overline{\operatorname{span}\mathcal{R}(F)}$ and let $(y_l)_{l\in\mathbb{N}}$ be a sequence in $\operatorname{span}\mathcal{R}(F)$ converging to y. Then

$$|\langle F(x_k), y\rangle| \leq \|F(x_k)\|\,\|y - y_l\| + |\langle F(x_k), y_l\rangle|$$

for all l and all k. Weak convergence of (x_k) implies boundedness of (x_k) and thus $\|F(x_k)\| \leq c$ for all k with some $c > 0$. For arbitrary positive ε we may choose $\bar{l}$ such that $\|y - y_{\bar{l}}\| \leq \frac{\varepsilon}{2\,c}$ and $\bar{k}$ such that $|\langle F(x_k), y_{\bar{l}}\rangle| \leq \frac{\varepsilon}{2}$ for all $k \geq \bar{k}$. This shows $|\langle F(x_k), y\rangle| \leq \varepsilon$ for $k \geq \bar{k}$.

We deduce weak convergence $F(x_k) \rightharpoonup 0$ if F is considered as a mapping between X and $\overline{\operatorname{span}\mathcal{R}(F)}$, which proves weak-to-weak continuity of F at zero. Proposition 2.1 yields weak-to-weak continuity on X. □

3.2. Decomposition of quadratic mappings

Bounded linear operators can be decomposed into a partial isometry and a selfadjoint operator. In a similar spirit we suggest the decomposition of quadratic mappings into a quadratic isometry and a linear operator.

Theorem 3.8. *Each quadratic mapping F can be decomposed into a strong quadratic isometry $Q : X \to \ell^2(\mathbb{N})$ and a densely defined linear operator $A : \ell^2(\mathbb{N}) \to Y$ such that*

$$F(x) = A\,Q(x) \tag{3.1}$$

for all x in X.

The proof is constructive and will be given in the following. The two lemmas provide a possible choice of the quadratic part Q and the linear part A. But as we will discuss below, other choices are possible and maybe advantageous.

For easier handling of indices we define the mapping

$$\kappa : \{(i,j) \in \mathbb{N}\times\mathbb{N} : 1 \leq j \leq i\} \to \mathbb{N}$$

by

$$\kappa(i,j) := j + \frac{i\,(i-1)}{2}. \tag{3.2}$$

This is a bijection.

Lemma 3.9. *Let $(e_i)_{i\in\mathbb{N}}$ be an orthonormal basis of X. The mapping $Q : X \to \ell^2(\mathbb{N})$ defined by*

$$\big(Q(x)\big)_{\kappa(i,j)} := \begin{cases} \sqrt{2}\,\langle x, e_i\rangle\,\langle x, e_j\rangle, & \text{if } j < i, \\ \langle x, e_i\rangle^2, & \text{if } j = i \end{cases} \tag{3.3}$$

for (i, j) in $\mathbb{N} \times \mathbb{N}$ with $1 \leq j \leq i$ and x in X is a strong quadratic isometry.

Proof. The underlying symmetric bilinear mapping of Q is given by

$$\big(B_Q(x, u)\big)_{\kappa(i,j)} := \begin{cases} \frac{1}{\sqrt{2}}\big(\langle x, e_i\rangle\,\langle u, e_j\rangle + \langle x, e_j\rangle\,\langle u, e_i\rangle\big), & \text{if } j < i, \\ \langle x, e_i\rangle\,\langle u, e_i\rangle, & \text{if } j = i. \end{cases} \tag{3.4}$$

Thus, $B_Q(e_i, e_j)$ is one or $\frac{1}{\sqrt{2}}$ at position $\kappa(i, j)$ if $i = j$ or $i \neq j$, respectively, and zero at all other positions. The assertion of the lemma now follows from Proposition 3.3. □

Lemma 3.10. *Let $(e_i)_{i\in\mathbb{N}}$ be an orthonormal basis of X and let F be quadratic. Denote by $\mathcal{D}(A) \subseteq \ell^2(\mathbb{N})$ the set of all z in $\ell^2(\mathbb{N})$ for which*

$$A\,z := \sum_{i=1}^{\infty}\left(\sum_{j=1}^{i-1}\sqrt{2}\,z_{\kappa(i,j)}\,B_F(e_i, e_j) + z_{\kappa(i,i)}\,B_F(e_i, e_i)\right) \tag{3.5}$$

converges. Then the corresponding mapping $A : \mathcal{D}(A) \to Y$ is linear and its domain $\mathcal{D}(A)$ is dense in $\ell^2(\mathbb{N})$.

Proof. Linearity is obvious and since B_F is bounded we have $\|B_F(e_i, e_j)\| \leq \|F\|$ for all i and j. Thus, the dense subspace $\ell^1(\mathbb{N})$ of $\ell^2(\mathbb{N})$ belongs to the domain of A. □

Now the proof of the main theorem is quite simple.

Proof of Theorem 3.8. With Q from (3.3) and A from (3.5) we have $F(x) = A\,Q(x)$ for all x in X. Since $F(x)$ is defined for each x, in particular we see that the range of Q belongs to the domain of A. □

The mapping A in the decomposition may be unbounded, even if F is injective up to sign as the following example shows. In Section 3.5 we discuss a setting with bounded A.

Example 3.11. This example is based on an idea by Steven Bürger (TU Chemnitz, November 2016), which has not been published elsewhere. Let $X = Y = \ell^2_{\mathbb{R}}(\mathbb{N})$ and define F by

$$[F(x)]_1 := \|x\|^2 \qquad \text{and} \qquad [F(x)]_{1+k} := [Q(x)]_k, \quad k \in \mathbb{N},$$

for all $x \in \ell^2_{\mathbb{R}}(\mathbb{N})$ with Q from (3.3). This mapping is injective up to sign because Q is injective up to sign (cf. Remark 3.5) and $A : \ell^2_{\mathbb{R}}(\mathbb{N}) \to \ell^2_{\mathbb{R}}(\mathbb{N})$ is given by

$$[A\,z]_1 = \sum_{k=1}^{\infty} z_{\kappa(k,k)} \qquad \text{and} \qquad [A\,z]_{1+k} = z_k, \quad k \in \mathbb{N}.$$

If we choose $z^{(n)}$ in $\ell^2_{\mathbb{R}}(\mathbb{N})$ with

$$z^{(n)}_{\kappa(k,l)} := \begin{cases} \frac{1}{\sqrt{n}}, & \text{if } k = l \text{ and } k \leq n, \\ 0, & \text{else,} \end{cases}$$

then $\|z^{(n)}\| = 1$ and

$$\|A\, z^{(n)}\|^2 \geq \left|[A\, z^{(n)}]_1\right|^2 = n.$$

Thus, $\|A\, z^{(n)}\| \to \infty$ for $n \to \infty$, which proves unboundedness of A. □

The constructed decomposition in the proof of the theorem is not the only one. Choosing another quadratic isometry Q one can improve the properties of A. If, for instance, A in the proof is injective and bounded, we can write it as $A = \tilde{A}\, U$ with selfadjoint $\tilde{A} : Y \to Y$ and a linear isometry $U : \ell^2(\mathbb{N}) \to Y$. This follows from the polar decomposition of the adjoint A^*. Then $\tilde{Q} := U\, Q$ is again a strong quadratic isometry and $F = \tilde{A}\, \tilde{Q}$.

The following proposition states that the converse is also true: each strong quadratic isometry is the composition of a linear isometry and Q from (3.3), with the underlying basis $(e_i)_{i\in\mathbb{N}}$ chosen arbitrarily. Note that restriction to the image space $\ell^2(\mathbb{N})$ comes from the context of the present section, but from the proof we immediately see that $\ell^2(\mathbb{N})$ can be replaced by any separable Hilbert space.

Proposition 3.12. *Let $\tilde{Q} : X \to \ell^2(\mathbb{N})$ be a strong quadratic isometry and let $(e_i)_{i\in\mathbb{N}}$ be an orthonormal basis in X. Then there is a linear isometry $U : \overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}} \to \ell^2(\mathbb{N})$ such that the strong quadratic isometry $U\,\tilde{Q} : X \to \ell^2(\mathbb{N})$ attains the form of Q in* (3.3).

Proof. By Proposition 3.3 the set

$$\left\{\tilde{Q}(e_i) : i \in \mathbb{N}\right\} \cup \left\{\sqrt{2}\, B_{\tilde{Q}}(e_i, e_j) : i, j \in \mathbb{N},\, j < i\right\}$$

is an orthonormal basis in $\overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}}$. Let $(f_i)_{i\in\mathbb{N}}$ be the standard orthonormal basis in $\ell^2(\mathbb{N})$ and define the linear mapping $U : \overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}} \to \ell^2(\mathbb{N})$ by

$$U\,\tilde{Q}(e_i) := f_{\kappa(i,i)} \qquad \text{and} \qquad U\, B_{\tilde{Q}}(e_i, e_j) := f_{\kappa(i,j)}$$

for all i and j with $j < i$ (the index map κ is defined in (3.2)). Then U transfers an orthonormal basis to an orthonormal basis and thus U is an isometry. From

$$U\,\tilde{Q}(x) = \sum_{i=1}^{\infty} \left(\sum_{j=1}^{i-1} \sqrt{2}\, \langle x, e_i\rangle\, \langle x, e_j\rangle\, f_{\kappa(i,j)} + \langle x, e_i\rangle^2\, f_{\kappa(i,i)} \right) \qquad \text{for all } x \text{ in } X$$

we see that $U\,\tilde{Q}$ attains the form of Q in (3.3). □

Regarding different choices of Q in the decomposition (3.1) the question arises whether boundedness of corresponding operators A depends on the choice of Q. The answer is ‘No’.

Proposition 3.13. *Let $Q : X \to \ell^2(\mathbb{N})$ be as in* (3.3) *and let $\tilde{Q} : X \to \ell^2(\mathbb{N})$ be a second strong quadratic isometry. Further let $A, \tilde{A} : \ell^2(\mathbb{N}) \to Y$ be two linear mappings such that $\tilde{A}\,\tilde{Q} = A\,Q$. Then $\tilde{A}$ is bounded on $\overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}}$ if and only if A is bounded on $\ell^2(\mathbb{N})$.*

Proof. By Proposition 3.12 there is a linear isometry $U : \overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}} \to \ell^2(\mathbb{N})$ such that $U\,\tilde{Q} = Q$. Consequently $\tilde{A}\,\tilde{Q} = A\,U\,\tilde{Q}$ and hence $\tilde{A} = A\,U$ on $\overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}}$ and also $A = \tilde{A}\,U^{-1}$ on $\ell^2(\mathbb{N})$. If A is bounded, then $\tilde{A}$ is bounded on $\overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}}$. If $\tilde{A}$ is bounded on $\overline{\operatorname{span}\{\mathcal{R}(\tilde{Q})\}}$, then A is bounded on $\ell^2(\mathbb{N})$. □

The isometry Q defined by (3.3) has the advantage that it is weak-to-weak continuous. This property will be used in the next section. As a consequence its range is closed (cf. Proposition 3.6). The same is true for all isometries $\tilde{Q} := U\,Q$ constructed as described above. The weak-to-weak continuity of Q is a direct consequence of Proposition 3.7.

Proposition 3.14. *The isometry Q defined by* (3.3) *is weak-to-weak continuous.*

Proof. We only have to show that the range of Q spans the whole space $\ell^2(\mathbb{N})$ (cf. Proposition 3.7). Let $(e_i)_{i\in\mathbb{N}}$ be the orthonormal basis used in the definition of Q and let $(f_i)_{i\in\mathbb{N}}$ be the standard orthonormal basis in $\ell^2(\mathbb{N})$. The range of Q spans the whole space if (f_i) belongs to $\operatorname{span}\mathcal{R}(Q)$. But this can easily be seen because

$$f_{\kappa(i,j)} = \begin{cases} \frac{1}{\sqrt{2}}\big(Q(e_i + e_j) - Q(e_i) - Q(e_j)\big), & \text{if } j < i, \\ Q(e_i), & \text{if } j = i. \end{cases}$$

□

3.3. Inversion of quadratic isometries

With the decomposition (3.1) at hand regularization of a quadratic mapping F reduces to regularization of one possibly unbounded linear operator. At least if A is bounded this can be done by standard techniques. The interested reader finds information about regularization of unbounded linear operators in [HMvW09].

After inverting A by some regularization method we have to invert the strong quadratic isometry Q. As shown in Proposition 3.4 such mappings are continuously invertible and therefore no regularization is required. Only the fact that the solution z of the regularized linear problem typically lies in the orthogonal complement of the null space of A and possibly not in the range of Q has to be handled somehow. This can be done by projecting z onto the range of Q. A more advanced approach to tackle this problem will be presented in the next section.

In the present section we state results at first for general quadratic isometries Q and then we apply them to the concrete Q defined in (3.3).

Projection of some element z in $\ell^2(\mathbb{N})$ onto the range of an isometry $Q : X \to \ell^2(\mathbb{N})$ can be realized by solving

$$\|Q(x) - z\| \to \min_{x\in X}. \tag{3.6}$$

Existence and stability of minimizers can be shown for weak-to-weak continuous quadratic mappings, especially for the strong isometry introduced in (3.3) (cf. Proposition 3.14).

Proposition 3.15. *Let $Q : X \to \ell^2(\mathbb{N})$ be a weak-to-weak continuous quadratic mapping. Then the minimization problem* (3.6) *has at least one solution. If $(z_k)_{k\in\mathbb{N}}$ is a sequence in $\ell^2(\mathbb{N})$ with limit z and if $(x_k)_{k\in\mathbb{N}}$ is a sequence of corresponding minimizers, then (x_k) has a convergent subsequence and the limits of all convergent subsequences are solutions to* (3.6)*.*

Proof. To prove existence, take a minimizing sequence $(x_k)_{k\in\mathbb{N}}$ and observe

$$\|x_k\|^2 = \|Q(x_k)\| \leq \|Q(x_k) - z\| + \|z\| \to \inf_{x\in X} \|Q(x) - z\| + \|z\|,$$

that is, (x_k) is bounded. Thus, there is a weakly convergent subsequence and due to weak lower semi-continuity of the norm limits of convergent subsequences are minimizers of (3.6).

For proving stability take a sequence $(z_k)_{k\in\mathbb{N}}$ with limit z and a sequence of corresponding minimizers $(x_k)_{k\in\mathbb{N}}$. Then

$$\|x_k\|^2 = \|Q(x_k)\| \leq \|Q(x_k) - z_k\| + \|z_k\| \leq \|Q(0) - z\| + \|z\| = 2\,\|z\|,$$

that is, (x_k) is bounded. Thus, there is a weakly convergent subsequence and the limit $\bar{x}$ of each weakly convergent subsequence satisfies (with (x_k) denoting the subsequence)

$$\|Q(\bar{x}) - z\| \leq \liminf_{k\to\infty} \|Q(x_k) - z_k\| \leq \liminf_{k\to\infty} \|Q(x) - z_k\| = \|Q(x) - z\|$$

for all x. To obtain convergence in norm we observe

$$\begin{aligned}\|Q(\bar{x}) - z\| &\leq \liminf_{k\to\infty} \|Q(x_k) - z\| \leq \limsup_{k\to\infty} \|Q(x_k) - z\| \\ &\leq \limsup_{k\to\infty} \bigl(\|Q(x_k) - z_k\| + \|z_k - z\|\bigr) \leq \limsup_{k\to\infty} \bigl(\|Q(\bar{x}) - z_k\| + \|z_k - z\|\bigr) \\ &= \|Q(\bar{x}) - z\|,\end{aligned}$$

which yields

$$\|Q(\bar{x}) - z\| = \lim_{k\to\infty} \|Q(x_k) - z\|.$$

Equivalently we may write

$$\|\bar{x}\|^4 - 2\,\mathrm{Re}\,\langle Q(\bar{x}), z\rangle = \lim_{k\to\infty} \bigl(\|x_k\|^4 - 2\,\mathrm{Re}\,\langle Q(x_k), z\rangle\bigr).$$

Weak convergence $Q(x_k) \rightharpoonup Q(\bar{x})$ implies $\mathrm{Re}\,\langle Q(x_k), z\rangle \to \mathrm{Re}\,\langle Q(\bar{x}), z\rangle$ and thus $\|x_k\|$ converges to $\|\bar{x}\|$. Now, weak convergence in combination with convergence of the norms yields $\|x_k - \bar{x}\| \to 0$. □

Next we show how to calculate the minimizers of (3.6). We start with a lemma and then provide two theorems, one for real Hilbert spaces and one for complex Hilbert spaces. The adjoint B_Q^* of B_Q appearing in the two theorems has been defined in (1.4).

Lemma 3.16. *Let Q be a weak quadratic isometry. If* $\operatorname{Re}\langle Q(x), z\rangle \le 0$ *for all x, then zero is a solution to* (3.6). *Else the normalized minimizers of* (3.6) *coincide with the maximizers of*

$$\operatorname{Re}\langle Q(x), z\rangle \to \max_{\|x\|=1} . \tag{3.7}$$

and the minimizers of (3.6) *have norm* $\sqrt{\operatorname{Re}\langle Q(\tilde{x}), z\rangle}$, *where $\tilde{x}$ is a maximizer of* (3.7).

Proof. If $\operatorname{Re}\langle Q(x), z\rangle \le 0$ for all x, then

$$\|Q(0) - z\|^2 = \|z\|^2 \le \|Q(x)\|^2 - 2\operatorname{Re}\langle Q(x), z\rangle + \|z\|^2 = \|Q(x) - z\|^2,$$

that is, zero is a minimizer of (3.6). Else set $x := t\,u$ where $t = \|x\| \ge 0$ and $\|u\| = 1$, and minimize with respect to t for each u. The minimum of

$$h_u(t) := \|Q(t\,u) - z\|^2 = t^4 - 2\,t^2 \operatorname{Re}\langle Q(u), z\rangle + \|z\|^2$$

is at

$$t = \begin{cases} 0, & \text{if } \operatorname{Re}\langle Q(u), z\rangle \le 0, \\ \sqrt{\operatorname{Re}\langle Q(u), z\rangle}, & \text{if } \operatorname{Re}\langle Q(u), z\rangle > 0. \end{cases} \tag{3.8}$$

Thus, for each u with $\|u\| = 1$ we have

$$\min_{t\ge 0} h_u(t) = \begin{cases} \|z\|^2, & \text{if } \operatorname{Re}\langle Q(u), z\rangle \le 0, \\ \|z\|^2 - (\operatorname{Re}\langle Q(u), z\rangle)^2, & \text{if } \operatorname{Re}\langle Q(u), z\rangle > 0, \end{cases}$$

and the minimization problem (3.6) turns out to be equivalent to

$$(\operatorname{Re}\langle Q(u), z\rangle)^2 \to \max_{\substack{\|u\|=1 \\ \operatorname{Re}\langle Q(u),z\rangle > 0}} ,$$

which can be rewritten as

$$\operatorname{Re}\langle Q(u), z\rangle \to \max_{\|u\|=1} .$$

□

Theorem 3.17. *Let Q be a weak quadratic isometry between real Hilbert spaces X and $\ell^2_{\mathbb{R}}(\mathbb{N})$. If* $\langle Q(x), z\rangle \le 0$ *for all x, then zero is a solution to* (3.6). *Else each minimizer is of the form* $\sqrt{\lambda}\,\tilde{x}$ *where λ is the largest eigenvalue of the selfadjoint bounded linear operator $C : X \to X$ defined by*

$$C\,x := B^*_Q(x, z) \quad \textit{for all } x \textit{ in } X$$

and $\tilde{x}$ is a corresponding normalized eigenelement. In particular, C has positive eigenvalues.

Proof. By Lemma 3.16 the minimization problem (3.6) is equivalent to (3.7). If x is a maximizer of (3.7), which exists by Proposition 3.15, then it is a stationary point of the Lagrange function (cf. [Zei85, Theorem 43.A]), that is, there is some non-zero real number λ such that

$$C\,x - \lambda\,x = 0.$$

Here we use that C is the Fréchet derivative of $u \mapsto \frac{1}{2}\langle Q(u), z\rangle$.

It remains to show that the stationary points of the Lagrange function with largest Lagrange multiplier λ are indeed maximizers of (3.7). Taking some stationary point x with multiplier λ this follows from

$$\langle Q(x), z\rangle = \langle x, C\,x\rangle = \langle x, \lambda\,x\rangle = \lambda.$$

In addition we see that existence of a maximizer implies existence of positive eigenvalues.

□

Theorem 3.17 holds, in principle, also if X and $\ell^2(\mathbb{N})$ are considered over the complex numbers. But we are faced with two more or less technical difficulties, which force us to take some additional care of the complex case. On the one hand the mapping $x \mapsto B_Q^*(x,z)$ is antilinear, that is, $B_Q^*(a\,x,z) = \overline{a}\,B_Q^*(x,z)$ for complex numbers a. Thus we would have to use spectral theory for antilinear operators, which is of course quite similar to spectral theory of linear operators, but hardly covered in the literature. On the other hand, optimization over complex Hilbert spaces is hardly covered in the literature, too.

Our aim is to reduce the complex case to a problem in real Hilbert spaces. Before we state the theorem we have to introduce some notation. Let $(e_k)_{k\in\mathbb{N}}$ be an orthonormal basis in the complex Hilbert space X. By

$$X_{\mathbb{R}} := \left\{\sum_{k\in\mathbb{N}} a_k\,e_k : a \in \ell^2_{\mathbb{R}}(\mathbb{N})\right\}$$

we denote the real Hilbert space spanned by (e_k). For x in X we define

$$\operatorname{Re} x := \sum_{k\in\mathbb{N}} (\operatorname{Re}\langle x, e_k\rangle)\,e_k, \qquad \operatorname{Im} x := \sum_{k\in\mathbb{N}} (\operatorname{Im}\langle x, e_k\rangle)\,e_k.$$

Thus, $\operatorname{Re} x \in X_{\mathbb{R}}$ and $\operatorname{Im} x \in X_{\mathbb{R}}$. Analogously, for z in $\ell^2_{\mathbb{C}}(\mathbb{N})$ we define

$$\operatorname{Re} z := (\operatorname{Re} z_1, \operatorname{Re} z_2, \ldots), \qquad \operatorname{Im} z := (\operatorname{Im} z_1, \operatorname{Im} z_2, \ldots),$$

which are elements of $\ell^2_{\mathbb{R}}(\mathbb{N})$.

The isometry $Q : X \to \ell^2_{\mathbb{C}}(\mathbb{N})$ can be considered as a mapping on $X_{\mathbb{R}}$ and we define its real part $R : X_{\mathbb{R}} \to \ell^2_{\mathbb{R}}(\mathbb{N})$ and its imaginary part $S : X_{\mathbb{R}} \to \ell^2_{\mathbb{R}}(\mathbb{N})$ by

$$R(x) := \operatorname{Re} Q(x) \qquad \text{and} \qquad S(x) := \operatorname{Im} Q(x) \qquad \text{for } x \text{ in } X_{\mathbb{R}}. \tag{3.9}$$

The mappings R and S are quadratic mappings, too.

Theorem 3.18. *Let Q be a weak quadratic isometry between complex Hilbert spaces X and $\ell^2_{\mathbb{C}}(\mathbb{N})$ with real part R and imaginary part S as defined in* (3.9). *If* $\operatorname{Re}\langle Q(x), z\rangle \le 0$ *for all x, then zero is a solution to* (3.6). *Else each minimizer is of the form $\sqrt{\lambda}\,\tilde{x}$ where λ is the largest eigenvalue of the selfadjoint bounded linear operator $C : X_{\mathbb{R}} \times X_{\mathbb{R}} \to X_{\mathbb{R}} \times X_{\mathbb{R}}$ defined by*

$$C\begin{bmatrix} u \\ v \end{bmatrix} := \begin{bmatrix} B_R^*(u, \operatorname{Re} z) - B_S^*(v, \operatorname{Re} z) + B_S^*(u, \operatorname{Im} z) + B_R^*(v, \operatorname{Im} z) \\ -B_R^*(v, \operatorname{Re} z) - B_S^*(u, \operatorname{Re} z) - B_S^*(v, \operatorname{Im} z) + B_R^*(u, \operatorname{Im} z) \end{bmatrix}$$

for all u and v in $X_{\mathbb{R}}$ and $\tilde{x} = \tilde{u} + \mathrm{i}\,\tilde{v}$ with $[\tilde{u}, \tilde{v}]^T$ being a corresponding normalized eigenelement. In paricular, C has positive eigenvalues.

Proof. The mapping C is obviously linear and bounded. Selfadjointness follows from

$$\begin{aligned}\left\langle C\begin{bmatrix}u\\v\end{bmatrix},\begin{bmatrix}\tilde{u}\\\tilde{v}\end{bmatrix}\right\rangle &= \big\langle B_R^*(u,\operatorname{Re}z) - B_S^*(v,\operatorname{Re}z) - B_S^*(u,\operatorname{Im}z) - B_R^*(v,\operatorname{Im}z),\tilde{u}\big\rangle\\ &\quad + \big\langle -B_R^*(v,\operatorname{Re}z) - B_S^*(u,\operatorname{Re}z) + B_S^*(v,\operatorname{Im}z) - B_R^*(u,\operatorname{Im}z),\tilde{v}\big\rangle\\ &= \langle \operatorname{Re}z, B_R(u,\tilde{u})\rangle - \langle \operatorname{Re}z, B_S(v,\tilde{u})\rangle - \langle \operatorname{Im}z, B_S(u,\tilde{u})\rangle\\ &\quad - \langle \operatorname{Im}z, B_R(v,\tilde{u})\rangle - \langle \operatorname{Re}z, B_R(v,\tilde{v})\rangle - \langle \operatorname{Re}z, B_S(u,\tilde{v})\rangle\\ &\quad + \langle \operatorname{Im}z, B_S(v,\tilde{v})\rangle - \langle \operatorname{Im}z, B_R(u,\tilde{v})\rangle\\ &= \langle B_R^*(\tilde{u},\operatorname{Re}z),u\rangle - \langle B_S^*(\tilde{u},\operatorname{Re}z),v\rangle - \langle B_S^*(\tilde{u},\operatorname{Im}z),u\rangle\\ &\quad - \langle B_R^*(\tilde{u},\operatorname{Im}z),v\rangle - \langle B_R^*(\tilde{v},\operatorname{Re}z),v\rangle - \langle B_S^*(\tilde{v},\operatorname{Re}z),u\rangle\\ &\quad + \langle B_S^*(\tilde{v},\operatorname{Im}z),v\rangle - \langle B_R^*(\tilde{v},\operatorname{Im}z),u\rangle\\ &= \left\langle C\begin{bmatrix}\tilde{u}\\\tilde{v}\end{bmatrix},\begin{bmatrix}u\\v\end{bmatrix}\right\rangle = \left\langle \begin{bmatrix}u\\v\end{bmatrix}, C\begin{bmatrix}\tilde{u}\\\tilde{v}\end{bmatrix}\right\rangle.\end{aligned}$$

By the definition of R and S we have

$$\begin{aligned}Q(x) &= Q(\operatorname{Re}x + \mathrm{i}\operatorname{Im}x)\\ &= Q(\operatorname{Re}x) - Q(\operatorname{Im}x) + 2\,\mathrm{i}\,B_Q(\operatorname{Re}x,\operatorname{Im}x)\\ &= R(\operatorname{Re}x) + \mathrm{i}\,S(\operatorname{Re}x) - R(\operatorname{Im}x) - \mathrm{i}\,S(\operatorname{Im}x)\\ &\qquad + 2\,\mathrm{i}\,B_R(\operatorname{Re}x,\operatorname{Im}x) - 2\,B_S(\operatorname{Re}x,\operatorname{Im}x)\end{aligned}$$

and thus

$$\begin{aligned}\operatorname{Re}Q(x) &= R(\operatorname{Re}x) - R(\operatorname{Im}x) - 2\,B_S(\operatorname{Re}x,\operatorname{Im}x),\\ \operatorname{Im}Q(x) &= S(\operatorname{Re}x) - S(\operatorname{Im}x) + 2\,B_R(\operatorname{Re}x,\operatorname{Im}x).\end{aligned}$$

From the definition of the inner product in $\ell^2_{\mathbb{C}}(\mathbb{N})$ we see

$$\operatorname{Re}\langle Q(x),z\rangle = \langle \operatorname{Re}Q(x),\operatorname{Re}z\rangle + \langle \operatorname{Im}Q(x),\operatorname{Im}z\rangle,$$

where the first inner product is in $\ell^2_{\mathbb{C}}(\mathbb{N})$ and the second and third are in $\ell^2_{\mathbb{R}}(\mathbb{N})$. Now, remembering the definition of C, we rewrite the objective function as

$$\begin{aligned}\operatorname{Re}\langle Q(x),z\rangle &= \big\langle \operatorname{Re}z, R(\operatorname{Re}x) - R(\operatorname{Im}x) - 2\,B_S(\operatorname{Re}x,\operatorname{Im}x)\big\rangle\\ &\qquad + \big\langle \operatorname{Im}z, S(\operatorname{Re}x) - S(\operatorname{Im}x) + 2\,B_R(\operatorname{Re}x,\operatorname{Im}x)\big\rangle\\ &= \langle B_R^*(\operatorname{Re}x,\operatorname{Re}z),\operatorname{Re}x\rangle - \langle B_R^*(\operatorname{Im}x,\operatorname{Re}z),\operatorname{Im}x\rangle\\ &\quad - \langle B_S^*(\operatorname{Re}x,\operatorname{Re}z),\operatorname{Im}x\rangle - \langle B_S^*(\operatorname{Im}x,\operatorname{Re}z),\operatorname{Re}x\rangle\\ &\quad + \langle B_S^*(\operatorname{Re}x,\operatorname{Im}z),\operatorname{Re}x\rangle - \langle B_S^*(\operatorname{Im}x,\operatorname{Im}z),\operatorname{Im}x\rangle\\ &\quad + \langle B_R^*(\operatorname{Re}x,\operatorname{Im}z),\operatorname{Im}x\rangle + \langle B_R^*(\operatorname{Im}x,\operatorname{Im}z),\operatorname{Re}x\rangle\\ &= \left\langle \begin{bmatrix}\operatorname{Re}x\\\operatorname{Im}x\end{bmatrix}, C\begin{bmatrix}\operatorname{Re}x\\\operatorname{Im}x\end{bmatrix}\right\rangle.\end{aligned}$$

From here on the proof is analogous to the proof of Theorem 3.17. □

We now specify the mapping C from Theorems 3.17 and 3.18 for the strong isometry Q from Lemma 3.9.

Lemma 3.19. *Let X be a real Hilbert space and let $Q : X \to \ell^2_{\mathbb{R}}(\mathbb{N})$ be defined by* (3.3) *with an orthonormal basis $(e_i)_{i\in\mathbb{N}}$. The mapping $x \mapsto B^*_Q(x,z)$ then has the symmetric matrix representation D_z in $\mathbb{R}^{\mathbb{N}\times\mathbb{N}}$ with respect to (e_i) given by*

$$[D_z]_{i,j} := \begin{cases} \frac{1}{\sqrt{2}} z_{\kappa(i,j)}, & \text{if } j < i, \\ z_{\kappa(i,j)}, & \text{if } j = i, \end{cases}$$

where κ is the index map defined in (3.2).

Proof. From

$$\begin{aligned} \langle B^*_Q(x,z), e_i\rangle = \langle z, B_Q(x,e_i)\rangle &= \left\langle z, B_Q\left(\sum_{j=1}^{\infty} \langle x, e_j\rangle\, e_j, e_i\right)\right\rangle \\ &= \sum_{j=1}^{\infty} \langle x, e_j\rangle\, \langle z, B_Q(e_j, e_i)\rangle \end{aligned}$$

we see

$$[D_z]_{i,j} = \langle z, B_Q(e_i,e_j)\rangle = \sum_{k=1}^{\infty}\sum_{l=1}^{k} z_{\kappa(k,l)} \left[B_Q(e_i,e_j)\right]_{\kappa(k,l)}$$

and by the definition of B_Q, see (3.4), for $j \le i$ we have

$$\left[B_Q(e_i,e_j)\right]_{\kappa(k,l)} = \begin{cases} \frac{1}{\sqrt{2}}, & \text{if } j < i \text{ and } k = i \text{ and } l = j, \\ 1, & \text{if } j = i \text{ and } k = l = i, \\ 0, & \text{else,} \end{cases}$$

which together with the previous equation shows the assertion of the lemma. □

Proposition 3.20. *Let D_z be the (infinite) matrix from Lemma 3.19 and let Q be defined by* (3.3) *with an orthonormal basis $(e_i)_{i\in\mathbb{N}}$. Then the mapping C in Theorem 3.17 has the matrix representation D_z with respect to (e_i) and the mapping C in Theorem 3.18 has the matrix representation*

$$\begin{bmatrix} D_{\operatorname{Re} z} & D_{\operatorname{Im} z} \\ D_{\operatorname{Im} z} & -D_{\operatorname{Re} z} \end{bmatrix},$$

which has to be understood as a mapping from $\ell^2_{\mathbb{R}}(\mathbb{N}) \times \ell^2_{\mathbb{R}}(\mathbb{N})$ into $\ell^2_{\mathbb{R}}(\mathbb{N}) \times \ell^2_{\mathbb{R}}(\mathbb{N})$.

Proof. The matrix representation of C in Theorem 3.17 is a direct consequence of Lemma 3.19. To obtain the matrix representation for C in Theorem 3.18 we note that the imaginary part S of Q is zero and the real part R is the mapping Q restricted to the real Hilbert space $X_{\mathbb{R}}$. Consequently,

$$C = \begin{bmatrix} B^*_Q(\cdot, \operatorname{Re} z) & B^*_Q(\cdot, \operatorname{Im} z) \\ B^*_Q(\cdot, \operatorname{Im} z) & -B^*_Q(\cdot, \operatorname{Re} z) \end{bmatrix},$$

where B_Q is considered as a mapping from $X_{\mathbb{R}} \times X_{\mathbb{R}}$ into $\ell^2_{\mathbb{R}}(\mathbb{N})$. Applying Lemma 3.19 to the four B^*_Q-mappings in real spaces completes the proof. □

Remark 3.21. From the matrix representation of C (real and complex case) and the fact that z is in $\ell^2(\mathbb{N})$ we immediately see that C is a Hilbert–Schmidt operator and thus compact.

We close this section with two properties of projections onto the range of an quadratic isometry.

Proposition 3.22. *Denote by ζ an orthogonal projection of some z in $\ell^2(\mathbb{N})$ onto the range of a weak quadratic isometry $Q : X \to \ell^2(\mathbb{N})$. Then $\|\zeta\| \leq \|z\|$.*

Proof. Let x in X be such that $Q(x) = \zeta$. If $x = 0$ then $\zeta = 0$ and the assertion is obviously true. If $x \neq 0$, Lemma 3.16 in combination with the Cauchy–Schwarz inequality yields

$$\|\zeta\| = \|x\|^2 = \operatorname{Re}\left\langle Q\left(\frac{x}{\|x\|}\right), z\right\rangle \leq \|z\|. \qquad \square$$

Proposition 3.23. *If zero is an orthogonal projection of some z in $\ell^2(\mathbb{N})$ and also of $-z$ onto the range of a weak quadratic isometry $Q : X \to \ell^2(\mathbb{N})$, then $z = 0$.*

Proof. By Lemma 3.16 we have

$$\operatorname{Re}\langle Q(x), z\rangle \leq 0 \qquad \text{and} \qquad \operatorname{Re}\langle Q(x), -z\rangle \leq 0$$

for all x in X. Thus, $\operatorname{Re}\langle \tilde{z}, z\rangle = 0$ for all $\tilde{z}$ in $\overline{\operatorname{span}\mathcal{R}(Q)}$, that is, for all z in $\ell^2(\mathbb{N})$ (cf. proof of Proposition 3.14), which is equivalent to $z = 0$. □

As a consequence of the last proposition we immediately see that projecting a non-trivial subspace onto the range of a quadratic isometry always yields a non-trivial set of projections.

3.4. A regularization method

In view of Theorem 3.8 regularized inversion of a quadratic mapping can be realized in two steps: regularized inversion of a linear mapping and inversion of a quadratic isometry. The second has been discussed in the previous section. The first can be done by standard regularization methods in Hilbert spaces, for example Tikhonov regularization, Landweber iteration or spectral cut-off. Here we focus on Tikhonov regularization

$$\|A\,z - y\|^2 + \alpha\,\|z\|^2 \to \min_{z \in \ell^2(\mathbb{N})}$$

with positive regularization parameter α and data y in Y for approximate but stable solution of

$$A\,z = y^\dagger, \quad , z \in \ell^2(\mathbb{N}). \tag{3.10}$$

For Q we choose (3.3).

Throughout the present section we assume that the linear mapping A is bounded. From Example 3.11 we know that this is not always the case, but in the next section we apply the described regularization method to a mapping F for which one can prove boundedness of A.

An issue that only becomes visible at the second sight requires our attention and will have essential influence on the regularization procedure to be developed: From regularization theory in Hilbert spaces (cf. [EHN96]) we know that given noisy data y^δ with positive noise level δ, that is $\|y^\delta - y^\dagger\| \leq \delta$, corresponding regularized solutions z_α^δ converge to the norm minimizing solution $z^\dagger$ of (3.10) if δ tends to zero and the regularization parameter α is chosen in the right way depending on δ. Thus, the projections onto the range of Q converge to the projections of $z^\dagger$. Here the question arises, whether $Q(x^\dagger)$ is a projection of $z^\dagger$ onto the range of Q, where $x^\dagger$ denotes a solution of (1.1). Else we cannot expect convergence of the minimizers x_α^δ of (3.6) with $z = z_\alpha^\delta$ to $x^\dagger$.

If A is injective then obviously $z^\dagger$ belongs to the range of Q. If A is not injective we have to modify the regularization procedure in a way which forces the regularized solutions z_α^δ to converge to some solution of (3.10) which lies in the range of Q. Such a solution always exists because $y^\dagger$ belongs to the range of F by assumption. Tikhonov regularization can be modified to shift the limit of regularized solutions:

$$\|A\,z - y^\delta\|^2 + \alpha\,\|z - \zeta\|^2 \to \min_{z \in \ell^2(\mathbb{N})}.$$

Corresponding minimizers then converge to a solution of (3.10) which minimizes the distance to ζ over the set of all solutions. Choosing a suitable reference element ζ in the penalty term may force convergence to $Q(x^\dagger)$. In the method we propose below ζ is chosen iteratively.

We first present the algorithm and then discuss its construction:

1. Choose $\alpha_0 > 0$, $q \in (0,1)$ and $\tau > 1$. Set $\zeta^{(0)} = 0 \in \ell^2(\mathbb{N})$ and $k = 1$.
2. Set $\alpha_k = q\,\alpha_{k-1}$.
3. Solve $(A\,A^* + \alpha_k\,I)\,y_k = y^\delta - A\,\zeta^{(k-1)}$ for y_k.
4. Set $z^{(k)} = A^*\,y_k + \zeta^{(k-1)}$.
5. Find a minimizer x_k of (3.6) with $z = z^{(k)}$.
6. Set $\zeta^{(k)} = Q(x_k)$.
7. If $\|A\,\zeta^{(k)} - y^\delta\| \leq \tau\,\delta$ then stop. Else increase k by one and go to 2.

The three main ingredients are Tikhonov regularization as described above, the discrepancy principle for stopping the algorithm and an alternating projections approach to determine reference elements for the Tikhonov penalty.

Solving the Tikhonov minimization problem

$$\|A\,z - y^\delta\|^2 + \alpha_k\,\|z - \zeta^{(k-1)}\|^2 \to \min_{z \in \ell^2(\mathbb{N})}$$

is equivalent to solving the first order optimality condition

$$(A^*\,A + \alpha_k\,I)\,z = A^*\,y^\delta + \alpha_k\,\zeta^{(k-1)}. \tag{3.11}$$

This holds for both real and complex Hilbert spaces $\ell^2(\mathbb{N})$ and Y. In a finite-dimensional situation (e. g. after discretization) the matrix A^*A is extremely large. To see this take an orthonormal basis $(e_i)_{i\in\mathbb{N}}$ in X and consider only the n-dimensional subspace spanned by $e_1,\ldots,e_n$. Applying Q to this subspace yields a subset of $\ell^2(\mathbb{N})$ which spans a subspace of dimension $\frac{n\,(n+1)}{2}$. Consequently A would map this subspace to some finite-dimensional subspace of Y with dimension $m \leq \frac{n\,(n+1)}{2}$. The corresponding matrix A^*A then would have dimension $\frac{n\,(n+1)}{2} \times \frac{n\,(n+1)}{2}$ or, in other words, the matrix would have more than $\frac{n^4}{4}$ entries.

To avoid solving such a large system we rewrite the first order optimality condition (3.11) in a way which involves $A\,A^*$ instead of $A^*\,A$. In finite dimensions the corresponding matrix would have $m\times m$ entries and typically m is much smaller than $\frac{n\,(n+1)}{2}$. In the example presented in the next section we will have $m = 2\,n$.

We start rewriting (3.11) by substituting $\tilde{z} := z - \zeta^{(k-1)}$. Then (3.11) becomes

$$(A^*\,A + \alpha_k\,I)\,\tilde{z} = A^*\,y^\delta - A^*\,A\,\zeta^{(k-1)}$$

or, equivalently,

$$\tilde{z} = (A^*\,A + \alpha_k\,I)^{-1}\,A^*\,\big(y^\delta - A\,\zeta^{(k-1)}\big).$$

A simple calculation shows

$$(A^*\,A + \alpha_k\,I)^{-1}\,A^* = A^*\,(A\,A^* + \alpha_k\,I)^{-1},$$

which yields

$$\tilde{z} = A^*\,(A\,A^* + \alpha_k\,I)^{-1}\,\big(y^\delta - A\,\zeta^{(k-1)}\big).$$

To obtain $\tilde{z}$ we thus have to solve

$$(A\,A^* + \alpha_k\,I)\,y_k = y^\delta - A\,\zeta^{(k-1)}$$

for y_k and then calculate $\tilde{z} = A^*\,y_k$ or, equivalently, $z = A^*\,y_k + \zeta^{(k-1)}$. This is done in steps 3 and 4 of the algorithm.

Step 5 is based on Theorems 3.17 and 3.18 in connection with Proposition 3.20. Here we have to find the largest eigenvalue and a corresponding eigenelement of a symmetric Hilbert–Schmidt operator (or a symmetric matrix in finite dimensions). There are several standard algorithms for this purpose. We mention the power iteration (also known as Von Mises iteration).

The iteration is stopped in step 7 if the so called discrepancy principle is satisfied, that is, if the obtained approximate solution x_k fulfils $\|F(x_k) - y^\delta\| \leq \tau\delta$ where τ should be slightly greater than one. Note that $F(x_k) = A\,\zeta^{(k)}$. The idea behind this well-known principle is that there is no reason to get closer to the noisy data y^δ than the exact data $y^\dagger$ is.

To complete the description of the algorithm the interplay of the Tikhonov minimization problem and the projection in steps 5 and 6 has to be discussed. At first we consider the idea without regularization. So the question is: How to iteratively approximate $Q(x^\dagger)$? Having approximations of $Q(x^\dagger)$ we also have approximations of $x^\dagger$. Thus, everything can be considered in $\ell^2(\mathbb{N})$. Obviously, $Q(x^\dagger)$ belongs to the intersection of the range of Q and the shifted null space $z^\dagger + \mathcal{N}(A)$. The alternating projections method can be used to find points in the intersection of two sets. One starts at some point (we use $z^\dagger$) by projecting it orthogonally onto one of the sets. Then this projected point is projected orthogonally onto the other set. With this second projection point the procedure is repeated. One can show that this alternating projections method converges weakly to a point in the intersection of the two sets if both sets are closed and convex (see [KR12, Theorem 1.3(a)]). The set $z^\dagger + \mathcal{N}(A)$ is closed and convex, but the range of Q is only closed and not convex. Thus, it is not clear whether the method converges and we were not able to solve this issue.

Now we incorporate regularization into the idea of alternating projections. Projecting some $\zeta^{(k-1)}$ onto $z^\dagger + \mathcal{N}(A)$ means that we search for the solution of (3.10) with minimal distance to the original point $\zeta^{(k-1)}$. Such a solution can be found approximately but stable by minimizing a Tikhonov functional with reference element $\zeta^{(k-1)}$ in the penalty term. For fixed k the regularization parameter α in the Tikhonov functional can be chosen by starting with a large value α_0 and then decreasing α by some factor q until some stopping rule (e. g. discrepancy principle) is fulfilled. In principle this parameter choice has to be realized for each k. To save computation time parameter choice and (outer) iteration with respect to k can be combined because in the first outer iterations we do not need maximal accuracy. Thus, solving the Tikhonov minimization problem only for one regularization parameter and decreasing the regularization parameter for the next iteration with new reference element seems to be a valid approach. A similar idea is used for instance in the TIGRA method (see [Ram03]).

Due to the lack of convexity of the range of Q we do not have a full convergence proof for our proposed algorithm. We only have proven that it is stable, which is a consequence of standard regularization theory in Hilbert spaces and of Proposition 3.15, and we gave a precise motivation why we expect that the algorithm yields useful results.

3.5. Numerical example

To demonstrate practicality of the algorithm proposed in the previous section we implemented it to solve the complex-valued autoconvolution problem with full data described in Subsection 1.2.1. The corresponding quadratic mapping was introduced in (1.5). This problem is very similar to the autoconvolution problem for the SD-SPIDER method described in Subsection 1.2.2. The only difference is that in (1.9) we have an additional kernel function k. Assuming that this kernel has a multiplicative structure

$$k(s,t) = \tilde{k}(t)\,\tilde{k}(s-t), \quad (s,t) \in \mathcal{D}(k)$$

with a function $\tilde{k} : (0,1) \to \mathbb{C}$ inverting the SD-SPIDER-related mapping is equivalent to inverting (1.5) and dividing by $\tilde{k}$. Of course our method can be applied to the kernel-based autoconvolution problem even if the kernel has not such a multiplicative structure. The only additional difficulty would be that we no longer would be able to explicitly calculate several expressions when discretizing our spaces and mappings. Instead we would have to use numerical integration.

We discretize functions in $X = L^2_{\mathbb{C}}(0,1)$ by cutting off their Fourier coefficients. For k in $\mathbb{N}$ set

$$\gamma_k := (-1)^k \left\lfloor \frac{k}{2} \right\rfloor$$

and

$$e_k(t) := \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,t}, \quad t \in (0,1).$$

Then $(e_k)_{k\in\mathbb{N}}$ is the usual Fourier basis, but re-indexed to avoid negative indices. The indices $k = 1, 2, \ldots$ are mapped to $\gamma_k = 0, 1, -1, 2, -2, \ldots$. For our computations we only use the first n Fourier coefficients, that is, approximate solutions belong to the span of $e_1, \ldots, e_n$.

The mapping Q from (3.3) satisfies $[Q(x)]_{\kappa(k,l)} = 0$ for $k > n$, that is, only the first $\frac{n\,(n+1)}{2}$ components of $Q(x)$ are non-zero.

To obtain the linear mapping A we need $B_F(e_k, e_l)$ for $k = 1, 2, \ldots$ and $l = 1, \ldots, k$. Evaluating the integral in (1.6) we obtain

$$\big(F(e_k)\big)(s) = \begin{cases} s\,\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s}, & \text{if } s \in (0,1), \\ (2-s)\,\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s}, & \text{if } s \in (1,2), \end{cases} \tag{3.12}$$

and

$$\big(B_F(e_k, e_l)\big)(s) = \begin{cases} \frac{1}{2\,\pi\,\mathrm{i}\,(\gamma_k-\gamma_l)}\,\big(\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s} - \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_l\,s}\big), & \text{if } s \in (0,1), \\ \frac{-1}{2\,\pi\,\mathrm{i}\,(\gamma_k-\gamma_l)}\,\big(\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s} - \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_l\,s}\big), & \text{if } s \in (1,2), \end{cases} \tag{3.13}$$

for $k = 1, 2, \ldots$ and $l = 1, \ldots, k-1$. Here we see that the images of $F(x) = A\,Q(x)$ for x in span $\{e_1, \ldots, e_n\}$ are continuous. Thus, to discretize them we may use piecewise linear interpolation on an equispaced grid with $m+1$ nodes in $(0,2)$ at $\frac{2\,k}{m}$ for $k = 0, 1, \ldots, m$.

Proposition 3.24. *The mapping A is bounded.*

Proof. For z in $\mathcal{D}(A)$ we have

$$(A\,z)(s) = \sum_{k=1}^{\infty}\left(\sum_{l=1}^{k-1}\frac{\sqrt{2}\,z_{\kappa(k,l)}}{2\,\pi\,\mathrm{i}\,(\gamma_k-\gamma_l)}\,\big(\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s} - \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_l\,s}\big) + z_{\kappa(k,k)}\,s\,\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s}\right).$$

We show that this series converges in $L^2_{\mathbb{C}}(0,2)$ for each z in $\ell^2_{\mathbb{C}}(\mathbb{N})$, that is, $\mathcal{D}(A) = \ell^2_{\mathbb{C}}(\mathbb{N})$. Then A is obviously bounded. We restrict our attention to convergence in $L^2_{\mathbb{C}}(0,1)$. Analogous steps lead to convergence in $L^2_{\mathbb{C}}(1,2)$ and thus to convergence in $L^2_{\mathbb{C}}(0,2)$.

At first we show that

$$a_n(s) := \sum_{k=1}^{n}\sum_{l=1}^{k-1}\frac{z_{\kappa(k,l)}}{\gamma_k-\gamma_l}\,\big(\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s} - \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_l\,s}\big), \qquad s \in (0,1),$$

defines a Cauchy sequence $(a_n)_{n\in\mathbb{N}}$ and then that

$$b_n(s) := \sum_{k=1}^{n} z_{\kappa(k,k)}\,s\,\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s}, \qquad s \in (0,1),$$

defines a Cauchy sequence $(b_n)_{n\in\mathbb{N}}$.

We have

$$-\sum_{k=1}^{n}\sum_{l=1}^{k-1}\frac{z_{\kappa(k,l)}}{\gamma_k-\gamma_l}\,\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_l\,s} = -\sum_{l=1}^{n}\sum_{k=l+1}^{n}\frac{z_{\kappa(k,l)}}{\gamma_k-\gamma_l}\,\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_l\,s} = \sum_{k=1}^{n}\sum_{l=k+1}^{n}\frac{z_{\kappa(l,k)}}{\gamma_k-\gamma_l}\,\mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\,s}$$

and hence

$$a_n(s) = \sum_{k=1}^{n} \left(\sum_{l=1}^{k-1} \frac{z_{\kappa(k,l)}}{\gamma_k - \gamma_l} + \sum_{l=k+1}^{n} \frac{z_{\kappa(l,k)}}{\gamma_k - \gamma_l} \right) \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\, s}.$$

For m in $\mathbb{N}$ (without loss of generality we assume $m < n$) we obtain

$$\begin{aligned} \|a_n - a_m\|^2 &= \Bigg\| \sum_{k=1}^{m} \left(\sum_{l=m+1}^{n} \frac{z_{\kappa(l,k)}}{\gamma_k - \gamma_l} \right) \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\, \cdot} \\ &\qquad + \sum_{k=m+1}^{n} \left(\sum_{l=1}^{k-1} \frac{z_{\kappa(k,l)}}{\gamma_k - \gamma_l} + \sum_{l=k+1}^{n} \frac{z_{\kappa(l,k)}}{\gamma_k - \gamma_l} \right) \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\, \cdot} \Bigg\|^2_{L^2_{\mathbb{C}}(0,1)} \\ &= \sum_{k=1}^{m} \left| \sum_{l=m+1}^{n} \frac{z_{\kappa(l,k)}}{\gamma_k - \gamma_l} \right|^2 + \sum_{k=m+1}^{n} \left| \sum_{l=1}^{k-1} \frac{z_{\kappa(k,l)}}{\gamma_k - \gamma_l} + \sum_{l=k+1}^{n} \frac{z_{\kappa(l,k)}}{\gamma_k - \gamma_l} \right|^2 \end{aligned}$$

and the Cauchy–Schwarz inequality yields

$$\left| \sum_{l=m+1}^{n} \frac{z_{\kappa(l,k)}}{\gamma_k - \gamma_l} \right|^2 \leq \left(\sum_{l=m+1}^{n} \frac{1}{|\gamma_k - \gamma_l|^2} \right) \left(\sum_{l=m+1}^{n} |z_{\kappa(l,k)}|^2 \right)$$

as well as

$$\begin{aligned} &\left| \sum_{l=1}^{k-1} \frac{z_{\kappa(k,l)}}{\gamma_k - \gamma_l} + \sum_{l=k+1}^{n} \frac{z_{\kappa(l,k)}}{\gamma_k - \gamma_l} \right|^2 \\ &\qquad \leq \left(\sum_{l=1}^{k-1} \frac{1}{|\gamma_k - \gamma_l|^2} + \sum_{l=k+1}^{n} \frac{1}{|\gamma_k - \gamma_l|^2} \right) \left(\sum_{l=1}^{k-1} |z_{\kappa(k,l)}|^2 + \sum_{l=k+1}^{n} |z_{\kappa(l,k)}|^2 \right). \end{aligned}$$

Together with

$$\sum_{\substack{l=1 \\ l \neq k}}^{\infty} \frac{1}{|\gamma_k - \gamma_l|^2} \leq 2 \sum_{l=1}^{\infty} \frac{1}{l^2} = \frac{\pi^2}{3}$$

and an extensive re-indexing we see

$$\begin{aligned} \|a_n - a_m\|^2 &\leq \frac{\pi^2}{3} \left(\sum_{k=1}^{m} \sum_{l=m+1}^{n} |z_{\kappa(l,k)}|^2 + \sum_{k=m+1}^{n} \left(\sum_{l=1}^{k-1} |z_{\kappa(k,l)}|^2 + \sum_{l=k+1}^{n} |z_{\kappa(l,k)}|^2 \right) \right) \\ &= \frac{2\,\pi^2}{3} \left(\sum_{k=1}^{n} \sum_{l=1}^{k-1} |z_{\kappa(k,l)}|^2 - \sum_{k=1}^{m} \sum_{l=1}^{k-1} |z_{\kappa(k,l)}|^2 \right). \end{aligned}$$

Since

$$\sum_{k=1}^{\infty} \sum_{l=1}^{k-1} |z_{\kappa(k,l)}|^2 \leq \|z\|^2 < \infty$$

we obtain $\|a_n - a_m\| \to 0$ if $m, n \to \infty$.

Now we come to $(b_n)_{n\in\mathbb{N}}$. Here for m in $\mathbb{N}$ (again $m < n$) we have

$$\begin{aligned}\|b_n - b_m\|^2 &= \int\limits_0^1 \left|\sum_{m+1}^{n} z_{\kappa(k,k)}\, s\, \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\, s}\right|^2 \mathrm{d}s = \int\limits_0^1 |s|^2 \left|\sum_{m+1}^{n} z_{\kappa(k,k)}\, \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\, s}\right|^2 \mathrm{d}s\\ &\leq \int\limits_0^1 \left|\sum_{m+1}^{n} z_{\kappa(k,k)}\, \mathrm{e}^{2\,\pi\,\mathrm{i}\,\gamma_k\, s}\right|^2 \mathrm{d}s = \sum_{k=m+1}^{n} |z_{\kappa(k,k)}|^2,\end{aligned}$$

which together with

$$\sum_{k=1}^{\infty} |z_{\kappa(k,k)}|^2 \leq \|z\|^2 < \infty$$

implies $\|b_n - b_m\| \to 0$ if $m, n \to \infty$.

To complete the proof note that

$$A\, z = \lim_{n\to\infty} \left(\frac{\sqrt{2}}{2\,\pi\,\mathrm{i}}\, a_n + b_n\right)$$

and this limit exists for all z in $\ell^2_{\mathbb{C}}(\mathbb{N})$. □

The adjoint $A^* : L^2_{\mathbb{C}}(0,2) \to \ell^2_{\mathbb{C}}(\mathbb{N})$ of A is given by

$$[A^*\, y]_{\kappa(k,l)} = \begin{cases} \langle y, F(e_k)\rangle, & l = k,\\ \sqrt{2}\,\langle y, B_F(e_k, e_l)\rangle, & l < k \end{cases} \tag{3.14}$$

for $k = 1, 2, \ldots$ and $l = 1, \ldots, k$.

To obtain a discretized version of $A\,A^*$ mapping piecewise linear functions to piecewise linear functions we calculate the inner products in (3.14) for $k = 1, 2, \ldots, n_{A\,A^*}$ and $l \leq k$ analytically and then apply A and piecewise linear interpolation. These calculations are elementary and were carried out with the help of a computer algebra system. Results were copied directly to the source code of the implemented algorithm and we do not provide the extensive formulas here.

From (3.12) and (3.13) we see that F applied to $\operatorname{span}\{e_1, \ldots, e_n\}$ spans a subspace of dimension $2\,n - 1$ and that $(F(x))(0) = 0 = (F(x))(2)$ for all x in $\operatorname{span}\{e_1, \ldots, e_n\}$. Thus, we choose $2\,n - 1 + 2 = 2\,n + 1$ nodes for linear interpolation in Y, that is $m = 2\,n$. In our numerical experiments we use $n = 200$, $n_{A\,A^*} = 200$ and $m = 400$. Further, choice of the regularization parameter is controlled by $\alpha_0 = 1$, $q = 0.9$ and $\tau = 1.6$ in all experiments.

Given a function $x^\dagger$ in $L^2_{\mathbb{C}}(0,1)$ we calculate a piecewise linear approximation of $y^\dagger = F(x^\dagger)$ by numerical integration (to avoid an 'inverse crime' we do not use the discretized A and Q). Then we add random noise following a Gaussian distribution to the function values of $y^\dagger$ at the interpolation nodes. This yields noisy data y^δ. The noise is scaled such that $\|y^\delta - y^\dagger\| = \delta$. In the experiments below we provide the relative noise level

$$\delta_{\mathrm{rel}} := \frac{\delta}{\|y^\dagger\|}.$$

Remember that $F(x) = F(-x)$ for all x. The algorithm outputs only one of the two solutions. To improve presentation we manually flip the sign so that the signs of calculated and exact solution coincide.

We consider three examples. The table below shows the exact solution $x^\dagger$, the relative noise level δ_{rel}, the number of iterations the algorithm performed until the discrepancy principle was satisfied and where to find corresponding plots.

	exact solution	noise level	iterations	figures
example 1	$x^\dagger = t\,\mathrm{e}^{8\pi \mathrm{i}\, t}$	$\delta_{\mathrm{rel}} = 0.05$	21	3.2–3.9
example 2	$x^\dagger = t\,\mathrm{e}^{8\pi \mathrm{i}\, t}$	$\delta_{\mathrm{rel}} = 0.2$	13	3.10–3.17
example 3	$x^\dagger = t^2 + \mathrm{i}\,(\lvert t - \frac{1}{2}\rvert - \frac{1}{2})$	$\delta_{\mathrm{rel}} = 0.1$	22	3.18–3.25

We do not give extensive numerical results here. We only demonstrate with few examples that the described method works well and discuss a number of features the reconstructions show. The interested reader finds further numerical experiments and a comparison to other methods in [BF15].

Figure 3.2 shows the exact solution $x^\dagger$ for example 1. Figures 3.3 and 3.4 show corresponding exact and noisy data. The output of the algorithm is depicted in Figure 3.5. For better comparability real and imaginary parts are plotted in Figures 3.6 and 3.7 for exact and noisy data and in Figures 3.8 and 3.9 for exact and reconstructed solution. The same system is used for the figures belonging to examples 2 and 3.

The real part of the reconstructed solution in example 1 (Figure 3.8) shows some oscillations near zero and one. The Fourier basis, which we used for discretization in X, only contains 1-periodic functions. But if we look at the function we used in example 1 as a periodic function, we see a jump at zero (or one). To obtain a good reconstruction of this jump we would have to reconstruct much more Fourier coefficients because the elements of the Fourier basis are very smooth. Same effect can be seen in Figures 3.16 and 3.24.

In example 2 we use a four times higher noise level than in example 1, but the reconstructions still are quite good. Here the reason lies in the fact that inverting a quadratic mapping is of the same nature as taking the square root of a real number. Thus, if the noise level in the data is multiplied by four, the solution error is multiplied only by two. To our regret we did not find a rigorous proof for this intuitive explanation.

The exact solution in example 3 is not differentiable. Approximation of the kink by a smooth basis requires more coefficients than we reconstruct. The result produced by our algorithm is to some extent underregularized. This can be regarded as a compromise between approximation of the kink and regularization of the problem.

Page intentionally left blank.

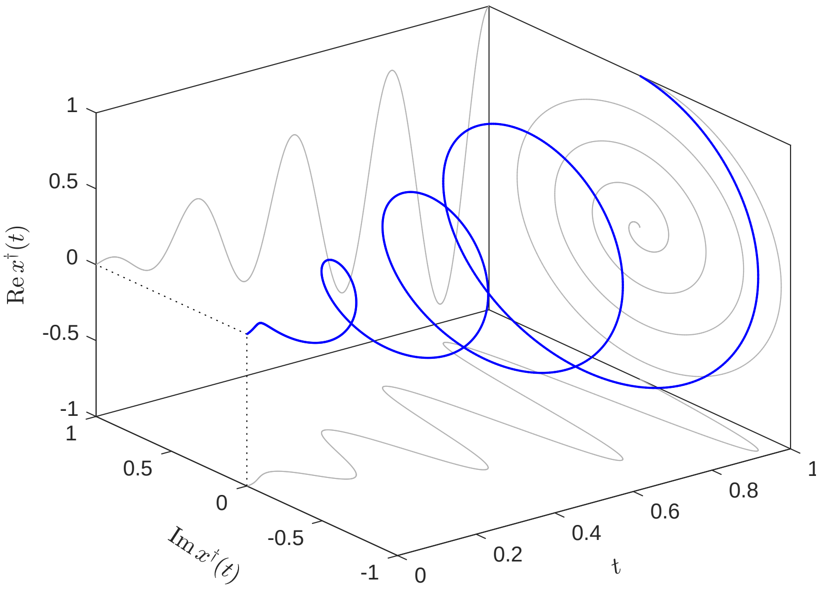

Figure 3.2.: Exact solution for example 1.

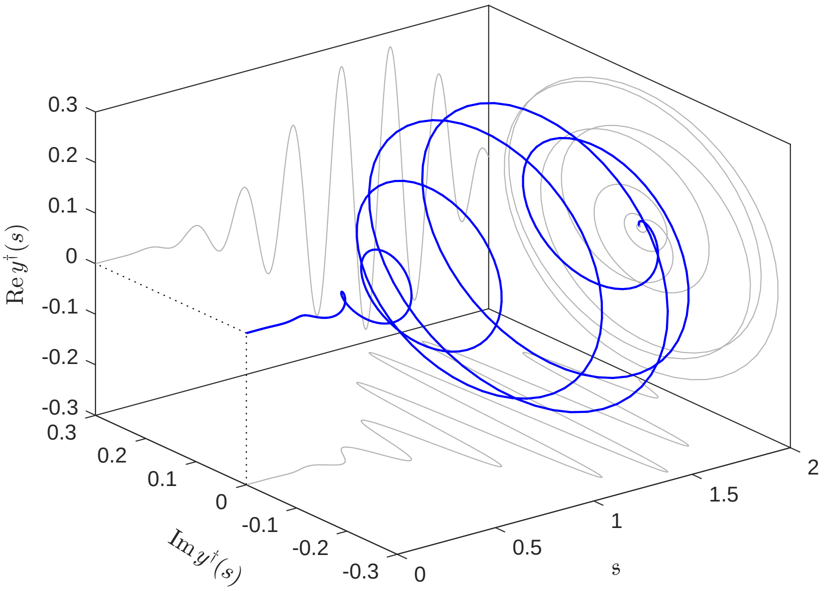

Figure 3.3.: Exact data for example 1.

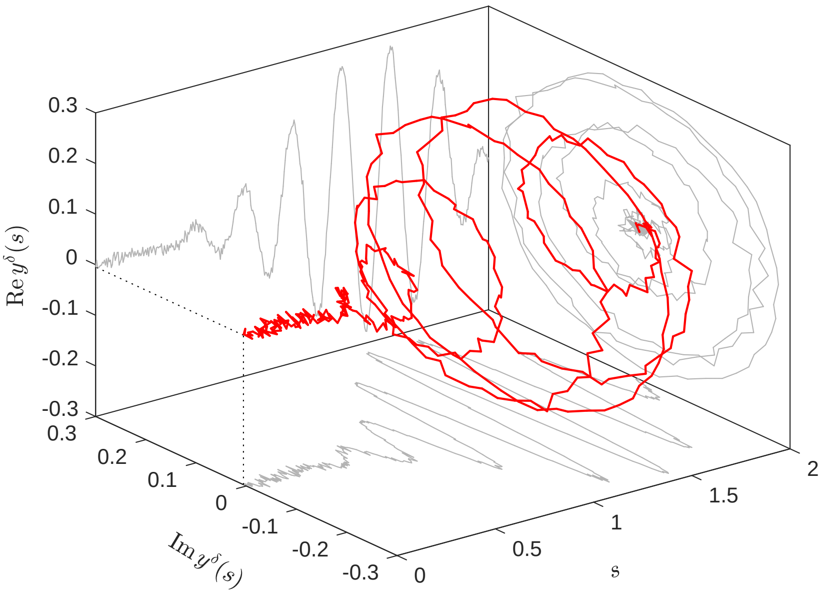

Figure 3.4.: Noisy data for example 1.

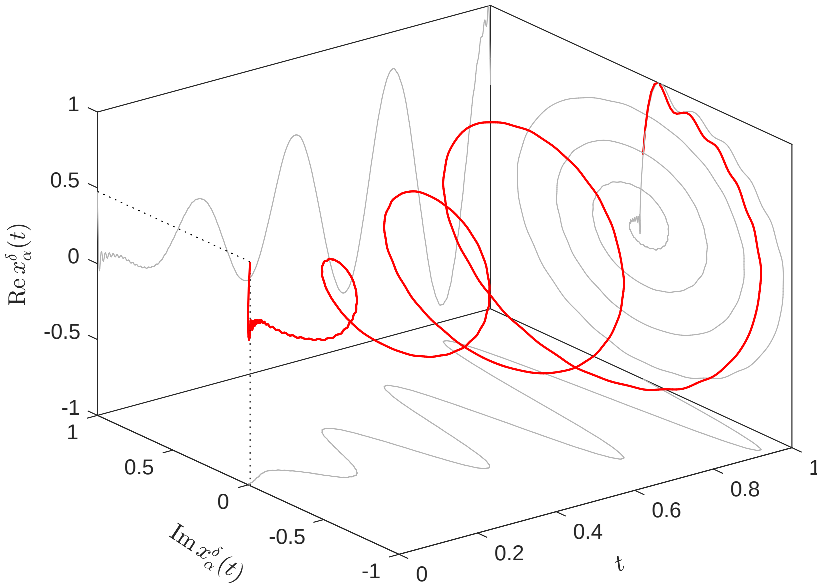

Figure 3.5.: Reconstructed solution for example 1.

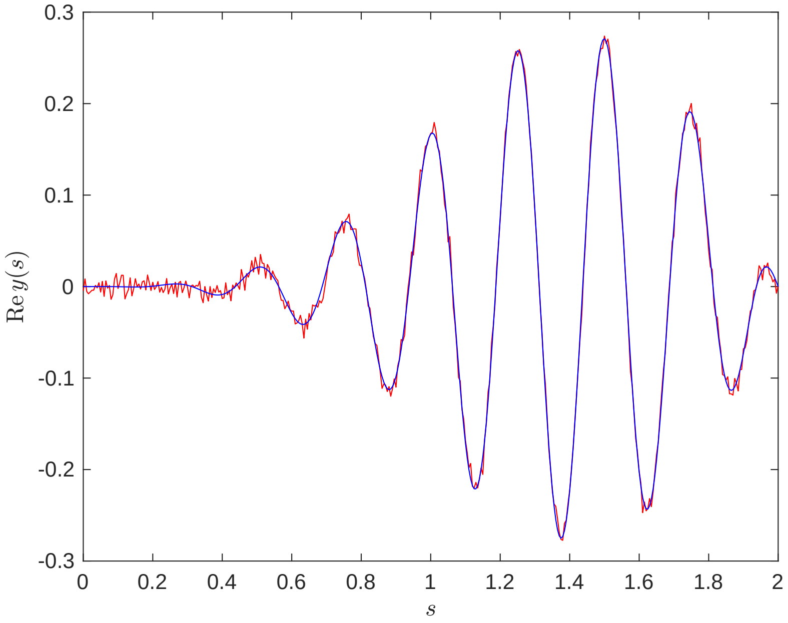

Figure 3.6.: Real parts of exact (blue) and noisy (red) data for example 1.

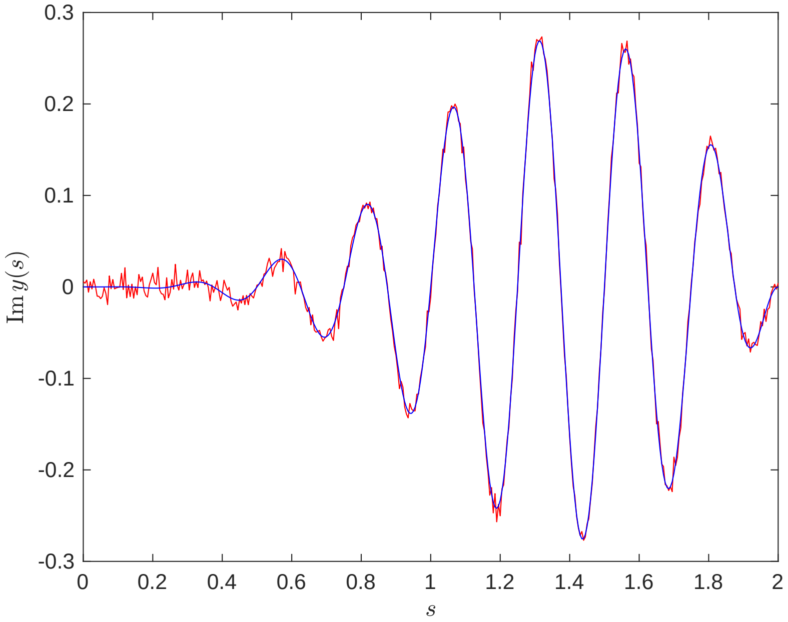

Figure 3.7.: Imaginary parts of exact (blue) and noisy (red) data for example 1.

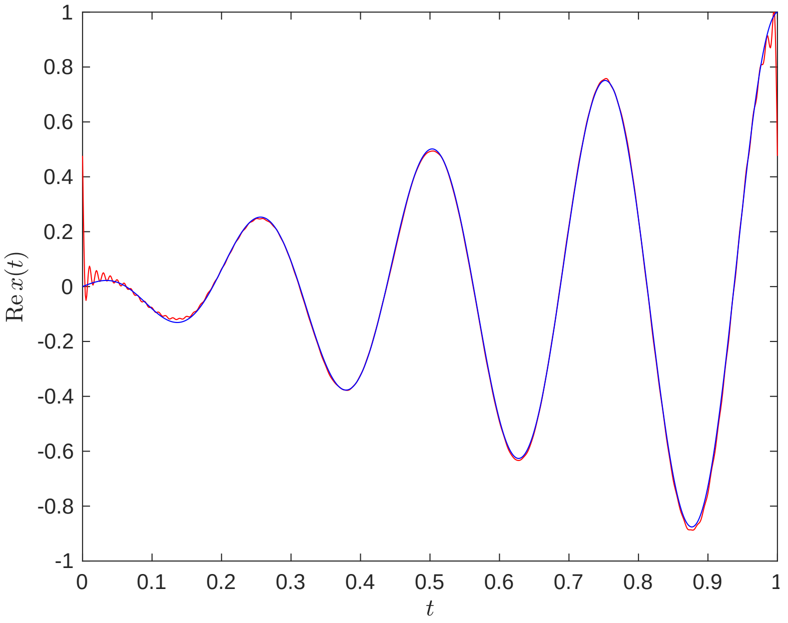

Figure 3.8.: Real parts of exact (blue) and reconstructed (red) solution for example 1.

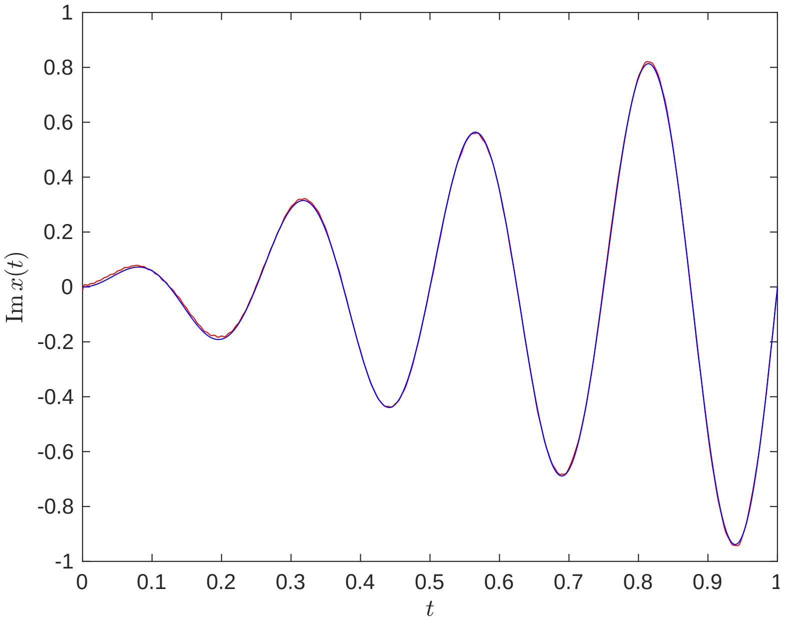

Figure 3.9.: Imaginary parts of exact (blue) and reconstructed (red) solution for example 1.

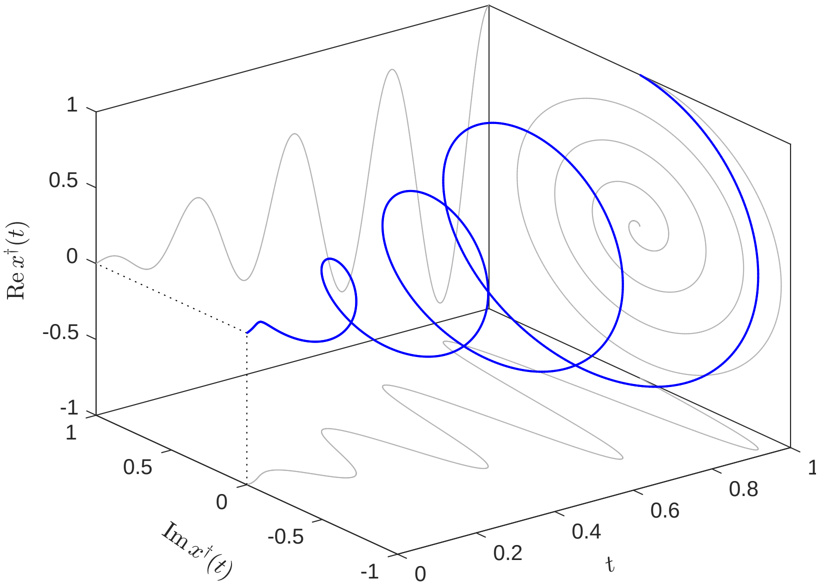

Figure 3.10.: Exact solution for example 2.

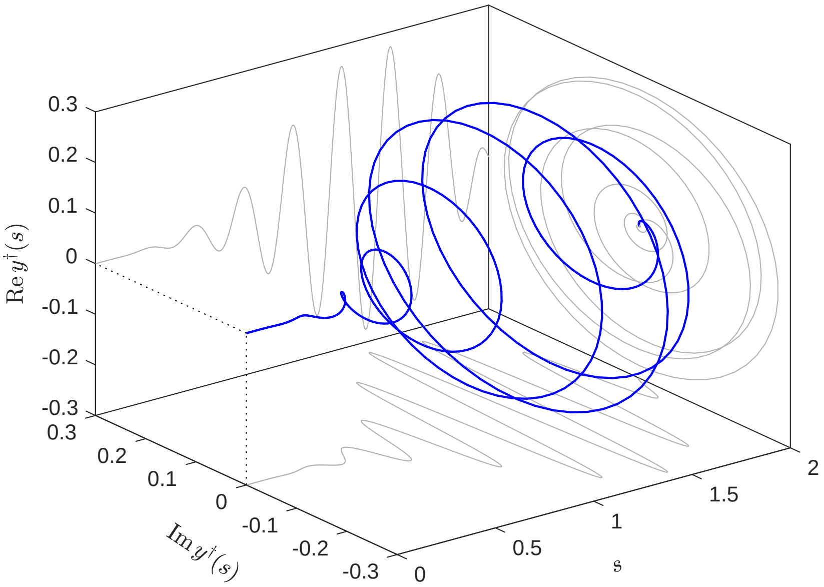

Figure 3.11.: Exact data for example 2.

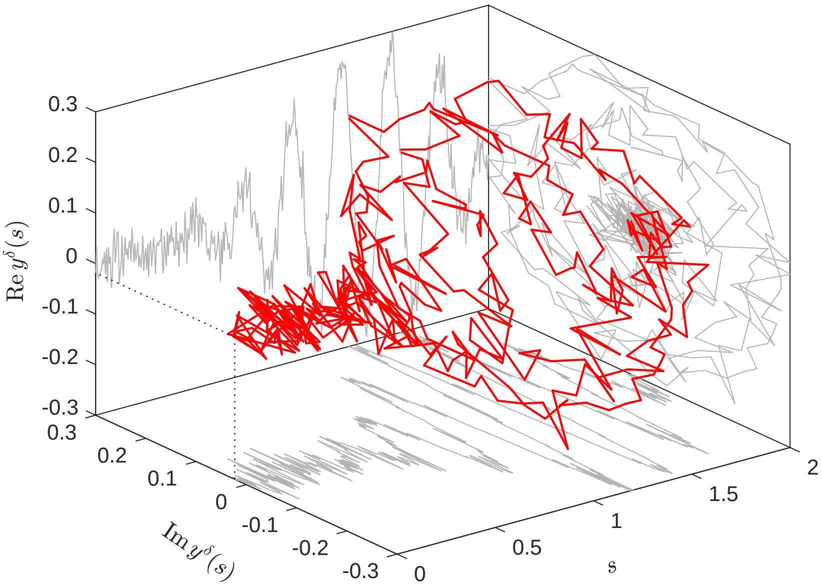

Figure 3.12.: Noisy data for example 2.

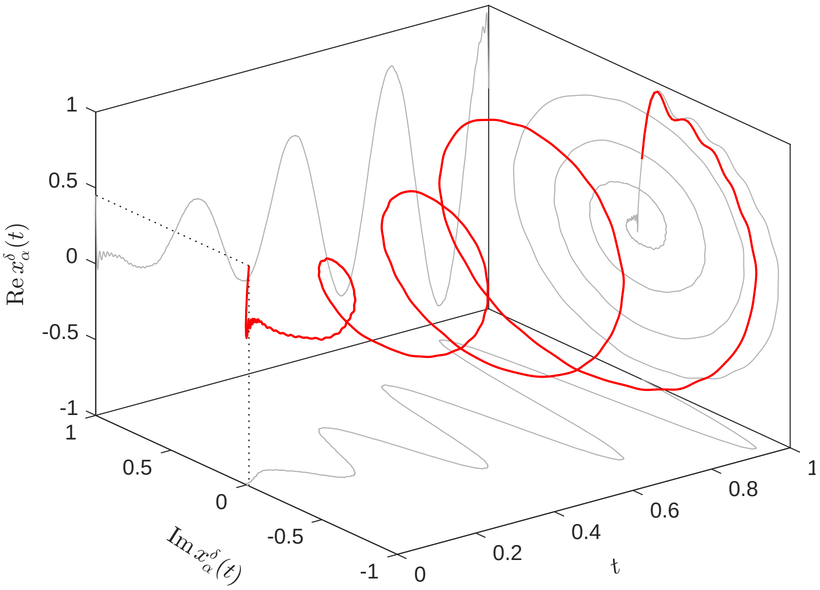

Figure 3.13.: Reconstructed solution for example 2.

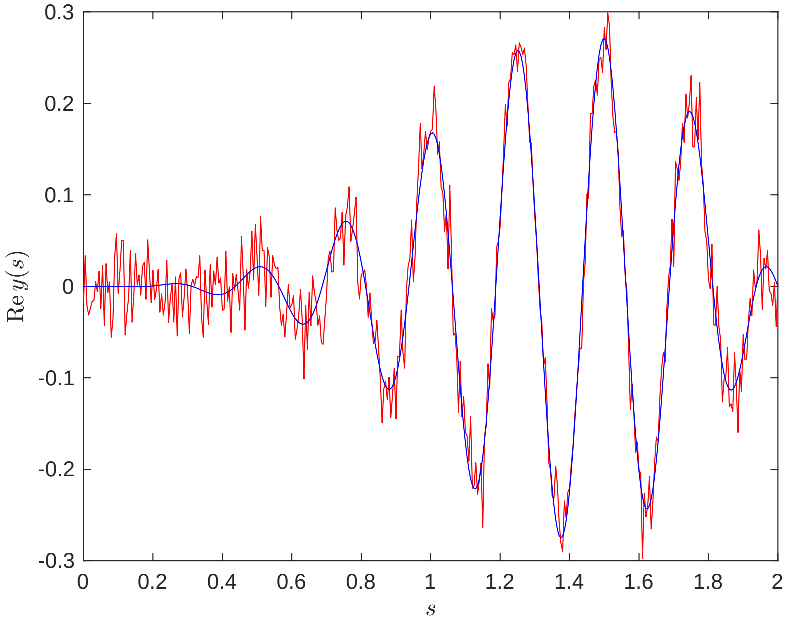

Figure 3.14.: Real parts of exact (blue) and noisy (red) data for example 2.

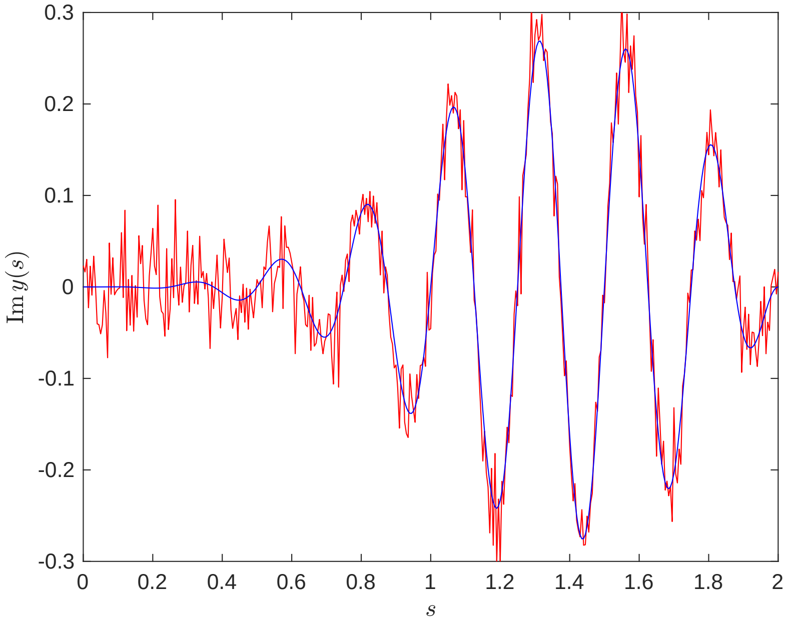

Figure 3.15.: Imaginary parts of exact (blue) and noisy (red) data for example 2.

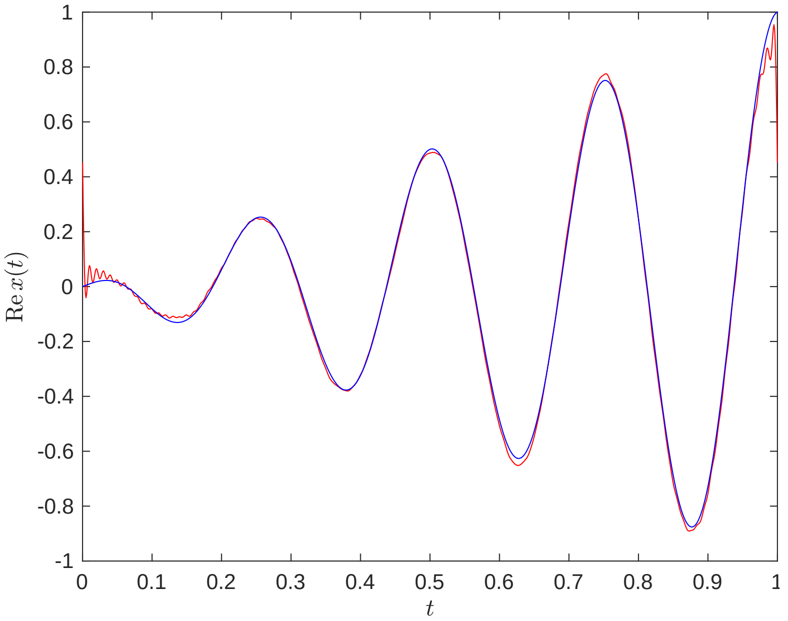

Figure 3.16.: Real parts of exact (blue) and reconstructed (red) solution for example 2.

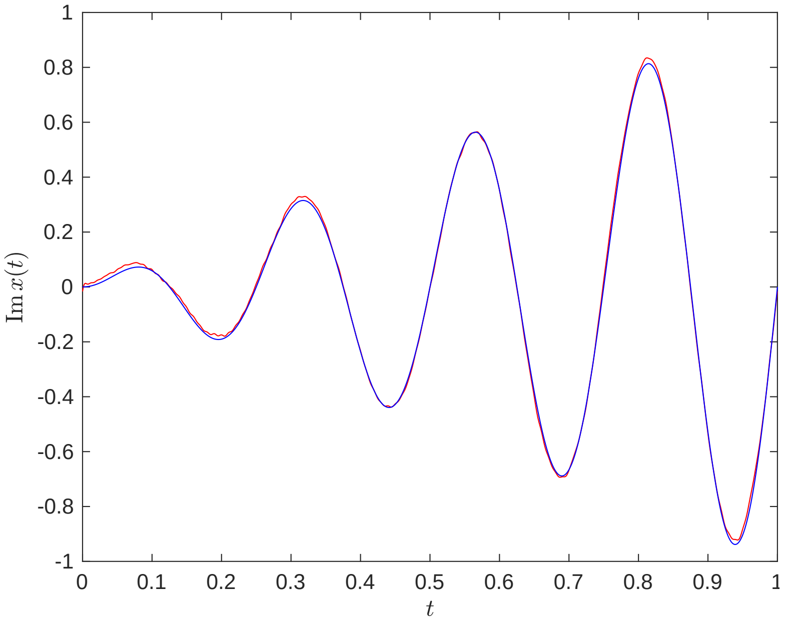

Figure 3.17.: Imaginary parts of exact (blue) and reconstructed (red) solution for example 2.

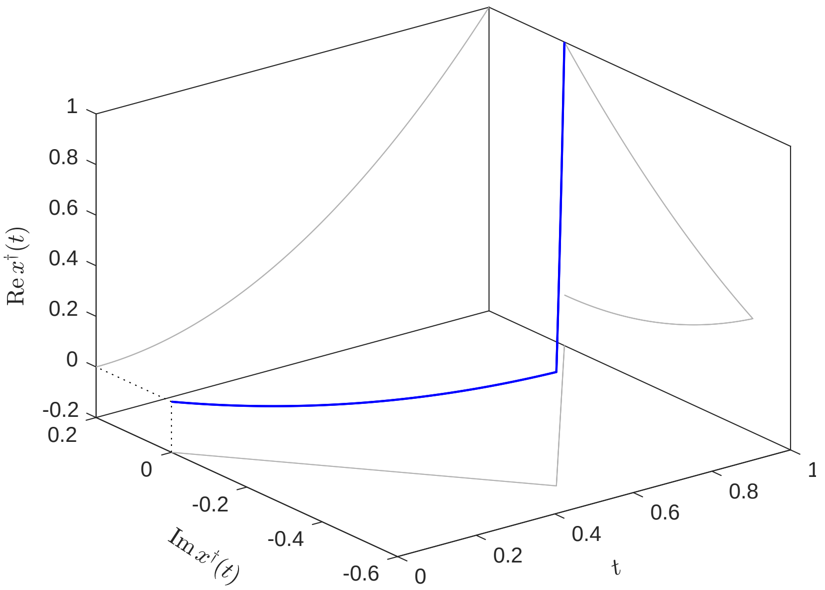

Figure 3.18.: Exact solution for example 3.

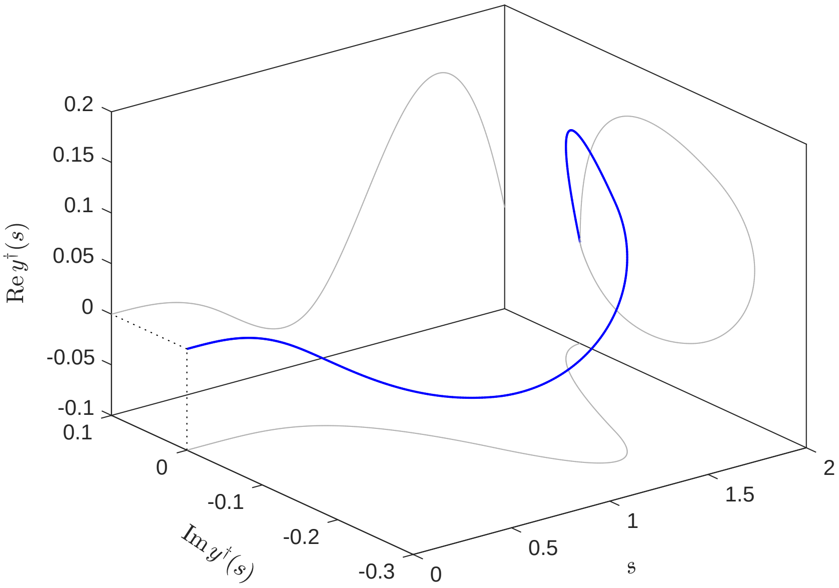

Figure 3.19.: Exact data for example 3.

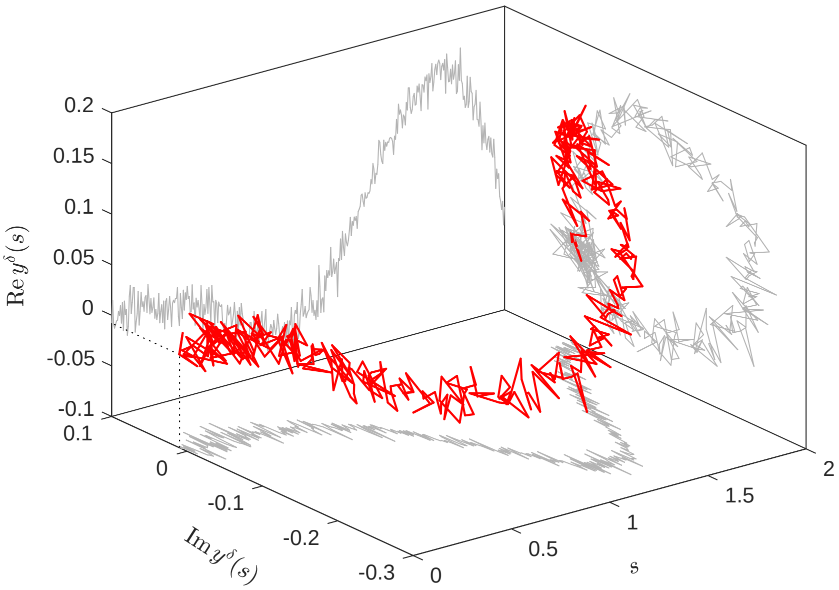

Figure 3.20.: Noisy data for example 3.

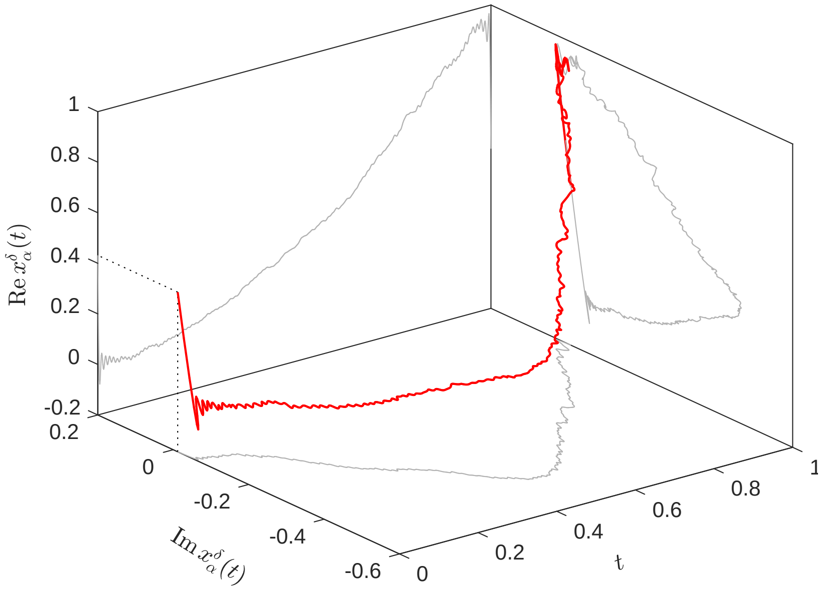

Figure 3.21.: Reconstructed solution for example 3.

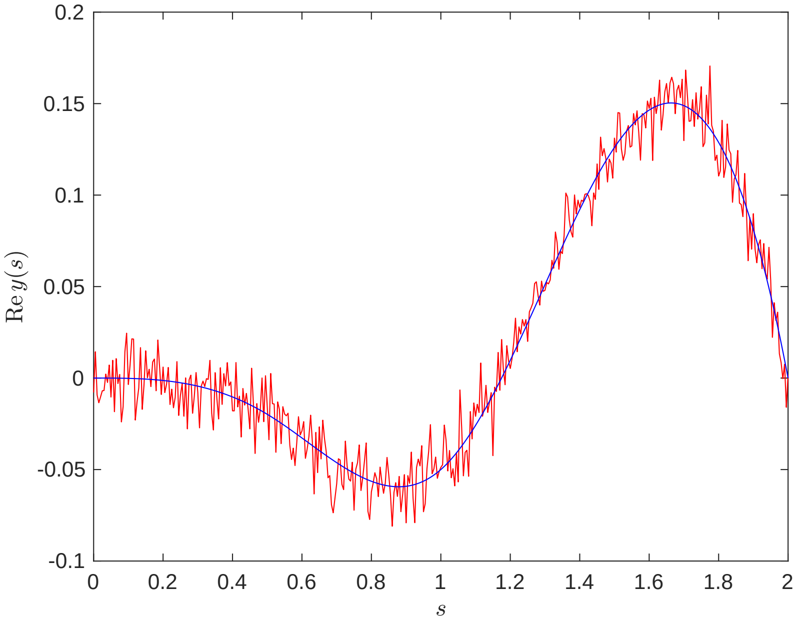

Figure 3.22.: Real parts of exact (blue) and noisy (red) data for example 3.

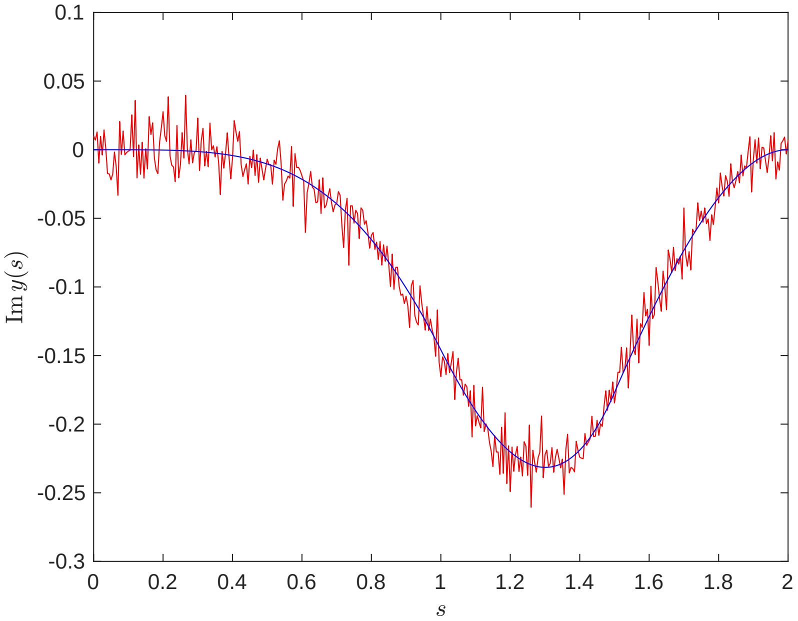

Figure 3.23.: Imaginary parts of exact (blue) and noisy (red) data for example 3.

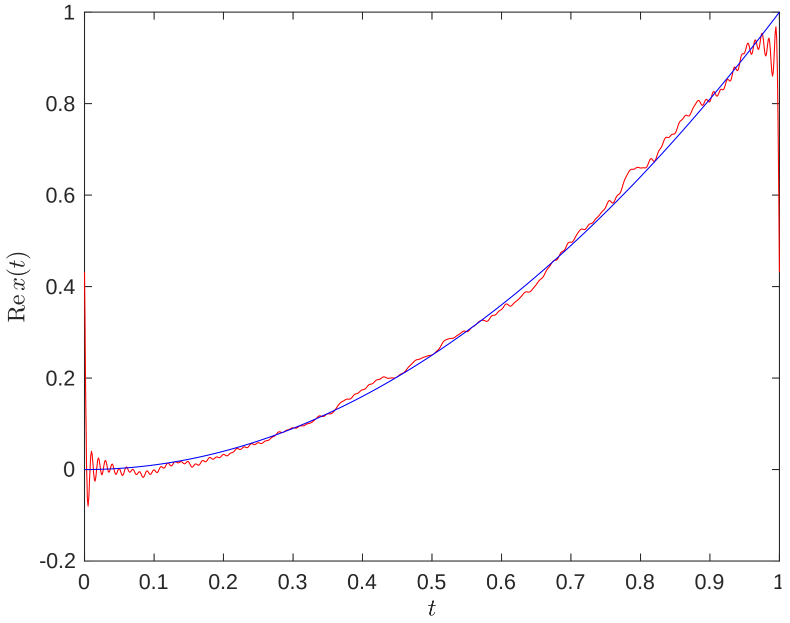

Figure 3.24.: Real parts of exact (blue) and reconstructed (red) solution for example 3.

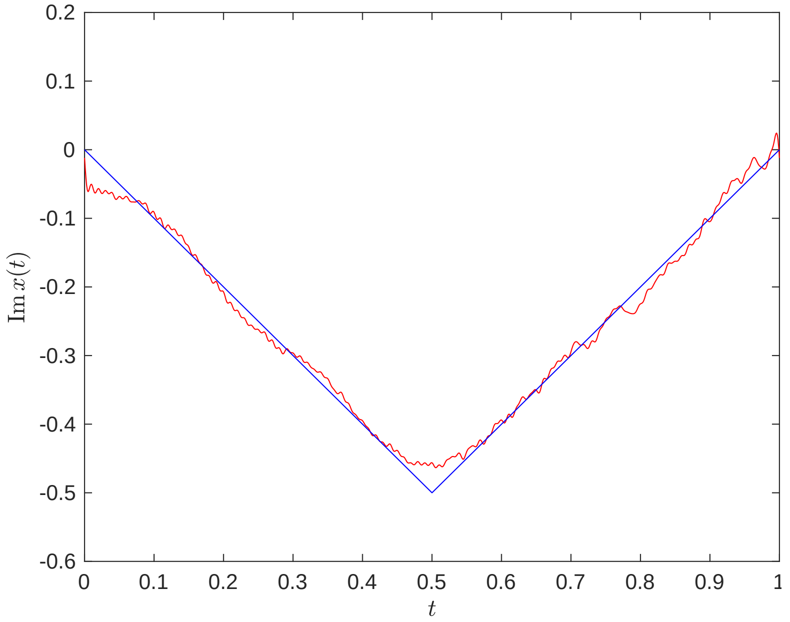

Figure 3.25.: Imaginary parts of exact (blue) and reconstructed (red) solution for example 3.

4. Variational source conditions

In this chapter we discuss a mathematical tool called variational source conditions in the context of quadratic inverse problems. We start with a short introduction to the topic and then discuss more or less classical alternatives and their deficiencies. The first of the two main results of this chapter will state that variational source conditions are the right tool for our purposes, and the second demonstrates a way to obtain concrete variational source conditions for quadratic mappings.

4.1. About variational source conditions

Even for linear ill-posed inverse problems convergence of regularized solutions to an exact solution of (1.1) can be arbitrarily slow. Thus, dependence of convergence speed on properties of the exact solutions and also of the mapping F has to be investigated in detail to obtain information about validity of a regularization method. Upper bounds for the distance of regularized solutions to exact solutions in terms of some data noise level are the typical form to express such convergence rates.

As before we restrict our attention to quadratic mappings F mapping between real or complex separable Hilbert spaces. Denoting exact data in Y by $y^\dagger$ and by δ the positive noise level, available noisy data in Y will be denoted by y^δ and we assume that

$$\|y^\dagger - y^\delta\| \leq \delta.$$

If x_α^δ is a regularized solution obtained from some regularization method based on noisy data y^δ with regularization parameter α and if S is the set of all solutions to (1.1) with exact right-hand side $y^\dagger$, then we aim at estimates

$$\operatorname{dist}(x_\alpha^\delta, S)^2 = \mathcal{O}(\varphi(\delta)) \qquad \text{if } \delta \to 0. \tag{4.1}$$

Here, the function φ should be an index function:

Definition 4.1. A function $\varphi : [0, \infty) \to [0, \infty)$ is called *index function* if it is continuous, monotonically increasing, strictly increasing in a neighborhood of the origin and satisfies $\varphi(0) = 0$.

A sufficient condition for convergence rates (4.1), which has a much wider scope of applicability compared to previous techniques, was introduced in [HKPS07]. In principle one assumes that a certain inequality holds for all (or sufficiently many) elements of the space X. Several different names were given to this technique. Often such inequalities are referred to as 'variational inequalities', but this term conflicts with the already existing mathematical field with the same name. A second frequently used term is 'variational source condition', which rouses associations to classical source conditions. Because the new concept has no similarity to classical source conditions, most notably there is no source element, an alternative was introduced in the book [Fle12], which bases on the author's PhD thesis [Fle11]. There the term 'variational smoothness assumption' is used, because several kinds of smoothness (not only of the underlying exact solution as is the case for classical source conditions) are jointly described by one expression. In most recent literature 'variational source condition' seems to be used more often than 'variational inequality', whereas 'variational smoothness assumption' is not used by other authors. Thus, in the present thesis we use 'variational source condition' to name the technique described and applied below and, to avoid drawing to many parallels to source conditions, we read it as 'variational replacement for source conditions'.

The original concept of variational source conditions has been extended in [Pös08, HH09, Gei09, FH10, BH10, Gra10a, Fle10, Fle11], for details on those extensions we refer to [Fle12, page 37]. In this chapter we use the following form of variational source conditions: Fixing $y^\dagger$ and denoting by S the set of all norm minimizing solutions to (1.1) we say that a variational source condition is satisfied if there are a positive constant β and an index function φ such that

$$\beta\,\mathrm{dist}(x,S)^2 \leq \|x\|^2 - \|x^\dagger\|^2 + \varphi(\|F(x) - F(x^\dagger)\|) \qquad \text{for all } x \text{ in } X, \tag{4.2}$$

where $x^\dagger$ is an arbitrary element of S. Such a variational source condition is known to yield convergence rates (4.1) with concave φ if Tikhonov regularization is used to obtain regularized solutions, see, e. g., [Fle12, HM12]. But iterative methods can be used, too, see [HW13, Wer15].

From here on we restrict our considerations to quadratic mappings which are injective up to sign. Thus, given an exact right-hand side $y^\dagger$, there are only two solutions to (1.1). We denote them by $x^\dagger$ and $-x^\dagger$. The distance on the left-hand side of (4.2) then becomes

$$\mathrm{dist}(x,S)^2 = \min\{\|x - x^\dagger\|^2,\, \|x + x^\dagger\|^2\} = \|x\|^2 - 2\,|\mathrm{Re}\,\langle x, x^\dagger\rangle| + \|x^\dagger\|^2.$$

4.2. Nonlinearity conditions

The classical way to prove convergence rates for nonlinear inverse problems is to verify some nonlinearity condition for the mapping F, which in combination with a source condition yields the desired rate. The same way one can prove variational source conditions, which then imply the same rate. Here we briefly discuss nonlinearity conditions for quadratic mappings and in the next section we have a closer look at source conditions.

Starting with the paper [EKN89] a number of inequalities has been suggested to describe or restrict the nonlinear behavior of nonlinear mappings. We do not list those inequalities here, because in [BH15, Bür16] it was shown that only one of them can be verified for (truncated or untruncated) autoconvolution of functions with bounded support. Thus, we cannot expect them to hold for general quadratic mappings. The interested reader finds relevant references in [BH15].

Only the original nonlinearity condition used in [EKN89] (cf. equations (2.8) and (2.9) there) can be verified for autoconvolutions and also for general quadratic mappings:

Proposition 4.2. *For each $x^\dagger$ in X the inequality*

$$\|F(x) - F(x^\dagger) - F'[x^\dagger](x - x^\dagger)\| \leq \|F\| \, \|x - x^\dagger\|^2 \tag{4.3}$$

holds for all x in X.

Proof. By $F'[x^\dagger] \, x = 2 \, B_F(x^\dagger, x)$ we have

$$\|F(x) - F(x^\dagger) - F'[x^\dagger](x - x^\dagger)\| = \|F(x - x^\dagger)\|.$$

The assertion thus follows immediately from (1.3). □

4.3. Classical source conditions

The second ingredient used in [EKN89] to obtain convergence rates is a source condition. The authors considered Tikhonov regularization with penalty term $x \mapsto \|x - x_0\|^2$, where x_0 is some fixed reference element in X, and assumed that the exact solution $x^\dagger$ satisfies

$$x^\dagger - x_0 = F'[x^\dagger]^* \, v \tag{4.4}$$

for some v in Y. The element v is called source element. If v satisfies the smallness condition

$$2 \, \|F\| \, \|v\| < 1, \tag{4.5}$$

then one obtains that the distance between regularized solutions and $x^\dagger$ converges to zero with the rate $\sqrt{\delta}$ if α is chosen in the right way.

Verification of source conditions of the above type for autoconvolutions has been discussed in [BH15]. We extend the results to general quadratic mappings here.

Proposition 4.3. *For each reference element x_0 in X there exist $x^\dagger$ satisfying a source condition* (4.4) *with some v. If $x^\dagger$ and v satisfy* (4.4) *and* (4.5)*, then*

$$\begin{aligned} \operatorname{Re} \langle x^\dagger, x_0 \rangle \leq 0 \quad &\Leftrightarrow \quad x^\dagger = 0, \, x_0 = 0, \\ \operatorname{Re} \langle x^\dagger, x_0 \rangle > 0 \quad &\Leftrightarrow \quad \|x^\dagger - x_0\| < \|x^\dagger\|. \end{aligned}$$

In particular, $x_0 = 0$ implies $x^\dagger = 0$.

Proof. Let v satisfy (4.5) and define $G : X \to X$ by $G(x) := x_0 + F'[x]^* \, v$. Then G is a contraction, because

$$\begin{aligned} \|G(x) - G(u)\| &= 2 \, \|B_F^*(x - u, v)\| \leq 2 \, \|B_F^*(x - u, \cdot \,)\| \, \|v\| \\ &= 2 \, \|B_F(x - u, \cdot \,)\| \, \|v\| \leq 2 \, \|F\| \, \|v\| \, \|x - u\| \end{aligned}$$

holds for all x and u and we have $2 \, \|F\| \, \|v\| < 1$. Thus, Banach's fixed point theorem implies existence of $x^\dagger$ with $x^\dagger = G(x^\dagger)$, which is equivalent to (4.4).

Now let $x^\dagger$ and v satisfy (4.4). Then

$$\|x^\dagger - x_0\| = \|F'[x^\dagger]^* \, v\| = 2 \, \|B_F^*(x^\dagger, v)\| \leq 2 \, \|F\| \, \|v\| \, \|x^\dagger\| \leq \|x^\dagger\|$$

and strict inequality $\|x^\dagger - x_0\| < \|x^\dagger\|$ holds if $x^\dagger \neq 0$. Simple manipulations lead to

$$\|x_0\|^2 \leq 2\,\mathrm{Re}\,\langle x^\dagger, x_0\rangle$$

and $\|x_0\|^2 < 2\,\mathrm{Re}\,\langle x^\dagger, x_0\rangle$ if $x^\dagger \neq 0$. From these two inequalities one easily obtains the first equivalence. The second equivalence is not hard to prove, too, since $\mathrm{Re}\,\langle x^\dagger, x_0\rangle > 0$ implies $x^\dagger \neq 0$. □

The proposition shows that for non-trivial $x^\dagger$ the a priori chosen reference element x_0 has to be close enough to the unknown solution $x^\dagger$. Typically, no useful a priori information is available and one chooses $x_0 = 0$. Then source condition and smallness condition are only satisfied if $x^\dagger = 0$. Thus, the classical concept of source conditions for obtaining convergence rates in case of nonlinear inverse problems is not applicable for quadratic mappings.

4.4. Variational source conditions are the right tool

We now show that variational source conditions (4.2) are always satisfied in case of quadratic inverse problems and thus are a suitable tool for proving convergence rates (4.1). This results has not been published elsewhere.

Theorem 4.4. *Let F be quadratic, weakly continuous and injective up to sign. Then for each $x^\dagger$ in X and for each positive β with $\beta < 1$ there is a concave index function φ such that the variational source condition* (4.2) *holds on X.*

Proof. Fix $x^\dagger$ and note that due to injectivity up to sign we have $S = \{x^\dagger, -x^\dagger\}$ for the set of solutions in (4.2). Thus,

$$\begin{aligned}\mathrm{dist}(x,S)^2 &= \min\bigl\{\|x - x^\dagger\|^2,\ \|x + x^\dagger\|^2\bigr\}\\ &= \|x\|^2 + \|x^\dagger\|^2 + 2\,\min\{-\mathrm{Re}\,\langle x, x^\dagger\rangle,\ \mathrm{Re}\,\langle x, x^\dagger\rangle\}\\ &= \|x\|^2 + \|x^\dagger\|^2 - 2\,|\mathrm{Re}\,\langle x, x^\dagger\rangle|\end{aligned}$$

for all x in X.

The case $x^\dagger = 0$ will be excluded in the sequel because in this case a variational source condition holds with arbitrary index function φ.

To obtain a variational source condition (4.2) we use the concept of approximate variational source conditions introduced in [Fle12, Section 12.1.5] for a slightly different setting. For fixed β with $\beta < 1$ we define a distance function $D_\beta : [0,\infty) \to [0,\infty)$ by

$$D_\beta(r) := \sup_{x \in X}\bigl(\beta\,\mathrm{dist}(x,S) - \|x\|^2 + \|x^\dagger\|^2 - r\,\|F(x) - F(x^\dagger)\|\bigr).$$

We immediately see $0 \leq D_\beta(r) < \infty$ for all r and $D_\beta(0) > 0$, else $x^\dagger = 0$. Further, D_β is convex, monotonically decreasing and continuous. The distance function D_β expresses the violation of a variational source condition with linear φ and allows to derive a variational source condition with some (nonlinear) φ if $D_\beta(r) \to 0$ for $r \to \infty$. To see this we estimate

$$\begin{aligned}
&\beta \operatorname{dist}(x,S)^2 - \|x\|^2 + \|x^\dagger\|^2 \\
&\quad = \inf_{r\geq 0}\bigl(\beta \operatorname{dist}(x,S)^2 - \|x\|^2 + \|x^\dagger\|^2 - r\,\|F(x)-F(x^\dagger)\| + r\,\|F(x)-F(x^\dagger)\|\bigr) \\
&\quad \leq \inf_{r\geq 0}\bigl(D_\beta(r) + r\,\|F(x)-F(x^\dagger)\|\bigr)
\end{aligned}$$

and show that

$$\varphi(t) := \inf_{r\geq 0}\bigl(D_\beta(r) + r\,t\bigr), \qquad t \geq 0,$$

defines a concave index function. Obviously, $0 \leq \varphi(t) < \infty$ and φ is monotonically increasing. Since φ is an infimum of affine functions, it is concave, upper semi-continuous and continuous on $(0,\infty)$. Monotonicity and upper semi-continuity imply continuity on $[0,\infty)$. The decay of D_β to zero yields $\varphi(0) = 0$ and by $D_\beta(0) > 0$ we see $\varphi(t) > 0$ for $t > 0$, that is, φ is strictly increasing in a neighborhood of zero.

Next, we show that for each r the supremum in the definition of $D_\beta(r)$ is attained at some x. Rearranging the terms in the supremum and flipping the sign we obtain a functional

$$x \mapsto (1-\beta)\,\|x\|^2 - (1+\beta)\,\|x^\dagger\|^2 + 2\,\beta\,|\mathrm{Re}\,\langle x, x^\dagger\rangle| + r\,\|F(x)-F(x^\dagger)\|.$$

A sequence for which the values of the functional become arbitrarily close to the functional's infimum is bounded and thus has a weakly convergent subsequence. Weak lower semi-continuity of the functional shows that the limit of this subsequence is a minimizer of the functional. Consequently, the supremum in the definition of D_β is attained for each r.

Now let $(r_n)_{n\in\mathbb{N}}$ be a sequence in $[0,\infty)$ with $r_n \to \infty$ and let $(x_n)_{n\in\mathbb{N}}$ be a sequence of corresponding maximizers in the definition of D_β. To complete the proof we have to show $D_\beta(r_n) \to 0$. From

$$0 \leq D_\beta(r_n) \leq -(1-\beta)\,\|x_n\|^2 + (1+\beta)\,\|x^\dagger\|^2$$

we see that $(x_n)_{n\in\mathbb{N}}$ is bounded. Thus, there is a weakly convergent subsequence with limit $\tilde{x}$. The subsequence again will be denoted by $(x_n)_{n\in\mathbb{N}}$. Since x_n realizes the supremum in the definition of $D_\beta(r_n)$ and because we have $D_\beta(r_n) \geq 0$, we see

$$r_n\,\|F(x_n)-F(x^\dagger)\| \leq \beta \operatorname{dist}(x_n,S)^2 - \|x_n\|^2 + \|x^\dagger\|^2 \leq (1+\beta)\,\|x^\dagger\|^2.$$

This implies $F(x_n) \to F(x^\dagger)$ and together with the weak continuity of F we obtain $F(\tilde{x}) = F(x^\dagger)$, that is, $\tilde{x} = x^\dagger$ or $\tilde{x} = -x^\dagger$. Eventually,

$$\begin{aligned}
0 &\leq \liminf_{n\to\infty} D_\beta(r_n) \leq \limsup_{n\to\infty} D_\beta(r_n) = -\liminf_{n\to\infty}\bigl(-D_\beta(r_n)\bigr) \\
&= -\liminf_{n\to\infty}\bigl((1-\beta)\,\|x_n\|^2 - (1+\beta)\,\|x^\dagger\|^2 + 2\,\beta\,|\mathrm{Re}\,\langle x_n, x^\dagger\rangle| + r_n\,\|F(x_n)-F(x^\dagger)\|\bigr) \\
&\leq -\liminf_{n\to\infty}\bigl((1-\beta)\,\|x_n\|^2 - (1+\beta)\,\|x^\dagger\|^2 + 2\,\beta\,|\mathrm{Re}\,\langle x_n, x^\dagger\rangle|\bigr) \\
&\leq -\bigl((1-\beta)\,\|x^\dagger\|^2 - (1+\beta)\,\|x^\dagger\|^2 + 2\,\beta\,|\mathrm{Re}\,\langle x^\dagger, x^\dagger\rangle|\bigr) = 0,
\end{aligned}$$

which proves $D_\beta(r_n) \to 0$. □

Note that the technique used in the proof can also be applied to prove validity of variational source conditions in quite general Banach space settings. This will be done in Appendix C.

4.5. Sparsity yields variational source conditions

In the previous section we saw that there are always a constant β and an index function φ such that a variational source condition (4.2) is satisfied for a fixed quadratic mapping F. Now we demonstrate a method to obtain a concrete function φ depending on the behavior of the exact solutions and of the mapping F. The core of this section has been published in [BFH16].

First we consider general quadratic mappings and later we apply the results to autoconvolution of functions with uniformly bounded support. The following lemma provides a variational source condition for very specific quadratic mappings, which will be the basis for the more general result.

Lemma 4.5. *If there are an exact solution $x^\dagger$ to* (1.1) *and an orthonormal basis $(e_k)_{k\in\mathbb{N}}$ in X such that*

(i) $\big(F(e_k)\big)_{k\in\mathbb{N}}$ *is an orthonormal system in Y,*

(ii) $B_F(e_k,e_l)=0$ *for all k and l in $\mathbb{N}$ with $k\neq l$,*

(iii) $\{k\in\mathbb{N}:\,\langle x^\dagger,e_k\rangle\neq 0\}$ *is finite, that is, $x^\dagger$ is sparse with respect to $(e_k)_{k\in\mathbb{N}}$,*

then a variational source condition (4.2) *is satisfied with*

$$\beta=1 \qquad \textit{and} \qquad \varphi(t)=2\sqrt{n}\,t,$$

where n is the number of non-zero coefficients of $x^\dagger$.

Proof. Set $x_k^\dagger:=\langle x^\dagger,e_k\rangle$ and $x_k:=\langle x,e_k\rangle$ for all x and all k. Further set $f_k:=F(e_k)$. Then $(f_k)_{k\in\mathbb{N}}$ is an orthonormal system in Y. From

$$F(x)=F(x^\dagger) \qquad \Leftrightarrow \qquad (x_k^\dagger)^2=x_k^2 \quad \text{for all } k$$

we see that the set S of all solutions to (1.1) is

$$S=\{x\in X:\,x_k=x_k^\dagger \text{ or } x_k=-x_k^\dagger \text{ for all } k \text{ in } \mathbb{N}\}.$$

Thus,

$$\operatorname{dist}(x,S)^2=\sum_{k=1}^{\infty}\min\big\{|x_k-x_k^\dagger|^2,\,|x_k+x_k^\dagger|^2\big\}.$$

To simplify this expression further, for each x in X we define a sequence $(\xi_k(x))_{k\in\mathbb{N}}$ by

$$\xi_k(x):=\begin{cases}1, & \text{if } \operatorname{Re}(\overline{x_k}\,x_k^\dagger)\geq 0,\\ -1, & \text{else.}\end{cases}$$

Then we easily derive

$$\operatorname{dist}(x,S)^2=\sum_{k=1}^{\infty}|x_k-\xi_k(x)\,x_k^\dagger|^2. \tag{4.6}$$

Fix x and set

$$\tilde{x}^\dagger:=\sum_{k=1}^{\infty}\xi_k(x)\,x_k^\dagger\,e_k.$$

Obviously, $\tilde{x}^\dagger\in S$. From (i) and (ii) in the lemma we thus obtain

$$\begin{aligned}\|F(x)-F(x^\dagger)\|^2=\|F(x)-F(\tilde{x}^\dagger)\|^2&=\left\|\sum_{k=1}^{\infty}x_k^2\,f_k-\sum_{k=1}^{\infty}(\tilde{x}_k^\dagger)^2\,f_k\right\|^2\\&=\sum_{k=1}^{\infty}\left|x_k^2-(\tilde{x}_k^\dagger)^2\right|^2=\sum_{k=1}^{\infty}\left|x_k-\xi_k(x)\,x_k^\dagger\right|^2\left|x_k+\xi_k(x)\,x_k^\dagger\right|^2\end{aligned}$$

and a simple calculation shows

$$\left|x_k+\xi_k(x)\,x_k^\dagger\right|^2\geq\left|x_k^\dagger\right|^2.$$

Denoting by $I:=\{k\in\mathbb{N}:\,x_k^\dagger\neq0\}$ the support of $(x_k^\dagger)_{k\in\mathbb{N}}$ with cardinality n we obtain

$$\|F(x)-F(x^\dagger)\|^2\geq\sum_{k=1}^{\infty}\left|x_k-\xi_k(x)\,x_k^\dagger\right|^2\left|x_k^\dagger\right|^2=\sum_{k=1}^{\infty}\left|\left(x_k-\xi_k(x)\,x_k^\dagger\right)x_k^\dagger\right|^2$$

and by applying the Cauchy–Schwarz inequality and the triangle inequality we obtain the estimate

$$\begin{aligned}\|F(x)-F(x^\dagger)\|^2&\geq\frac{1}{n}\left(\sum_{k=1}^{\infty}\left|\left(x_k-\xi_k(x)\,x_k^\dagger\right)x_k^\dagger\right|\right)^2\\&\geq\frac{1}{n}\left|\sum_{k=1}^{\infty}\overline{\left(x_k-\xi_k(x)\,x_k^\dagger\right)}\,\xi_k(x)\,x_k^\dagger\right|^2=\frac{1}{n}\left|\left\langle\tilde{x}^\dagger,x-\tilde{x}^\dagger\right\rangle\right|^2.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\operatorname{dist}(x,S)^2-\|x\|^2+\|x^\dagger\|^2&=\|x-\tilde{x}^\dagger\|^2-\|x\|^2+\|\tilde{x}^\dagger\|^2\\&=-2\operatorname{Re}\left\langle\tilde{x}^\dagger,x-\tilde{x}^\dagger\right\rangle\leq2\left|\left\langle\tilde{x}^\dagger,x-\tilde{x}^\dagger\right\rangle\right|,\end{aligned}$$

completing the proof. □

Note that if F is the autoconvolution of periodic functions (see Subsection (1.2.1)), then (i) and (ii) in the lemma are satisfied with $(e_k)_{k\in\mathbb{N}}$ being the Fourier basis in $L^2(0,1)$.

Theorem 4.6. *Let $F:X\to Y$ be quadratic and injective up to sign and denote the solutions to* (1.1) *by $x^\dagger$ and $-x^\dagger$. If there exist an orthonormal basis $(e_k)_{k\in\mathbb{N}}$ in X and a bounded linear operator $A:Y\to Y$ such that* *(i)* $\big(A\,F(e_k)\big)_{k\in\mathbb{N}}$ *is an orthonormal system in* Y*,*

(ii) $A\,B_F(e_k,e_l)=0$ *for all* k *and* l *in* $\mathbb{N}$ *with* $k\neq l$*,*

(iii) $\{k\in\mathbb{N}:\,\langle x^\dagger,e_k\rangle\neq 0\}$ *is finite, that is,* $x^\dagger$ *is sparse with respect to* $(e_k)_{k\in\mathbb{N}}$*,*

then a variational source condition (4.2) *is satisfied for all* x *in*

$$M:=\{x\in X:\,\operatorname{Re}(\overline{x_k}\,x_k^\dagger)\geq 0 \text{ for all } k \text{ or } \operatorname{Re}(\overline{x_k}\,x_k^\dagger)\leq 0 \text{ for all } k\}$$

with

$$\beta=1 \qquad \text{and} \qquad \varphi(t)=2\,\|A\|\,\sqrt{n}\,t,$$

where n *is the number of non-zero coefficients of* $x^\dagger$*. The solutions* $x^\dagger$ *and* $-x^\dagger$ *are interior points of* M*.*

Proof. The solution $x^\dagger$ of (1.1) is also a solution of $A\,F(x)=A\,y^\dagger$, $x\in X$. Thus we can apply Lemma 4.5 to the quadratic mapping $A\,F$ and obtain the variational source condition

$$\operatorname{dist}(x,S_{A\,F})^2\leq\|x\|^2-\|x^\dagger\|^2+2\sqrt{n}\,\|A\,F(x)-A\,F(x^\dagger)\|$$

for all x in X, where $S_{A\,F}$ denotes the set of all solutions to $A\,F(x)=A\,y^\dagger$, $x\in X$. The estimate

$$\|A\,F(x)-A\,F(x^\dagger)\|\leq\|A\|\,\|F(x)-F(x^\dagger)\|$$

yields the right-hand side of the desired variational source condition.

From (4.6) in the proof of Lemma 4.5 we know that

$$\operatorname{dist}(x,S_{A\,F})^2=\sum_{k=1}^{\infty}|x_k-x_k^\dagger|^2=\|x-x^\dagger\|^2\geq\operatorname{dist}(x,\{x^\dagger,-x^\dagger\})^2$$

if $\operatorname{Re}(\overline{x_k}\,x^\dagger)\geq 0$ for all k. On the other hand, if $\operatorname{Re}(\overline{x_k}\,x^\dagger)\leq 0$ for all k, then

$$\operatorname{dist}(x,S_{A\,F})^2=\sum_{k=1}^{\infty}|x_k+x_k^\dagger|^2=\|x+x^\dagger\|^2\geq\operatorname{dist}(x,\{x^\dagger,-x^\dagger\})^2.$$

Thus, for all x in M we have

$$\operatorname{dist}(x,S_{A\,F})^2\geq\operatorname{dist}(x,\{x^\dagger,-x^\dagger\})^2,$$

which proves the variational source condition on M.

To show that $x^\dagger$ and $-x^\dagger$ are interior points of M we show that balls centered at the solutions with radius

$$r:=\min_{k\in\mathbb{N}:\,x_k^\dagger\neq 0}|x_k^\dagger|$$

lie in M. From $\|x-x^\dagger\|\leq r$ we obtain $|x_k-x_k^\dagger|\leq|x_k|$ for all k with $x_k^\dagger\neq 0$. Thus, $|x_k|^2-2\operatorname{Re}(\overline{x_k}\,x_k^\dagger)\leq 0$, which yields $\operatorname{Re}(\overline{x_k}\,x_k^\dagger)\geq 0$. Analogously, $\|x+x^\dagger\|\leq r$ implies $\operatorname{Re}(\overline{x_k}\,x_k^\dagger)\leq 0$. □

Note that the variational source condition in the theorem does not hold on the whole space, but only on a smaller set M. Nevertheless convergence rates can be obtained from such a variational source condition because the exact solutions are interior points of M and, thus, regularized solutions lie in M if the data noise level is small enough, see, e. g. [Fle12].

Example 4.7. We consider autoconvolution of functions with uniformly bounded support, that is, $X = L^2_{\mathbb{C}}(0,1)$, $Y = L^2_{\mathbb{C}}(0,2)$ and F is given by (1.5). Let $(e_k)_{k\in\mathbb{N}}$ be the Fourier basis as described in Subsection 3.5 and let $A : Y \to Y$ be defined by

$$(A\,y)(s) := y(s) + y(s+1), \qquad s \in (0,1).$$

Then one easily verifies that $A\,F$ is the autoconvolution of periodic functions defined in (1.8) and that $\|A\| \leq \sqrt{2}$. Thus, Theorem 4.6 is applicable because the convolution theorem yields assumptions (i) and (ii) in the above theorem. □

Lemma 4.5 can be extended to non-sparse solutions $x^\dagger$. With a technique very similar to the proof of Theorem 8.9 in Part II we would obtain a variational source condition (4.2) with some positive β and with an index function

$$\varphi(t) = \inf_{n\in\mathbb{N}} \left(\frac{1}{1-\beta} \sum_{|k|>n} |x^\dagger_k|^2 + 2\sqrt{2\,n+1}\,t \right), \qquad t \geq 0, \tag{4.7}$$

which is essentially determined by the decay of the coefficients of $x^\dagger$. The variational source condition on a subset M of X in Theorem 4.6 can be proven for non-sparse $x^\dagger$, too. But $x^\dagger$ and $-x^\dagger$ are no longer interior points of M. Thus, we cannot ensure that regularized solutions belong to M and the usual convergence rates proof does not work.

Part II.

Sparsity promoting regularization

5. Aren't all questions answered?

Sparsity promoting regularization techniques for linear operator equations, which will be introduced in the next chapter, were a very active field of research during the first ten years of the new millennium. Methods were suggested, algorithms were developed and the theoretical backing had been extended extensively. The major sparsity promoting regularization method is ℓ^1-regularization and this method is used in many applications today and is an integral part of signal processing tool boxes.

Research in the inverse problems community started in 2004 with the influential paper [DDDM04] and went on for several years. From about 2010 on the focus changed to other, more involved sparsity promoting methods including regularization with non-convex penalties and it seemed that the theory of ℓ^1-regularization was more or less complete, at least complete enough to ensure proper behavior of corresponding algorithms in applications.

In the second part of this thesis we resume research on the fundamentals of ℓ^1-regularization and answer questions which were not posed in first years, but are of importance for understanding the method and for verifying its applicability in a wide field of problems. These questions include:

- What happens if the frequent assumption, that the sought-for solution has a sparse representation in a certain basis, fails? Is ℓ^1-regularization capable of producing sufficiently precise approximations to such non-sparse solutions?

- Next to sparsity of the solution, which additional conditions have to be satisfied to bound the distance between ℓ^1-regularized and exact solutions in terms of the data noise level? What is the weakest set of such conditions?

- What about non-injective operators? Is it possible to prove error estimates if the underlying linear operator is not injective, even not in a weakened sense?

These questions will be answered in detail in the following chapters. The presented results were published in [BFH13, FH15, FHV15, FHV16, Fle16, FG17].

If we would aim at brevity, we could formulate a very technical theorem first and then derive all results as corollaries. Instead, to increase readability, we start with the less technical results and then go step by step to the more involved questions and their answers. Following this longer path allows to reconstruct the core ideas of the final theorem on convergence rates for ℓ^1-regularization with non-injective operators, which else would be covered by technical and notational difficulties.

Depending on the reader's knowledge in the fields of set theoretic topology, functional analysis and convex analysis it might be beneficial to first look into Appendix A prior to reading the next chapters.

6. Sparsity and ℓ^1-regularization

In this section we describe the basic ideas of sparsity promoting regularization techniques, introduce ℓ^1-regularization and briefly discuss alternative methods. In addition we have a first look at several examples which will appear again later on in the text.

6.1. Sparse signals

Let $\tilde{X}$ and Y be real Banach spaces. For fixed $y^\dagger$ in Y we aim to solve the operator equation

$$\tilde{A}\,\tilde{x} = y^\dagger \tag{6.1}$$

for $\tilde{x}$ in $\tilde{X}$, where $A : \tilde{X} \to Y$ is assumed to be linear and bounded. If this equation is ill-posed, we have to take into account possible inaccuracies in the right-hand side, that is, only some y^δ in Y with

$$\|y^\delta - y^\dagger\|_Y \leq \delta \tag{6.2}$$

is accessible. Here, the positive constant δ denotes the noise level.

To stabilize the inversion process additional information on the exact solutions are required. This is the point where sparsity comes into play. Especially in signal processing applications, which include all kinds of image processing, one often knows a priori that the sought-for solution to (6.1) is sparse or almost sparse in the following sense.

Definition 6.1. Let $(\tilde{e}_k)_{k\in\mathbb{N}}$ be a Schauder basis in $\tilde{X}$ and denote be $(\tilde{x}_k)_{k\in\mathbb{N}}$ corresponding coefficients of some $\tilde{x}$ in $\tilde{X}$. The Banach space element $\tilde{x}$ is *sparse* with respect to the basis $(\tilde{e}_k)_{k\in\mathbb{N}}$ if only finitely many coefficients do not vanish. The element $\tilde{x}$ is *almost sparse* with respect to that basis if $(\tilde{x}_k)_{k\in\mathbb{N}}$ belongs to ℓ^1.

As usual, ℓ^1 denotes the Banach space of absolutely summable real sequences. If the underlying space $\tilde{X}$ is a sequence space and no basis is explicitly specified, then sparsity is considered with respect to standard basis $(e^{(k)})_{k\in\mathbb{N}}$, where

$$e_l^{(k)} := \begin{cases} 1, & \text{if } l = k, \\ 0, & \text{else.} \end{cases}$$

Note that existence of Schauder bases is not guaranteed for each Banach space. Obviously, if there is a Schauder basis, then the space is separable. However, there are separable Banach spaces which do not have a Schauder basis (see [Die84, page 35] and references therein). A Schauder basis $(\tilde{e}_k)_{k\in\mathbb{N}}$ might be unbounded, but yields a bounded Schauder basis if each element $\tilde{e}_k$ is divided by its norm $\|\tilde{e}_k\|_{\tilde{X}}$. Thus, existence of Schauder bases is equivalent to existence of bounded Schauder bases. But note that normalizing a Schauder basis as described changes the set of almost sparse elements, whereas the set of sparse elements remains unchanged.

In most applications (and publications) $\tilde{X}$ is a separable Hilbert space and $(\tilde{e}_k)_{k\in\mathbb{N}}$ is an orthonormal basis. Wavelet bases are a common choice. We do not restrict our considerations to Hilbert spaces, because, on the one hand, the Banach space setting does not imply ponderable additional expenses and, on the other hand, it allows us to test our results and their limitations with a much wider class of examples.

Sparsity is a property of the coefficient sequence with respect to a basis and not of a Banach space element itself. Thus, we may replace $\tilde{X}$ by some sequence space. A good choice is ℓ^1 for two reasons: first, it is large enough to contain the coefficient sequences of all almost sparse elements, and second, it is small enough to ensure that each sequence is indeed a coeffcient sequence of some element in $\tilde{X}$. More precisely, we could say that ℓ^1 contains exactly those sequences which are coefficient sequences of almost sparse elements in $\tilde{X}$. Because we are looking for (almost) sparse solutions to (6.1), there is no need to consider coefficient sequences which do not belong to ℓ^1.

Proposition 6.2. *Let $(\tilde{e}_k)_{k\in\mathbb{N}}$ be a Schauder basis in $\tilde{X}$ and define $L : \ell^1 \to \tilde{X}$ by*

$$L\,x := \sum_{k=1}^{\infty} x_k\,\tilde{e}_k \tag{6.3}$$

for x in ℓ^1. Then L is linear, injective and bounded with

$$\|L\|_{\mathcal{L}(\ell^1,\tilde{X})} = \sup_{k\in\mathbb{N}} \|\tilde{e}_k\|_{\tilde{X}}.$$

Proof. Linearity and injectivity are obvious. Boundedness and the upper bound for the norm of L follow from

$$\|L\,x\|_{\tilde{X}} \leq \sum_{k=1}^{\infty} |x_k|\,\|\tilde{e}_k\|_{\tilde{X}} \leq \|x\|_{\ell^1} \sup_{k\in\mathbb{N}} \|\tilde{e}_k\|_{\tilde{X}}$$

for x in ℓ^1. Choose $x = \tilde{e}_k$ to see that the upper bound is also a lower bound for the norm of L. □

Now define the bounded linear operator $A : \ell^1 \to Y$ by $A := \tilde{A}\,L$ and consider the equation

$$A\,x = y^{\dagger}, \qquad x \in \ell^1. \tag{6.4}$$

This is the equation we are going to deal with in the remaining chapters of this part of the thesis. If we know a solution to (6.4), then applying L yields a solution to the original equation (6.1).

6.2. ℓ^1-regularization

If equation (6.4) is ill-posed, exact solutions are not accessible, because they do not depend continuously on the (noisy) right-hand side. Instead we have to seek for approximate but stable solutions. Knowing a priori that the sought-for solutions to (6.4) are sparse, it is sensible to look for sparse approximations. One way to obtain such sparse approximations is to minimize the Tikhonov-type functional

$$T_\alpha^\delta(x) := \|A\,x - y^\delta\|_Y^p + \alpha\,\|x\|_{\ell^1} \tag{6.5}$$

over x in ℓ^1. Here, p is some positive exponent for simplifying numerical minimization. If Y is a Hilbert space, then $p = 2$ is a good choice. In any case we assume $p > 1$. The trade-off between data fitting and stabilization is controlled by the positive regularization parameter α. In [DDDM04] an algorithm for finding the minimizers numerically had been proposed. That article popularized ℓ^1-regularization in the inverse problems community and motivated numerous further publications on the subject.

The minimizers of T_α^δ are referred to as ℓ^1-regularized solutions. To show that this method is well-defined and regularizing and that the minimizers are sparse, we formulate the theorem below. The assertions of the theorem are well known in the literature and we repeat them here for the sake of completeness. The basic assumption required for the theorem can be stated in several equivalent ways as shown in the lemma. As usual, c_0 denotes the Banach space of real sequences converging to zero. Remember that $(c_0)^* = \ell^1$.

Lemma 6.3. *The following assertions are equivalent.*

(i) $(A\,e^{(k)})_{k\in\mathbb{N}}$ *converges weakly to zero.*

(ii) $\mathcal{R}(A^*) \subseteq c_0$.

(iii) A *is weak*-to-weak continuous.*

(iv) A *is sequentially weak*-to-weak continuous.*

Proof. Let (i) be satisfied. Then for each $A^*\,\eta$ from $\mathcal{R}(A^*)$ we have

$$[A^*\,\eta]_k = \langle A^*\,\eta, e^{(k)}\rangle_{\ell^\infty\times\ell^1} = \langle \eta, A\,e^{(k)}\rangle_{Y^*\times Y} \to 0 \quad \text{if } k\to\infty,$$

that is, $A^*\,\eta \in c_0$.

Now let (ii) be true. If we take a weakly* convergent net $(x^{(\kappa)})_{\kappa\in N}$ with limit x, then $\langle \eta, A\,x^{(\kappa)}\rangle_{Y^*\times Y} = \langle A^*\,\eta, x^{(\kappa)}\rangle_{\ell^\infty\times\ell^1}$ and, since $A^*\,\eta$ belongs to c_0 and ℓ^1 is the dual of c_0, we may write $\langle A^*\,\eta, x^{(\kappa)}\rangle_{\ell^\infty\times\ell^1} = \langle x^{(\kappa)}, A^*\,\eta\rangle_{\ell^1\times c_0}$. Thus,

$$\lim_\kappa \langle \eta, A\,x^{(\kappa)}\rangle_{Y^*\times Y} = \lim_\kappa \langle x^{(\kappa)}, A^*\,\eta\rangle_{\ell^1\times c_0} = \langle x, A^*\,\eta\rangle_{\ell^1\times c_0},$$

showing

$$\lim_\kappa \langle \eta, A\,x^{(\kappa)}\rangle_{Y^*\times Y} = \langle \eta, A\,x\rangle_{Y^*\times Y} \quad \text{for all } \eta\in Y^*.$$

Finally, (iii) immediately implies (iv) and from (iv) and the obvious fact that $(e^{(k)})_{k\in\mathbb{N}}$ converges weakly* to zero we immediately obtain (i). □

Theorem 6.4. *Let A be weak*-to-weak continuous. Then the following assertions are true.*

(i) Existence: There exist solutions to (6.4) *with minimal norm (referred to as norm minimizing solutions) and there exist minimizers of the Tikhonov-type functional* (6.5)*. Further, all minimizers of T_α^δ are sparse.*

(ii) Stability: If $(y_k)_{k\in\mathbb{N}}$ converges to y^δ and if $(x^{(k)})_{k\in\mathbb{N}}$ is a corresponding sequence of minimizers of (6.5) *with y^δ replaced by y_k, then this second sequence has a weakly* convergent subsequence and each weakly* convergent subsequence converges weakly* to a minimizer of T_α^δ.*

(iii) Convergence: If $(\delta_k)_{k\in\mathbb{N}}$ converges to zero and if $(y_k)_{k\in\mathbb{N}}$ satisfies $\|y_k - y^\dagger\| \leq \delta_k$, then there is a sequence $(\alpha_k)_{k\in\mathbb{N}}$ such that each corresponding sequence of minimizers of $T_{\alpha_k}^{\delta_k}$ contains a weakly convergent subsequence. Each such subsequence converges in norm to some norm minimizing solution of* (6.4)*.*

Proof. Closed balls in ℓ^1 are weakly* compact, see [Meg98, Theorem 2.6.18], which guarantees that bounded sequences have a weak* accumulation point and that the ℓ^1-norm is weakly* lower semi-continuous. Taking also the weak*-to-weak continuity of A and the weak lower semi-continuity of the norm in Y into account we may apply standard results on Tikhonov-type regularization methods in Banach spaces, see [Fle12, SKHK12]. Note that the ℓ^1-norm satisfies the so called weak* Kadec-Klee property, which yields convergence in norm in item (iii) of the theorem.

It only remains to show that each minimizer of (6.5) has only finitely many non-zero components. This is a consequence of $\mathcal{R}(A^*) \subseteq c_0$ (cf. Lemma 6.3). By standard arguments from convex analysis we see that some x is a minimizer of T_α^δ if and only if there is some ξ in ℓ^∞ such that

$$-\xi \in \alpha\,\partial\|\cdot\|(x) \quad \text{and} \quad \xi \in A^*\,\partial(\|A\,\cdot\, - y^\delta\|^p)(x).$$

Thus, $|\xi_k| = \alpha$ whenever $x_k \neq 0$ and ξ belongs to c_0. This is only possible if x has at most finitely many non-zero components. □

The assumption that A is weak*-to-weak continuous is very weak, because it is automatically satisfied if A has a bounded extension to some ℓ^q-space with $q > 1$. To see this, let $E_q : \ell^1 \to \ell^q$ be the bounded embedding of ℓ^1 into ℓ^q and let $A_q : \ell^q \to Y$ be the bounded extension of A to ℓ^q. Then $A\,e^{(k)} = A_q\,E_q\,e^{(k)} = A_q\,e^{(k)}$ and $(e^{(k)})_{k\in\mathbb{N}}$ converges weakly to zero in ℓ^q. The weak-to-weak continuity of bounded linear operators thus implies $A_q\,e^{(k)} \rightharpoonup 0$.

If the original space $\tilde{X}$ (cf. previous section) is a Hilbert space and $(\tilde{e}_k)_{k\in\mathbb{N}}$ is an orthonormal basis, then A has a bounded extension to ℓ^2 and thus is weak*-to-weak continuous.

Nevertheless there are bounded linear operators which are not weak*-to-weak continuous. An example is the identity mapping $A = I$ if $Y = \ell^1$. In ℓ^1 weak sequential convergence coincides with norm convergence, but not with weak* convergence. Thus, the identity mapping cannot be sequentially weak*-to-weak continuous. To cover also this special case, we may weaken the assumption of weak*-to-weak continuity slightly by demanding only weak*-to-weak* continuity. This is possible if Y is the dual space of some other Banach space. If Y is reflexive then weak and weak* convergence coincide.

If Y is not reflexive, as it is the case for $Y = \ell^1$, then weak convergence implies weak* convergence but not vice versa.

The above lemma can be modified as follows to deal with the weaker condition. The implication from (ii) to (iv) in the lemma was formulated and proven by Bernd Hofmann (TU Chemnitz) and up to now has not been published elsewhere.

Lemma 6.5. *Let Z be a Banach space and let $Y = Z^*$. Then the following assertions are equivalent.*

(i) $(A\,e^{(k)})_{k\in\mathbb{N}}$ converges weakly to zero.*

(ii) $\mathcal{R}(A^*|_Z) \subseteq c_0$.

(iii) A is weak-to-weak* continuous.*

(iv) A is sequentially weak-to-weak* continuous.*

Proof. Let (i) be satisfied. Then for each $A^*\,\eta$ from $\mathcal{R}(A^*|_Z)$, that is, $\eta \in Z \subseteq Y^*$, we have

$$[A^*\,\eta]_k = \langle A^*\,\eta, e^{(k)}\rangle_{\ell^\infty\times\ell^1} = \langle \eta, A\,e^{(k)}\rangle_{Y^*\times Y} = \langle A\,e^{(k)}, \eta\rangle_{Z^*\times Z} \to 0 \quad \text{if } k \to \infty,$$

that is, $A^*\,\eta \in c_0$.

Now let (ii) be true. If we take a weakly* convergent net $(x^{(\kappa)})_{\kappa\in N}$ with limit x, then $\langle A\,x^{(\kappa)}, \eta\rangle_{Z^*\times Z} = \langle \eta, A\,x^{(\kappa)}\rangle_{Y^*\times Y} = \langle A^*\,\eta, x^{(\kappa)}\rangle_{\ell^\infty\times\ell^1}$ for all η in Z and, since $A^*\,\eta$ belongs to c_0 and ℓ^1 is the dual of c_0, we may write $\langle A^*\,\eta, x^{(\kappa)}\rangle_{\ell^\infty\times\ell^1} = \langle x^{(\kappa)}, A^*\,\eta\rangle_{\ell^1\times c_0}$. Thus,

$$\lim_\kappa \langle A\,x^{(\kappa)}, \eta\rangle_{Z^*\times Z} = \lim_\kappa \langle \eta, A\,x^{(\kappa)}\rangle_{Y^*\times Y} = \lim \kappa \langle x^{(\kappa)}, A^*\,\eta\rangle_{\ell^1\times c_0} \to \langle x, A^*\,\eta\rangle_{\ell^1\times c_0},$$

showing

$$\lim_\kappa \langle A\,x^{(\kappa)}, \eta\rangle_{Z^*\times Z} = \langle A\,x, \eta\rangle_{Z^*\times Z} \quad \text{for all } \eta \in Z.$$

Finally, (iii) immediately implies (iv) and from (iv) and the obvious fact that $(e^{(k)})_{k\in\mathbb{N}}$ converges weakly* to zero we immediately obtain (i). □

The above theorem on existence, stability and convergence remains true, except for sparsity of minimizers, if weak*-to-weak continuity is replaced by weak*-to-weak* continuity, because the proof only relies on the fact that $\|A\ \cdot\ - y^\delta\|$ is weakly* lower semi-continuous. This is the case in both variants, with or without *, since the norm functional is weakly and also weakly* lower semi-continuous.

Note that the case $Y = \ell^1$, $A = I$ is mainly of theoretical interest. It is a tool for exploring the frontiers of the theoretic framework we have chosen for investigating ℓ^1-regularization. For practical applications it is irrelevant because one easily verifies that with the natural choice $p = 1$ in (6.5) the ℓ^1-regularized solutions coincide with the data y^δ if $\alpha < 1$.

For the sake of completeness we mention that there exist bounded linear operators which not even are weak*-to-weak* continuous.

Example 6.6. If $Y = \ell^1$ and

$$[A\,x]_k := \begin{cases} \sum\limits_{l=1}^{\infty} x_l, & \text{if } k = 1, \\ x_k, & \text{else,} \end{cases}$$

for all k in $\mathbb{N}$ and all x in ℓ^1, then $A\,e^{(k)} = e^{(1)} + e^{(k)}$ if $k > 1$. Thus, $A\,e^{(k)} \rightharpoonup^* e^{(1)}$ but $e^{(k)} \rightharpoonup^* 0$. The same operator A considered as mapping into $Y = \ell^2$ is an example of a not weak*-to-weak continuous bounded linear operator in the classical Hilbert space setting for ℓ^1-regularization. □

We close this section with an observation about the maximum sensible regularization parameter α. An analogous observation was made in Part I for standard Hilbert space Tikhonov regularization applied to quadratic mappings, cf. Proposition 2.3. To the author's best knowledge the following proposition does not appear in the literature, although its proof is quite simple.

Proposition 6.7. *Let Y be a Hilbert space, let $p = 2$ in* (6.5) *and set*

$$\alpha_{\max} := 2 \sup_{\substack{x \in X \\ \|x\| \leq 1}} \langle A\,x, y^\delta \rangle_{Y \times Y}.$$

If $\alpha \geq \alpha_{\max}$, then

$$0 \in \operatorname*{argmin}_{x \in X} T_\alpha^\delta(x).$$

If, in addition, $\alpha > \alpha_{\max}$ or A is injective, then

$$\operatorname*{argmin}_{x \in X} T_\alpha^\delta(x) = \{0\}.$$

Proof. If $x \neq 0$ we have

$$\begin{aligned} T_\alpha^\delta(x) &= \|A\,x\|_Y^2 - 2\,\langle A\,x, y^\delta \rangle_{Y \times Y} + \|y^\delta\|_Y^2 + \alpha\,\|x\|_{\ell^1} \\ &\geq \|A\,x)\|_Y^2 - 2\,\langle A\,x, y^\delta \rangle_{Y \times Y} + \|y^\delta\|_Y^2 + 2\left\langle A\left(\frac{x}{\|x\|_{\ell^1}}\right), y^\delta \right\rangle_{Y \times Y} \|x\|_{\ell^1} \\ &= \|A\,x\|_Y^2 + \|y^\delta\|_Y^2 \geq \|y^\delta\|_Y^2 = T_\alpha^\delta(0), \end{aligned}$$

proving the first assertion. If $\alpha > \alpha_{\max}$, the first inequality sign is strict. If A is injective, $x \neq 0$ implies $A\,x \neq 0$, making the second inequality sign a strict one. □

Remark 6.8. If $\alpha_{\max}$ is chosen greater than in the proposition, the proposition remains true. An easy to calculate replacement is $2\,\|A\|\,\|y^\delta\|_Y$.

Note that some of the results here and in the sequel can also be derived for nonlinear equations, if the nonlinearity is controlled by additional assumptions, see [BH13]

6.3. Other sparsity promoting regularization methods

We briefly mention three other regularization methods which, under suitable assumptions, produce sparse approximate solutions to linear operator equations (6.4). The first is the residual method, followed by so called non-convex regularization and the elastic net method. All three methods are closely related to ℓ^1-regularization.

The residual method consists in solving the minimization problem

$$\|x\|_{\ell^1} \to \min \quad \text{subject to} \quad \|A\,x - y^\delta\|_Y \leq \delta.$$

In finite dimensions the method is investigated, for instance, in [CRT06]. For results in a very general infinite-dimensional setting we refer to [GHS11b]. Under suitable assumptions, the residual method yields the same approximate solutions as ℓ^1-regularization with α chosen by the discrepancy principle, see [GHS11b, page 2] for a reference.

The term non-convex regularization typically refers to Tikhonov-type methods with non-convex penalty. That is, minimization problems of the form

$$\|A\,x - y^\delta\|_Y^p + \alpha\,R(x) \to \min_{x\in\ell^1}$$

are considered, where $R : \ell^1 \to [0,\infty]$ is some non-convex stabilizing functional. A common choice for R is

$$R(x) := \sum_{k=1}^{\infty} |x_k|^q$$

with q in $(0,1)$ and $R(x) = \infty$ if the series does not converge. Such methods yield sparse minimizers, but are numerically very challenging. For details and references we refer to [BL09, Gra10b].

Elastic net regularization is an extension of ℓ^1-regularization which aims at simplifying numerical minimization of the Tikhonov-type functional. The idea is to add a second, smooth regularization term, typically the ℓ^2 norm:

$$\|A\,x - y^\delta\|_Y^p + \alpha_1\,\|x\|_{\ell^1} + \alpha_2\,\|x\|_{\ell^2}^2 \to \min_{x\in\ell^1}.$$

Additional difficulties arise from the fact that here two regularization parameters have to be chosen in the right way to guarantee proper behavior of the method. For details we refer to [CHZ17] and references therein.

6.4. Examples

In this chapter we introduce several examples which will be discussed in more detail in subsequent chapters to demonstrate certain features of the obtained results.

6.4.1. Denoising

Reconstructing a sparse signal from noisy measurements is an important application of ℓ^1-regularization. Reduced to sequence spaces, that is, after representing the signal with respect to a suitable basis, the task is to find a sparse approximation x to given y^δ.

The noise is typically measured in a norm weaker than the ℓ^1-norm. Thus, let $Y := \ell^p$ with p in $(1, \infty)$. The operator A in (6.4) is the bounded embedding $E_p : \ell^1 \to \ell^p$ of ℓ^1 into ℓ^p.

Because E_p obviously has a bounded extension to ℓ^p, the operator is weak*-to-weak continuous, cf. discussion in Section 6.2. Thus, Theorem 6.4 applies. For later reference we note

$$\mathcal{R}(A^*) = \left\{\xi \in \ell^\infty : \sum_{k=1}^{\infty} |\xi_k|^{\frac{p}{p-1}} < \infty\right\}.$$

6.4.2. Bidiagonal operator

In [FH15, Example 2.6] a certain bidiagonal operator was considered. Set $Y = \ell^2$ and define $A_2 : \ell^2 \to \ell^2$ by

$$[A_2\, x]_k := \frac{x_k - x_{k+1}}{k}$$

for all k in $\mathbb{N}$ and all x in ℓ^2. Let A be the restriction of A_2 to ℓ^1. Both operators are bounded and injective. From

$$\|A_2\, e^{(1)}\|_Y^2 = \|e^{(1)}\|_Y^2 = 1$$

and

$$\|A_2\, e^{(k)}\|_Y^2 = \left\|\frac{1}{k}\, e^{(k)} - \frac{1}{k-1}\, e^{(k-1)}\right\|_Y^2 = \frac{1}{k^2} + \frac{1}{(k-1)^2} \qquad \text{if } k \geq 2$$

we see that $\sum_{k=1}^{\infty} \|A_2\, e^{(k)}\|_Y^2$ converges, that is, A_2 is a Hilbert–Schmidt operator and thus compact. Since the embedding of ℓ^1 into ℓ^2 is bounded, A is compact, too. Existence of the extension A_2 ensures weak*-to-weak continuity of A.

The adjoint operator $A^* : \ell^\infty \to \ell^1$ has the explicit representation

$$[A^*\, \eta]_1 = \eta_1 \qquad \text{and} \qquad [A^*\, \eta]_k = \frac{\eta_k}{k} - \frac{\eta_{k-1}}{k-1} \quad \text{if } k \geq 2$$

for all η in ℓ^2. For ξ in $\mathcal{R}(A^*)$ we have

$$\eta_k = k \sum_{l=1}^{k} \xi_l$$

for all k in $\mathbb{N}$. In particular, we see

$$\xi \in \mathcal{R}(A^*) \qquad \Leftrightarrow \qquad \left(k \mapsto k \sum_{l=1}^{k} \xi_l\right) \in \ell^2 \qquad \Rightarrow \qquad \sum_{k=1}^{\infty} \xi_k = 0.$$

6.4.3. Simple integration and Haar wavelets

The next example can be found in [FH15, Section 4]. With the notation of Section 6.1 we set $\tilde{X} = L^2(0,1)$ and $Y = L^2(0,1)$ and define $\tilde{A} : \tilde{X} \to Y$ by

$$(\tilde{A}\, \tilde{x})(s) := \int_0^s \tilde{x}(t)\, \mathrm{d}t, \quad s \in (0,1),$$

for $\tilde{x}$ in $\tilde{X}$. This operator is linear, injective and bounded. The adjoint $\tilde{A}^* : L^2(0,1) \to L^2(0,1)$ is given by

$$(\tilde{A}^* \eta)(t) = \int\limits_t^1 \eta(s)\,\mathrm{d}s, \quad t \in (0,1).$$

The range of $\tilde{A}^*$ consists exactly of those functions $\tilde{\xi}$ in the Sobolev space $H^1(0,1)$ which satisfy $\lim_{t\to 1-0} \tilde{x}(t) = 0$.

As basis, with respect to which we want to consider sparsity, we choose the Haar basis. The first element of the Haar system is given by $\tilde{e}^{(1)}(s) := 1$ for s in $(0,1)$. All other elements are scaled and translated versions of the function

$$\psi(s) := \begin{cases} 1, & \text{if } s \in (0,\frac{1}{2}), \\ -1, & \text{if } s \in (\frac{1}{2},1). \end{cases}$$

More precisely,

$$\tilde{e}^{(1+2^l+k)}(s) := \psi_{l,k}(s) := 2^{\frac{l}{2}}\,\psi(2^l\,s-k), \quad s \in (0,1),$$

for $l = 0,1,2\ldots$ and $k = 0,1,\ldots,2^l-1$.

Given the synthesis operator

$$E : \ell^1 \to L^2(0,1), \qquad E\,x := \sum_{k=1}^{\infty} x_k\,e^{(k)}$$

we define $A : \ell^1 \to L^2(0,1)$ by $A := \tilde{A}\,E$. Since the Haar system is an orthonormal basis in $L^2(0,1)$, the operator A has a bounded extension to ℓ^2 and, thus, is weak*-to-weak continuous. The adjoint $A^* : L^2(0,1) \to \ell^\infty$ is given by

$$[A^*\,\eta]_k = \langle \tilde{A}^*\,\eta, \tilde{e}^{(k)}\rangle_{L^2(0,1)\times L^2(0,1)}$$

for all k in $\mathbb{N}$ and for all η in $L^2(0,1)$.

6.4.4. Simple integration and Fourier basis

Let $Y := \ell^2$ and let $A := P\,V\,\tilde{A}\,U$ be the composition of the Fourier synthesis operator $U : \ell^1 \to L^2(0,1)$ defined by

$$(U\,x)(t) := x_1 + \sqrt{2}\sum_{l=1}^{\infty} x_{2\,l}\,\cos(2\,\pi\,l\,t) + \sqrt{2}\sum_{l=1}^{\infty} x_{2\,l+1}\,\sin(2\,\pi\,l\,t), \qquad t \in (0,1),$$

the integration operator $\tilde{A} : L^2(0,1) \to L^2(0,1)$ defined by

$$(\tilde{A}\,\tilde{x})(s) := \int\limits_0^s \tilde{x}(t)\,\mathrm{d}t, \qquad s \in (0,1),$$

the Fourier transform $V : L^2(0,1) \to \ell^2$ defined by

$$\begin{aligned}[V\,\tilde{y}]_1 &:= \int_0^1 \tilde{y}(s)\,\mathrm{d}s,\\ [V\,\tilde{y}]_{2\,l} &:= \int_0^1 \tilde{y}(s)\,\sqrt{2}\,\cos(2\,\pi\,l\,s)\,\mathrm{d}s,\\ [V\,\tilde{y}]_{2\,l+1} &:= \int_0^1 \tilde{y}(s)\,\sqrt{2}\,\sin(2\,\pi\,l\,s)\,\mathrm{d}s\end{aligned}$$

for l in $\mathbb{N}$, and the projection $P : \ell^2 \to \ell^2$ defined by

$$\begin{aligned}[P\,\bar{y}]_1 &= \bar{y}_1,\\ [P\,\bar{y}]_{2\,l} &= 0,\\ [P\,\bar{y}]_{2\,l+1} &= \bar{y}_{2\,l} + \bar{y}_{2\,l+1}\end{aligned}$$

for l in $\mathbb{N}$. In other words, we aim to reconstruct derivatives of functions from incomplete Fourier data under the a priori information that the derivatives are sparse or almost sparse with respect to the Fourier basis. Only sums of the data's cosine and sine coefficients are available, making the operator highly non-injective.

The operator $A : \ell^1 \to \ell^2$ turns out to map a sequence x to a sequence $A\,x$ defined by

$$\begin{aligned}[A\,x]_1 &= \frac{1}{2}\,x_1 + \sum_{l=1}^{\infty} \frac{1}{\sqrt{2}\,\pi\,l}\,x_{2\,l+1},\\ [A\,x]_{2\,l} &= 0,\\ [A\,x]_{2\,l+1} &= \frac{1}{2\,\pi\,l}\left(-\sqrt{2}\,x_1 + x_{2\,l} - x_{2\,l+1}\right)\end{aligned}$$

for l in $\mathbb{N}$. The adjoint $A^* = P^*\,V^*\,\tilde{A}^*\,U^* : \ell^2 \to \ell^\infty$ thus is given by

$$\begin{aligned}[A^*\,\eta]_1 &= \frac{1}{2}\,\eta_1 - \sum_{l=1}^{\infty} \frac{1}{\sqrt{2}\,\pi\,l}\,\eta_{2\,l+1},\\ [A^*\,\eta]_{2\,l} &= \frac{1}{2\,\pi\,l}\,\eta_{2\,l+1},\\ [A^*\,\eta]_{2\,l+1} &= \frac{1}{2\,\pi\,l}\left(\sqrt{2}\,\eta_1 - \eta_{2\,l+1}\right)\end{aligned}$$

for l in $\mathbb{N}$. The null space of A is

$$\mathcal{N}(A) = \left\{(0, w_1, w_1, w_2, w_2, \ldots) \in \ell^1 : \sum_{l=1}^{\infty} \frac{1}{l}\,w_l = 0\right\}.$$

7. Ill-posedness in the ℓ^1-setting

Looking for solutions in ℓ^1 and considering weak*-to-weak continuous operators A defined only on ℓ^1 makes equation (6.4), surprisingly, ill-posed regardless of ill-posedness or well-posedness of the original equation (6.1). Following the ideas of [Nas87] ill-posedness can be classified as type II. This chapter presents the details of these results. The contents of this chapter were published in [FHV15] under stronger assumptions. Here we content ourselves with the same assumptions as in Theorem 6.4.

Definition of the term *ill-posed* for linear equations in Banach spaces is a delicate issue. On the one hand, we have the well established definition in Hilbert spaces, which says that a linear equation (or its operator) is called ill-posed if the operator's range is not closed or, equivalently, if the Moore–Penrose inverse is unbounded. On the other hand, we have definitions for ill-posedness of nonlinear mappings in Banach spaces like Definition 1.13.

Having the Hilbert space concept in mind, one is tempted to say that linear equations in Banach spaces are ill-posed if corresponding operator ranges are not closed. But this definition is not equivalent to non-existence of bounded generalized inverses, due to the fact, that the operator's null space might be uncomplemented and there are no generalized inverses at all. The articles [Nas87] and [NV76] are a good starting point for the interested reader.

Taking a definition from the nonlinear theory is an alternative. But definitions in the literature differ and in part assume injectivity of A if specialized to linear operators, e. g. [HS98, Definition 1.1].

We are on the horns of a dilemma, which should be solved in future. In the following we show that the range of A is not closed, which by [Nas87, Proposition 2.1] at least implies that there is no bounded inner inverse. The following lemmas will prove useful in subsequent chapters, too. We start with a very useful characterization of $(\ell^\infty)^*$, which is a special case of [Tak02, Theorem 2.14].

Lemma 7.1. *Each element of $(\ell^\infty)^*$ is the sum of an element of ℓ^1 and an element of $c_0^\perp$, that is,*

$$(\ell^\infty)^* = \ell^1 \oplus c_0^\perp.$$

Proof. Let $u \in (\ell^\infty)^*$. Set

$$x_k := \langle u, e^{(k)} \rangle_{(\ell^\infty)^* \times \ell^\infty}.$$

Then $x = (x_k)_{k\in\mathbb{N}} \in \ell^1$ because

$$\sum_{k=1}^{n} |x_k| = \sum_{k=1}^{n} (\operatorname{sgn} x_k)\, \langle u, e^{(k)} \rangle_{(\ell^\infty)^* \times \ell^\infty} = \left\langle u, \sum_{k=1}^{n} (\operatorname{sgn} x_k)\, e^{(k)} \right\rangle_{(\ell^\infty)^* \times \ell^\infty} \leq \|u\|_{(\ell^\infty)^*}.$$

It remains to show $u - x \in c_0^\perp$. Indeed, for each ξ in c_0 we have

$$\begin{aligned}\langle u - x, \xi\rangle_{(\ell^\infty)^* \times \ell^\infty} &= \lim_{n\to\infty} \left\langle u - x, \sum_{k=1}^{n} \xi_k\, e^{(k)} \right\rangle_{(\ell^\infty)^* \times \ell^\infty} \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} \Big(\xi_k \, \langle u, e^{(k)}\rangle_{(\ell^\infty)^* \times \ell^\infty} - \xi_k \, \langle x, e^{(k)}\rangle_{\ell^1 \times \ell^\infty} \Big) \\ &= 0.\end{aligned}$$

□

Lemma 7.2. *If $A : \ell^1 \to Y$ is weak*-to-weak continuous, then $\mathcal{R}(A^{**}) = \mathcal{R}(A)$.*

Proof. From Lemma 7.1 we know that A^{**} maps $\ell^1 \oplus c_0^\perp$ into Y^{**}. On the one hand, for each $x \in \ell^1$ and each $\eta \in Y^*$ we see

$$\begin{aligned}\langle A^{**}\, x, \eta\rangle_{Y^{**} \times Y^*} &= \langle x, A^*\, \eta\rangle_{(\ell^\infty)^* \times \ell^\infty} = \langle x, A^*\, \eta\rangle_{\ell^1 \times \ell^\infty} = \langle A\, x, \eta\rangle_{Y \times Y^*} \\ &= \langle A\, x, \eta\rangle_{Y^{**} \times Y^*},\end{aligned}$$

that is, $A^{**}|_{\ell^1} = A$. On the other hand, for each u in $c_0^\perp$ and each η in Y^* we see

$$\langle A^{**}\, u, \eta\rangle_{Y^{**} \times Y^*} = \langle u, A^*\, \eta\rangle_{(\ell^\infty)^* \times \ell^\infty} = 0$$

because $A^*\, \eta \in \mathcal{R}(A^*) \subseteq c_0$ as a consequence of weak*-to-weak continuity (cf. Lemma 6.3). Thus, $A^{**}|_{c_0^\perp} = 0$ and consequently $\mathcal{R}(A^{**}) = \mathcal{R}(A)$. □

The following theorem was proven in [FHV15, Proposition 1.1] under the assumptions that A is injective and that A has a bounded extension to ℓ^2. Here we drop the injectivity assumption and we replace bounded extensibility by the weaker assumption of weak*-to-weak continuity, which we already used to prove Theorem 6.4.

Theorem 7.3. *If the bounded linear operator $A : \ell^1 \to Y$ is weak*-to-weak continuous, then either the range of A is finite-dimensional or the range is not closed.*

Proof. By Lemma 7.2 we have $\mathcal{R}(A^{**}) \subseteq Y$, which is equivalent to weak compactness of A, see [Meg98, Theorem 3.5.8]. Weak compactness of A is equivalent to weak compactness of A^*, see [Meg98, Theorem 3.5.13]. Thus, if $\mathcal{R}(A^*)$ would be closed, then $\mathcal{R}(A^*)$ would be reflexive, see [Meg98, Proposition 3.5.6]. But Lemma 6.3 states that $\mathcal{R}(A^*)$ is contained in c_0 and c_0 has no infinite-dimensional reflexive subspaces, see, e. g. [GT63, Remark (ii) on page 335]. Therefore $\mathcal{R}(A^*)$ is either finite-dimensional or not closed. To complete the proof we observe that $\mathcal{R}(A^*)$ is finite-dimensional or closed if and only if $\mathcal{R}(A)$ is finite-dimensional or closed, respectively, see [Meg98, Theorem 3.1.21]. □

In [Nas87, Definition 3.1] bounded linear operators with non-closed range were classified into operators of type I and operators of type II. More precisely, the type is assigned to regularizing families for an operator, but it turns out that the type only depends on the operator itself. Equations with operators of type I can be regarded as less ill-posed than equations with operators of type II. We use the following definition, which is an equivalent reformulation of the original definition, see [Nas87, Theorem 4.5].

Definition 7.4. A bounded linear operator $A : X \to Y$ between Banach spaces X and Y with non-closed range is of *type I*, if the range $\mathcal{R}(A)$ contains a closed infinite-dimensional subspace V and the nullspace $\mathcal{N}(A)$ is complemented in the full preimage $A^{-1}V$. Otherwise, A is of *type II*.

Theorem 7.5. *If the bounded linear operator $A : \ell^1 \to Y$ has infintite-dimensional range and is weak*-to-weak continuous, then A is of type II.*

Proof. By Theorem 7.3 the range of A is not closed and therefore Definition 7.4 applies.

Assume V is a closed infinite-dimensional subspace of $\mathcal{R}(A)$. Boundedness of A ensures closedness of $U := A^{-1}V$, that is, U is a Banach space itself. Since the weak topology of U as a Banach space is the same as the weak topology of ℓ^1 restricted to the subspace U, see [Meg98, Proposition 2.5.22], the restriction $A|_U : U \to Y$ inherits weak compactness from A, cf. proof of Theorem 7.3. Weak*-to-weak continuity of A is equivalent to $\mathcal{R}(A^*) \subseteq c_0$, which implies $\mathcal{R}(A|_U^*) \subseteq c_0$.

Repeating the arguments in the proof of Theorem 7.3 for $A|_U$ yields the contradiction that V cannot be closed. Consequently, A has to be of type II. □

8. Convergence rates

8.1. Results in the literature

The aim of this chapter is to derive error estimates for the distance between exact solutions $x^\dagger$ of (6.4) and ℓ^1-regularized solutions, that is, minimizers of (6.5). Such estimates shall bound the error in terms of the noise level δ introduced in (6.2). More specifically we aim at asymptotic estimates of the form

$$\|x_\alpha^\delta - x^\dagger\| = \mathcal{O}(\varphi(\delta)), \quad \text{if } \delta \to 0, \tag{8.1}$$

with an index function φ (cf. Definition 4.1), where the regularization parameter α has to be chosen appropriately depending on the noise level δ and noisy data y^δ. From Theorem 6.4 we know that ℓ^1-regularized solutions converge to norm minimizing solutions. Thus, $x^\dagger$ always denotes such a norm minimizing solution. If there are more than one norm minimizing solution, we aim at estimates

$$\operatorname{dist}(x_\alpha^\delta, S) = \mathcal{O}(\varphi(\delta)), \quad \text{if } \delta \to 0, \tag{8.2}$$

with S denoting the set of all norm minimizing solutions and $\operatorname{dist}(\cdot, S)$ denoting the ℓ^1-distance to this set. If there are more than one minimizer to (6.5), than estimates (8.1) and (8.2) typically hold for all minimizers.

Before we come to our own convergence rates results, we summarize approaches and results from the literature in the present section.

As already noted before, inverse problems related research on ℓ^1-regularization, including convergence rates, started with the influencial paper [DDDM04]. There the operator A is considered as defined on ℓ^2 and convergence rates with respect to the ℓ^2-norm are derived. The authors restrict their attention to injective operators and function spaces Y. Based on estimates with respect to Besov space norms they derive convergence rates.

In [Lor08] convergence rates (8.1) with $\varphi(\delta) = \sqrt{\delta}$ are obtained based on a typical Banach space source condition, which originally was used to obtain rates in terms of so called Bregman distances (see Appendix A). Such Bregman distances are no sensible error measure in the context of ℓ^1-regularization, cf. discussion in [Lor08, Section 4]. Using a Bregman distance related source condition in connection with the ℓ^1-norm might be the reason that the rate obtained this way turned out to be not optimal.

In [Gra09] a source-type condition requiring that the $e^{(k)}$ belong to the range of A^* is used to obtain a rate (8.1) with $\varphi(\delta) = \delta$, but only if the ℓ^q-norm with $q < 1$ is used as penalty in (6.5). The case $q = 1$, which is the one of interest in this thesis, has been added in [GHS08] based on a variational source condition quite similar to the variant we will use below.

Several different convergence rates results for sparsity promoting regularization are derived in [BL09] with variational source conditions as the main tool. Results differ with respect to the penalty functional and with respect to the error measure.

Without adhering to source-type conditions convergence rates for ℓ^q-penalties are proven in [Gra10b], but again only for $q < 1$. We mention this result here because we will add the rates result for $q = 1$ in this thesis.

Discussion of several details and extensions of the results to other penalty functionals can be found in [RR10, GHS11a]. For all results cited so far the authors assume that there is only one norm minimizing solution $x^\dagger$ and that this solution is sparse. In addition, the results rely on or imply some kind of injectivity of A. Either usual injectivity of A or injectivity on the support of the solution $x^\dagger$ (finite basis injectivity, restricted isometry property) is assumed.

In the remaining sections of this chapter we will extend convergence rates results for ℓ^1-regularization in several ways. Our main tool will be variational source conditions of the form

$$\beta \operatorname{dist}(x, S) \leq \|x\|_{\ell^1} - \|S\|_{\ell^1} + \varphi(\|A\,x - A\,S\|_Y) \qquad \text{for all } x \text{ in } \ell^1, \tag{8.3}$$

where β is some positive constant, φ is an index function and S denotes the set of all norm minimizing solutions. Note that all elements in S have the same norm and the same image with respect to A, which justifies the non-standard notation $\|S\|_{\ell^1}$ and $A\,S$ for norm and image of some norm minimizing solution. If there is only one norm minimizing solution, then the variational source condition becomes

$$\beta\,\|x - x^\dagger\|_{\ell^1} \leq \|x\|_{\ell^1} - \|x^\dagger\|_{\ell^1} + \varphi(\|A\,x - A\,x^\dagger\|_Y) \qquad \text{for all } x \text{ in } \ell^1. \tag{8.4}$$

The above variational source conditions are known to imply the corresponding convergence rates (8.1) and (8.2), see [Fle12, HM12]. For the reader's convenience we provide tailor-made proofs in Section 8.3.

8.2. Classical techniques do not work

Before we show how to obtain convergence rates for ℓ^1-regularization, we verify that techniques based on classical source conditions do not work in this context. We consider Banach space source conditions introduced in [BO04] and approximate source conditions introduced for Hilbert spaces in [Hof06] and extended to Banach spaces in [HH09].

In [BO04] it was shown that if an exact solution $x^\dagger$ to (6.4) satisfies

$$(\partial\|\cdot\|_{\ell^1})(x^\dagger) \cap \mathcal{R}(A^*) \neq \emptyset, \tag{8.5}$$

then linear convergence rates in terms of Bregman distances can be obtained. But condition (8.5) implies that $x^\dagger$ is sparse if A is weak*-to-weak continuous, which is a standard assumption to ensure proper behavior of ℓ^1-regularization. Indeed, if $\xi^\dagger$ is a subgradient of the ℓ^1-norm at $x^\dagger$ which belongs to $\mathcal{R}(A^*)$, then $\xi^\dagger_k \to 0$ by Lemma 6.3. Thus, only finitely many components of $\xi^\dagger$ can equal 1 or -1 and therefore $x^\dagger_k \neq 0$ is only possible for finitely many k.

Since we are interested not only in sparse solutions $x^\dagger$ but also in almost sparse solutions, the method of approximate source conditions seems to be a reasonable alternative. For a fixed subgradient $\xi^\dagger$ from $(\partial\|\cdot\|_{\ell^1})(x^\dagger)$ in [Hof06, HH09] the distance function $d_{\xi^\dagger}:[0,\infty)\to[0,\infty)$ defined by

$$d_{\xi^\dagger}(R) := \inf\bigl\{\|\xi^\dagger - A^*\eta\|_{\ell^\infty} : \eta\in Y^*,\ \|\eta\|_{Y^*}\le R\bigr\},\quad R\ge 0, \tag{8.6}$$

has been introduced.

If $d_{\xi^\dagger}(R)\to 0$ for $R\to\infty$, convergence rates for the Bregman distance as error measure can be shown. The rates then depend on the decay of the distance function at infinity. The following proposition states that in our ℓ^1-setting distance functions may only decay to zero if the solution $x^\dagger$ is sparse.

Proposition 8.1. *Let A be weak*-to-weak continuous, let $\xi^\dagger$ be in $(\partial\|\cdot\|_{\ell^1})(x^\dagger)$ and let $d_{\xi^\dagger}$ be as in* (8.6)*. If $d_{\xi^\dagger}(R)\to 0$ for $R\to\infty$, then $x^\dagger$ is sparse. Moreover, if $x^\dagger$ is not sparse, then $d_{\xi^\dagger}(R)\ge 1$ for all R.*

Proof. If $d_{\xi^\dagger}(R)\to 0$, then $\xi^\dagger\in\overline{\mathcal{R}(A^*)}$. By Lemma 6.3 we have $\mathcal{R}(A^*)\subseteq c_0$ and, since c_0 is closed in ℓ^∞, also $\overline{\mathcal{R}(A^*)}\subseteq c_0$. Therefore, only finitely many components of $\xi^\dagger$ can equal 1 or -1, which implies that $x^\dagger_k\neq 0$ is only possible for finitely many k.

If $x^\dagger$ is not sparse, we find a subsequence $(\xi^\dagger_{k_l})_{l\in\mathbb{N}}$ of the components of $\xi^\dagger$ with $|\xi^\dagger_{k_l}|=1$ for all l in $\mathbb{N}$. Since $\mathcal{R}(A^*)\subseteq c_0$, for each η in Y^* we obtain

$$\|\xi^\dagger - A^*\eta\|_{\ell^\infty}\ge\sup_{l\in\mathbb{N}}|\xi^\dagger_{k_l}-[A^*\eta]_{k_l}| = 1,$$

which completes the proof. □

In this section we saw two reasons, why (approximate) source conditions are not a suitable tool in the ℓ^1-setting. First, they are rarely satisfied. Second, if they are satisfied, they only yield rates for the Bregman distance, which carries almost no information in ℓ^1.

8.3. Variational source conditions imply convergence rates

In this section we prove that variational source conditions (8.3) imply rates (8.2). To choose the regularization parameter we consider an *a priori* parameter choice $\alpha=\alpha(\delta)$ specified below and an *a posteriori* parameter choice $\alpha=\alpha(\delta,y^\delta)$ known as discrepancy principle. The later consists in choosing α such that

$$\delta\le\|A\,x^\delta_\alpha - y^\delta\|_Y\le\tau\,\delta \tag{8.7}$$

with $\tau\ge 1$.

In case of the a priori choice we apply techniques from [HM12], but slightly improve the constants in the error estimate. For the discrepancy principle we take the proof from [Fle12] and specialize it to our setting. Both specialized proofs can also be found in [FG17].

To shorten the two proofs we mention two properties of concave index functions φ. Simple calculations show that $t\mapsto\frac{\varphi(t)}{t}$ is decreasing. As a consequence we see that $\varphi(c\,t)\le c\,\varphi(t)$ if $c\ge 1$. Both observations will be used without further notice.

Proposition 8.2. *Let the variational source condition* (8.3) *be satisfied and choose* α *in* (6.5) *such that*

$$c_1 \frac{\delta^p}{\varphi(\delta)} \leq \alpha \leq c_2 \frac{\delta^p}{\varphi(\delta)}$$

for all positive δ *with positive constants* c_1, c_2. *Then*

$$\operatorname{dist}(x_\alpha^\delta, S) \leq \frac{1}{\beta} \left(1 + \frac{1}{c_1} + (1 + 2\,c_2)^{\frac{1}{p-1}}\right) \varphi(\delta)$$

for all positive δ.

Proof. Because x_α^δ is a minimizer of (6.5) we have

$$\begin{aligned}
\|x_\alpha^\delta\|_{\ell^1} - \|x^\dagger\|_{\ell^1} &= \frac{1}{\alpha} \big(T_\alpha^\delta(x_\alpha^\delta) - \alpha\,\|x^\dagger\|_{\ell^1} - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\big)\\
&\leq \frac{1}{\alpha} \big(\|A\,x^\dagger - y^\delta\|_Y^p - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\big)\\
&\leq \frac{1}{\alpha} \big(\delta^p - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\big).
\end{aligned}$$

and thus the variational source condition (8.3) implies

$$\beta \operatorname{dist}(x_\alpha^\delta, S) \leq \frac{1}{\alpha} \big(\delta^p - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\big) + \varphi\big(\|A\,x_\alpha^\delta - A\,x^\dagger\|_Y\big). \tag{8.8}$$

Because $\beta \operatorname{dist}(x_\alpha^\delta, S) \geq 0$, we obtain

$$\|A\,x_\alpha^\delta - y^\delta\|_Y^p \leq \delta^p + \alpha\,\varphi\big(\|A\,x_\alpha^\delta - A\,x^\dagger\|_Y\big).$$

If $\|A\,x_\alpha^\delta - y^\delta\|_Y \leq \delta$, then the triangle inequality, the properties of φ and the parameter choice imply

$$\|A\,x_\alpha^\delta - y^\delta\|_Y^p \leq \delta^p + \alpha\,\varphi(2\,\delta) \leq \delta^p + 2\,\alpha\,\varphi(\delta) \leq (1 + 2\,c_2)\,\delta^p,$$

that is,

$$\|A\,x_\alpha^\delta - y^\delta\|_Y \leq (1 + 2\,c_2)^{\frac{1}{p}}\,\delta \leq (1 + 2\,c_2)^{\frac{1}{p-1}}\,\delta.$$

If, on the other hand, $\|A\,x_\alpha^\delta - y^\delta\|_Y > \delta$, then

$$\begin{aligned}
\|A\,x_\alpha^\delta - y^\delta\|_Y^p &\leq \delta^p + \alpha\,\varphi\big(\|A\,x_\alpha^\delta - y^\delta\|_Y + \delta\big)\\
&= \delta^p + \alpha\,\frac{\varphi\big(\|A\,x_\alpha^\delta - y^\delta\|_Y + \delta\big)}{\|A\,x_\alpha^\delta - y^\delta\|_Y + \delta} \big(\|A\,x_\alpha^\delta - y^\delta\|_Y + \delta\big)\\
&\leq \delta^p + \alpha\,\frac{\varphi(\delta)}{\delta} \big(\|A\,x_\alpha^\delta - y^\delta\|_Y + \delta\big)\\
&\leq \delta^{p-1}\,\|A\,x_\alpha^\delta - y^\delta\|_Y + 2\,\alpha\,\frac{\varphi(\delta)}{\delta}\,\|A\,x_\alpha^\delta - y^\delta\|_Y
\end{aligned}$$

and thus,

$$\|A\,x_\alpha^\delta - y^\delta\|_Y \leq \left(\delta^{p-1} + 2\,\alpha\,\frac{\varphi(\delta)}{\delta}\right)^{\frac{1}{p-1}} \leq (1 + 2\,c_2)^{\frac{1}{p-1}}\,\delta.$$

In both cases (8.8) can be further estimated to obtain

$$\begin{aligned}\beta\,\mathrm{dist}(x_\alpha^\delta, S) &\leq \frac{1}{\alpha}\left(\delta^p - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\right) + \varphi\big(\|A\,x_\alpha^\delta - y^\delta\|_Y + \delta\big)\\ &\leq \frac{\delta^p}{\alpha} + \varphi\left(\left(1 + (1 + 2\,c_2)^{\frac{1}{p-1}}\right)\delta\right)\\ &\leq \frac{\delta^p}{\alpha} + \left(1 + (1 + 2\,c_2)^{\frac{1}{p-1}}\right)\varphi(\delta)\end{aligned}$$

and the lower bound for α leads to

$$\beta\,\mathrm{dist}(x_\alpha^\delta, S) \leq \frac{\varphi(\delta)}{c_1} + \left(1 + (1 + 2\,c_2)^{\frac{1}{p-1}}\right)\varphi(\delta). \qquad \square$$

Note that in the proof we used arguments similar to the ones in [HM12], but made changes in the details leading to a better constant in the obtained error estimate. Corresponding estimates in [HM12, Theorem 1] lead to

$$\mathrm{dist}(x_\alpha^\delta, S) \leq \frac{1}{\beta}\left(1 + 2\,(2 + p)^{\frac{1}{p-1}}\right)\varphi(\delta),$$

which has a greater constant factor than our estimate. Our estimate with the parameter choice from [HM12], that is $c_1 = c_2 = 1$, reads

$$\mathrm{dist}(x_\alpha^\delta, S) \leq \frac{1}{\beta}\left(2 + 3^{\frac{1}{p-1}}\right)\varphi(\delta).$$

Proposition 8.3. *Let the variational source condition* (8.3) *be satisfied and choose α in* (6.5) *according to the discrepancy principle* (8.7)*. Then*

$$\mathrm{dist}(x_\alpha^\delta, S) \leq \frac{1 + \tau}{\beta}\,\varphi(\delta)$$

for all positive δ.

Proof. Because x_α^δ is a minimizer of (6.5) we have

$$\begin{aligned}\|x_\alpha^\delta\|_{\ell^1} - \|x^\dagger\|_{\ell^1} &= \frac{1}{\alpha}\left(T_\alpha^\delta(x_\alpha^\delta) - \alpha\,\|x^\dagger\|_{\ell^1} - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\right)\\ &\leq \frac{1}{\alpha}\left(\|A\,x^\dagger - y^\delta\|_Y^p - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\right)\\ &\leq \frac{1}{\alpha}\left(\delta^p - \|A\,x_\alpha^\delta - y^\delta\|_Y^p\right)\end{aligned}$$

and taking into account the left-hand inequality in (8.7) we obtain

$$\|x_\alpha^\delta\|_{\ell^1} - \|x^\dagger\|_{\ell^1} \leq 0.$$

The variational source condition (8.3) thus implies

$$\beta\,\mathrm{dist}(x_\alpha^\delta, S) \leq \varphi\big(\|A\,x_\alpha^\delta - A\,x^\dagger\|_Y\big) \leq \varphi\big(\|A\,x_\alpha^\delta - y^\delta\|_Y + \delta\big)$$

and the right-hand side in (8.7) yields

$$\beta\,\mathrm{dist}(x_\alpha^\delta, S) \leq \varphi\big((1 + \tau)\,\delta\big) \leq (1 + \tau)\,\varphi(\delta). \qquad \square$$

8.4. Smooth bases

As a first sufficient condition for convergence rates in ℓ^1-regularization we consider bounded linear operators $A : \ell^1 \to Y$ with respect to which the canonical basis $(e^{(k)})_{k\in\mathbb{N}}$ is smooth. We say that a basis is smooth with respect to the operator A if the basis is contained in the range of the adjoint $A^* : \ell^\infty \to Y^*$. This situation was considered in [BL09, Assumption 4.1] and [Gra08, Theorem 5] in connection with sparse exact solutions $x^\dagger$ to (6.4). We use smooth bases to obtain convergence rates also in case of almost sparse solutions and discuss implications of such a smoothness assumption as well as sufficient conditions for it. The core results of this section were published in [BFH13].

Assumption 8.4. For each k in $\mathbb{N}$ there is some f_k in Y^* such that

$$e^{(k)} = A^* f_k$$

holds.

At first we consider the relation of basis smoothness to injectivity of A.

Proposition 8.5. *If A satisfies Assumption 8.4, then A is injective.*

Proof. If $A\,x = 0$, then

$$x_k = \langle e^{(k)}, x\rangle_{\ell^\infty\times\ell^1} = \langle A^* f_k, x\rangle_{\ell^\infty\times\ell^1} = \langle f_k, A\,x\rangle_{Y^*\times Y} = 0$$

for all k. Therefore, $x = 0$. □

The following example shows that the converse of the proposition is not true. Injectivity does not imply smoothness of the basis.

Example 8.6 (bidiagonal operator)**.** We consider the operator introduced in Subsection 6.4.2. There we saw that

$$\xi \in \mathcal{R}(A^*) \qquad \Rightarrow \qquad \sum_{k=1}^{\infty} \xi_k = 0.$$

With $\xi = e^{(k)}$ we immediately see that Assumption 8.4 is violated. Nevertheless, A is injective because $A\,x = 0$ implies $x_1 = x_2 = \ldots$, which is only possible if all x_k are zero. □

In Section 6.1 we discussed operators $\tilde{A}$ on general Banach spaces $\tilde{X}$ and sparsity with respect to a Schauder basis $(\tilde{e}_k)_{k\in\mathbb{N}}$. Assumption 8.4 will be reformulated in that setting by the following proposition. For this purpose we denote by $\tilde{e}_k^*$ the coordinate functional associated with the basis element $\tilde{e}_k$, that is,

$$\left\langle \tilde{e}_k^*, \sum_{l=1}^{\infty} \tilde{x}_l\,\tilde{e}_l \right\rangle_{\tilde{X}^*\times\tilde{X}} = \tilde{x}_l.$$

For Schauder bases the coordinate functionals always are bounded, see [Meg98, Corollary 4.1.16].

Proposition 8.7. *Let $\tilde{X}$, $\tilde{A}$ and $(\tilde{e}_k)_{k\in\mathbb{N}}$ be as in Section 6.1 and denote by $(\tilde{e}_k^*)_{k\in\mathbb{N}}$ the coordinate functionals. Then Assumption 8.4 is satisfied if and only if for each k in $\mathbb{N}$ there is some f_k in Y^* such that*

$$\tilde{e}_k^* = \tilde{A}^*\,f_k.$$

The f_k here and in Assumption 8.4 coincide.

Proof. The adjoint $L^* : \tilde{X}^* \to \ell^\infty$ of the synthesis operator $L : \ell^1 \to \tilde{X}$ defined in (6.3) is given by

$$[L^*\,\tilde{\xi}]_k = \langle \tilde{\xi}, \tilde{e}_k \rangle_{\tilde{X}^* \times \tilde{X}}$$

for all k in $\mathbb{N}$ and $\tilde{\xi}$ in $\tilde{X}^*$. Thus, by $A^* = (\tilde{A}\,L)^* = L^*\,\tilde{A}^*$ the relation $e^{(k)} = A^*\,f_k$ is equivalent to

$$\langle \tilde{A}^*\,f_k, \tilde{e}_l \rangle_{\tilde{X}^* \times \tilde{X}} = \begin{cases} 1, & \text{if } l = k, \\ 0, & \text{else.} \end{cases}$$

In other words, $\tilde{A}^*\,f_k$ coincides with the k-th coordinate functional $\tilde{e}_k^*$. □

This reformulation shows, that the choice of the basis $(\tilde{e}_k)_{k\in\mathbb{N}}$ is the only factor influencing validity of Assumption 8.4 if we assume that $\tilde{A}$ and the Banach spaces $\tilde{X}$ and Y are preset by the application they model.

At the first glance Assumption 8.4 seems to be quite artificial and hard to satisfy. But in [AHR13] it was shown that this is not the case. There $\tilde{X}$ is a Hilbert space with a closed subspace $\tilde{U}$, which gives rise for a Gelfand triple of the form $\tilde{U} \subseteq \tilde{X} \subseteq \tilde{U}^*$. The authors have shown that if $\tilde{A}$ has a bounded and boundedly invertible extension to $\tilde{U}^*$ and if $(\tilde{e}_k)_{k\in\mathbb{N}}$ is a basis in $\tilde{U}$, then Assumption 8.4 is satisfied. An important example here is the Radon transform, but also other classes of ill-posed inverse problems can be handled this way, see [AHR13, Sections 3.1–3.5].

In Example 8.6 we already saw, that Assumption 8.4 may be violated. We provide a less academic example for the violation of this assumption next.

Example 8.8 (simple integration and Haar wavelets)**.** Consider the example introduced in Subsection 6.4.3 again. Since $\tilde{X}$ is a Hilbert space and $(\tilde{e}_k)_{k\in\mathbb{N}}$ is an orthonormal basis, the coordinate functionals $\tilde{e}_k^*$ can be identified with the original basis elements $\tilde{e}_k$. All functions $\tilde{x}$ in the range of $\tilde{A}^*$ are continuous and satisfy $\lim_{t\to 1-0} \tilde{x}(t) = 0$. By Proposition 8.7 Assumption 8.4 implies $\tilde{e}_1 \in \mathcal{R}(\tilde{A}^*)$. But $\tilde{e}_1$ is one on the whole interval $(0, 1)$ and therefore not a member of $\mathcal{R}(\tilde{A}^*)$. Consequently, Assumption 8.4 is not satisfied. □

Now we come the convergence rate result.

Theorem 8.9. *Assume that Assumption 8.4 is true and denote by $x^\dagger$ the uniquely determined solution to* (6.4)*. Then a variational source condition* (8.4) *with $\beta = 1$ and a concave index function φ given by*

$$\varphi(t) = 2 \inf_{n\in\mathbb{N}} \left(\sum_{k=n+1}^{\infty} |x_k^\dagger| + \gamma_n\,t \right), \qquad t \geq 0,$$

is satisfied, where

$$\gamma_n := \sup_{\sigma\in\{-1,0,1\}^n} \left\| \sum_{k=1}^{n} \sigma_k\, f_k \right\|_{Y^*}. \tag{8.9}$$

Proof. Uniqueness of the solution follows from Proposition 8.5.

Fix n in $\mathbb{N}$ and x in ℓ^1 and denote by $P_n : \ell^1 \to \ell^1$ the projection setting all but the first n components of a sequence to zero. Further, let $\xi := \operatorname{sgn} P_n\,(x - x^\dagger)$ be the sequence of signs of $P_n\,(x - x^\dagger)$. By Assumption 8.4 we may write

$$\begin{aligned} \|P_n\,(x - x^\dagger)\|_{\ell^1} &= \left\langle \sum_{k=1}^{n} \xi_k\, e^{(k)}, x - x^\dagger \right\rangle_{\ell^\infty\times\ell^1} = \left\langle \sum_{k=1}^{n} \xi_k\, f_k, A\,(x - x^\dagger) \right\rangle_{Y^*\times Y} \\ &\leq \left\| \sum_{k=1}^{n} \xi_k\, f_k \right\|_{Y^*} \|A\,x - A\,x^\dagger\|_Y \leq \gamma_n\, \|A\,x - A\,x^\dagger\|_Y. \end{aligned}$$

Now

$$\begin{aligned} &\|x - x^\dagger\|_{\ell^1} - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1} \\ &\quad = \|P_n\,(x - x^\dagger)\|_{\ell^1} + \|(I - P_n)\,(x - x^\dagger)\|_{\ell^1} - \|P_n\,x\|_{\ell^1} - \|(I - P_n)\,x\|_{\ell^1} \\ &\qquad + \|P_n\,x^\dagger\|_{\ell^1} + \|(I - P_n)\,x^\dagger\|_{\ell^1} \end{aligned}$$

together with

$$\|(I - P_n)\,(x - x^\dagger)\|_{\ell^1} \leq \|(I - P_n)\,x\|_{\ell^1} + \|(I - P_n)\,x^\dagger\|_{\ell^1}$$

and

$$\|P_n\,x^\dagger\|_{\ell^1} = \|P_n\,(x - x^\dagger - x)\|_{\ell^1} \leq \|P_n\,(x - x^\dagger)\|_{\ell^1} + \|P_n\,x\|_{\ell^1}$$

shows

$$\|x - x^\dagger\|_{\ell^1} - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1} \leq 2\,\|(I - P_n)\,x^\dagger\|_{\ell^1} + 2\,\|P_n\,(x - x^\dagger)\|_{\ell^1}.$$

Combining this estimate with the previous estimate for $\|P_n\,(x - x^\dagger)\|_{\ell^1}$ we obtain

$$\|x - x^\dagger\|_{\ell^1} - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1} \leq 2\,\|(I - P_n)\,x^\dagger\|_{\ell^1} + 2\,\gamma_n\,\|A\,x - A\,x^\dagger\|_Y.$$

Taking the infimum proves the asserted form of φ.

It remains to show that φ is a concave index function. As an infimum of affine functions φ is concave and upper semi-continuous. Concavity implies continuity on $(0, \infty)$ and from $\varphi(0) = 0$, non-negativity and upper semi-continuity we obtain continuity of φ on $[0, \infty)$. Monotonicity of φ follows from $\gamma_n > 0$ for all n. That φ is strictly increasing in a neighborhood of the origin follows from $\varphi(0) = 0$ and $\varphi(t) > 0$ for all positive t, where $\varphi(t) > 0$ is again a consequence of $\gamma_n > 0$ for all n. □

If $x^\dagger$ is sparse and $n^\dagger$ is the highest index at which $x^\dagger$ is not zero, the index function φ in the theorem satisfies

$$\varphi(t) \leq 2\,\gamma_{n^\dagger}\, t$$

for all t. This leads to a linear rate of convergence and coincides with the rates for sparse solutions already obtained in [GHS08].

For non-sparse solutions $x^\dagger$ the index function φ in the theorem is influenced by the decay of the components of $x^\dagger$ and by the growth of the source elements f_k. Obviously, $\sum_{k=n+1}^{\infty}|x_k^\dagger|$ decreases to zero and $(\gamma_n)_{n\in\mathbb{N}}$ grows. Moreover, we can show that the γ_n grow to infinity.

Lemma 8.10. *If $A : \ell^1 \to Y$ is weak*-to-weak continuous, then $A^* : Y^* \to \ell^\infty$ is weak*-to-weak continuous.*

Proof. Let $(\eta_\kappa)_{\kappa\in N}$ be a net in Y^* converging weakly* to zero and remember that by Lemma 7.2 we know $\mathcal{R}(A^{**}) \subseteq Y$. Then for each u in $(\ell^\infty)^*$ we have

$$\langle u, A^*\,\eta_\kappa\rangle_{(\ell^\infty)^*\times\ell^\infty} = \langle A^{**}\,u, \eta_\kappa\rangle_{Y^{**}\times Y^*} = \langle \eta_\kappa, A^{**}\,u\rangle_{Y^*\times Y} \to 0,$$

that is, $A^*\,\eta_\kappa \rightharpoonup 0$. □

Proposition 8.11. *Let A be weak*-to-weak continuous and let Assumption 8.4 be true. For $(\gamma_n)_{n\in\mathbb{N}}$ defined by* (8.9)*, we have $\gamma_n \to \infty$.*

Proof. Set

$$B_n := \left\{\sum_{k=1}^{n} a_k\,f_k : a_k \in [-1,1] \text{ for all } k\right\}$$

for all n. Then, by convexity of the norm in Y^*,

$$\gamma_n = \sup_{\eta\in B_n}\|\eta\|_{Y^*}.$$

By Theorem 7.3 and [Meg98, Theorem 3.1.21] the range of $\mathcal{R}(A^*)$ is not closed and by Lemma 6.3 we have $\mathcal{R}(A^*) \subseteq c_0$. Let ξ be some element from c_0 which does not belong to $\mathcal{R}(A^*)$ and satisfies $\|\xi\|_{\ell^\infty} \le 1$. Set $\xi^{(n)} := P_n\,\xi$ for all n in $\mathbb{N}$ with P_n as in the proof of Theorem 8.9. Then

$$\xi = \sum_{k=1}^{n}\xi_k\,A^*\,f_k = A^*\left(\sum_{k=1}^{n}\xi_k\,f_k\right)$$

and for

$$\eta_n := \sum_{k=1}^{n}\xi_k\,f_k$$

we see $\eta_n \in B_n$.

If $(\eta_n)_{n\in\mathbb{N}}$ would be bounded, then there would be a weakly* convergent subsequence with limit η. Denoting the subsequence again by $(\eta_n)_{n\in\mathbb{N}}$, Lemma 8.10 would imply $A^*\eta_n \rightharpoonup A^*\,\eta$ and from

$$\|A^*\,\eta_n - \xi\|_{\ell^\infty} = \|\xi^{(n)} - \xi\|_{\ell^\infty} = \|(I-P_n)\,\xi\|_{\ell^\infty} \to 0$$

we would obtain the contradiction $\xi = A^*\,\eta$. Thus, $\gamma_n \ge \|\eta_n\|_{Y^*} \to \infty$. □

Example 8.12. We simplify φ in Theorem 8.9 for polynomial decay and growth. Assume that

$$\sum_{k=1}^{\infty} |x_k^\dagger| \leq c_1\, n^{-\mu}$$

with positive constants c_1 and μ and that

$$\gamma_n \leq c_2\, n^{\nu}$$

with positive constants c_2 and ν. Then

$$\varphi(t) \leq c\, t^{\frac{\mu}{\mu+\nu}}$$

for all non-negative t with some positive constant c. □

Example 8.13 (denoising)**.** Consider again the example from Subsection 6.4.1. Assumption 8.4 is obviously satisfied with $f_k = e^{(k)}$ and we have $\gamma_n = n^{\frac{p-1}{p}}$. If

$$\sum_{k=1}^{\infty} |x_k^\dagger| \leq c_1\, n^{-\mu}$$

as in the previous example, then

$$\varphi(t) \leq c\, t^{\frac{\mu}{\mu+1-\frac{1}{p}}}$$

for all non-negative t. Here we see that the stronger the norm in the data space Y the better the obtained convergence rate. □

We close this chapter with a slight improvement of Theorem 8.9.

Remark 8.14. Let $\kappa : \mathbb{N} \to \mathbb{N}$ be a permutation of $\mathbb{N}$ such that $(|x_{\kappa(k)}^\dagger|)_{k\in\mathbb{N}}$ is decreasing. Then, slightly modifying the proofs, one can show that Theorem 8.9 holds with

$$\varphi(t) = 2 \inf_{n\in\mathbb{N}} \left(\sum_{k=n+1}^{\infty} |x_{\kappa(k)}^\dagger| + \gamma_n(x^\dagger)\, t \right), \qquad t \geq 0,$$

and

$$\gamma_n(x^\dagger) := \sup_{\sigma\in\{-1,0,1\}^n} \left\| \sum_{k=1}^{n} \sigma_k\, f_{\kappa(k)} \right\|_{Y^*}.$$

This function φ is bounded above by φ from Theorem 8.9 and, thus, possibly yields a better rate. The drawback of this approach is, that now γ_n depends on $x^\dagger$, because κ depends on $x^\dagger$.

8.5. Non-smooth bases

In Examples 8.6 and 8.8 we saw that the canonical basis $(e^{(k)})_{k\in\mathbb{N}}$ is not always smooth with respect to the bounded linear operator $A:\ell^1\to Y$, that is, Assumption 8.4 might be violated and Theorem 8.9 is not applicable. In the present section we introduce a weaker assumption for proving convergence rates, which will be satisfied by the operators from the above mentioned examples. Moreover, in the next section this assumption will turn out to be satisfied for all injective and weak*-to-weak continuous operators.

The core results of this section have been published in [FH15, FHV15, FHV16].

Throughout this section P_n denotes the mapping sending a sequence to the sequence with the same first n components and zeros else.

Assumption 8.15. For each n in $\mathbb{N}$ and each sequence σ taking values in $\{-1,1\}$ there exists a sequence $(\eta_k^{(n)})_{k\in\mathbb{N}}$ with

(i) $P_n\,A^*\,\eta_k^{(n)}=P_n\,\sigma$ for all k in $\mathbb{N}$,

(ii) $\lim\limits_{k\to\infty}(I-P_n)\,A^*\,\eta_k^{(n)}=0.$

The assumption does not require that all $e^{(k)}$ and thus all $P_n\,\sigma$ belong to $\mathcal{R}(A^*)$, but there have to be certain approximations $A^*\,\eta_k^{(n)}$ of $P_n\,\sigma$ in $\mathcal{R}(A^*)$. The first n components of $A^*\,\eta_k^{(n)}$ have to coincide with $P_n\,\sigma$ and the remaining components have to converge uniformly to $P_n\,\sigma$, where the latter is simply zero in the remaining components. Obviously, Assumption 8.15 is true if Assumption 8.4 is true.

To distinguish the assumptions from the previous and the present section we say that the basis $(e^{(k)})_{k\in\mathbb{N}}$ is non-smooth with respect to the operator A if Assumption 8.4 is not satisfied.

As for smooth bases, only injective A can be handled under Assumption 8.15.

Proposition 8.16. *If A satisfies Assumption 8.15, then A is injective.*

Proof. Let $\sigma=(1,1,\ldots)$ and let $(\eta_k^{(n)})_{k\in\mathbb{N}}$ and $(\eta_k^{(n-1)})_{k\in\mathbb{N}}$ be corresponding sequences as in Assumption 8.15 for n and $n-1$. Then

$$e^{(n)}=P_n\,\sigma-P_{n-1}\,\sigma=P_n\,A^*\,\eta_k^{(n)}-P_{n-1}\,A^*\,\eta_k^{(n-1)}$$

for $k=1,\ldots,n$. If $A\,x=0$, then

$$\begin{aligned}x_n=\langle e^{(n)},x\rangle_{\ell^\infty\times\ell^1}&=\left\langle P_n\,A^*\,\eta^{(n)}-P_{n-1}\,A^*\,\eta_k^{(n-1)},x\right\rangle_{\ell^\infty\times\ell^1}\\&=\left\langle A^*\,(\eta_k^{(n)}-\eta_k^{(n-1)}),x\right\rangle_{\ell^\infty\times\ell^1}\\&\quad-\left\langle (I-P_n)\,A^*\,\eta_k^{(n)}-(I-P_{n-1})\,A^*\,\eta_k^{(n-1)},x\right\rangle_{\ell^\infty\times\ell^1}\\&\leq\left\|(I-P_n)\,A^*\,\eta_k^{(n)}\right\|_{\ell^\infty}\|x\|_{\ell^1}+\left\|(I-P_{n-1})\,A^*\,\eta_k^{(n-1)}\right\|_{\ell^\infty}\|x\|_{\ell^1}\end{aligned}$$

for all k. Both summands in the last line converge to zero if $k\to\infty$. Thus, $x_n=0$. Since n was chosen arbitrarily, we obtain $x=0$. □

Here is the convergence rates result.

Theorem 8.17. *Assume that Assumption 8.15 is true and denote by $x^\dagger$ the uniquely determined solution to* (6.4)*. Then for each β in $(0,1)$ a variational source condition* (8.4) *with a concave index function φ given by*

$$\varphi(t) = 2\,\inf_{n\in\mathbb{N}}\left(\sum_{k=n+1}^{\infty}|x_k^\dagger| + \gamma_n\,t\right), \qquad t\geq 0,$$

is satisfied, where the γ_n are of the following structure.

For fixed n in $\mathbb{N}$ and each σ in $\{-1,1\}^{\mathbb{N}}$ denote by $\bigl(\eta_k^{(n)}(\sigma)\bigr)_{k\in\mathbb{N}}$ a sequence as in Assumption 8.15. Choose $k(\sigma)$ large enough to ensure $\bigl\|(I-P_n)\,A^*\,\eta_{k(\sigma)}^{(n)}\bigr\|_{\ell^\infty}\leq\frac{1-\beta}{1+\beta}$. Then

$$\gamma_n := \sup_{\sigma\in\{-1,1\}^{\mathbb{N}}}\left\|\eta_{k(\sigma)}^{(n)}(\sigma)\right\|_{Y^*}. \tag{8.10}$$

Proof. Fix n in $\mathbb{N}$ and x in ℓ^1 and let σ in $\{-1,1\}^{\mathbb{N}}$ such that

$$\sigma_k = \begin{cases}1, & \text{if } x_k - x_k^\dagger\geq 0 \text{ and } k\leq n\\ -1, & \text{if } x_k - x_k^\dagger < 0 \text{ and } k\leq n.\end{cases}$$

Choose $\eta := \eta_{k(\sigma)}^{(n)}$ as described in the theorem. Then

$$\begin{aligned}\|P_n\,(x-x^\dagger)\|_{\ell^1} &= \langle P_n\,\sigma, x-x^\dagger\rangle_{\ell^\infty\times\ell^1} = \langle P_n\,A^*\,\eta, x-x^\dagger\rangle_{\ell^\infty\times\ell^1}\\ &= \langle P_n\,A^*\,\eta - A^*\,\eta, x-x^\dagger\rangle_{\ell^\infty\times\ell^1} + \langle A^*\,\eta, x-x^\dagger\rangle_{\ell^\infty\times\ell^1}\\ &= -\langle (I-P_n)\,A^*\,\eta, (I-P_n)\,(x-x^\dagger)\rangle_{\ell^\infty\times\ell^1} + \langle \eta, A\,x - A\,x^\dagger\rangle_{Y^*\times Y}\\ &\leq \frac{1-\beta}{1+\beta}\,\|(I-P_n)\,(x-x^\dagger)\|_{\ell^1} + \gamma_n\,\|A\,x-A\,x^\dagger\|_Y\end{aligned}$$

and the triangle inequality yields

$$\|P_n\,(x-x^\dagger)\|_{\ell^1}\leq\frac{1-\beta}{1+\beta}\bigl(\|(I-P_n)\,x\|_{\ell^1} + \|(I-P_n)\,x^\dagger\|_{\ell^1}\bigr) + \gamma_n\,\|A\,x-A\,x^\dagger\|_Y. \tag{8.11}$$

Now

$$\begin{aligned}&\beta\,\|x-x^\dagger\|_{\ell^1} - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1}\\ &\quad = \beta\,\|P_n\,(x-x^\dagger)\|_{\ell^1} + \beta\,\|(I-P_n)\,(x-x^\dagger)\|_{\ell^1} - \|P_n\,x\|_{\ell^1} - \|(I-P_n)\,x\|_{\ell^1}\\ &\qquad + \|P_n\,x^\dagger\|_{\ell^1} + \|(I-P_n)\,x^\dagger\|_{\ell^1}\end{aligned}$$

together with

$$\beta\,\|(I-P_n)\,(x-x^\dagger)\|_{\ell^1}\leq\beta\,\|(I-P_n)\,x\|_{\ell^1} + \beta\,\|(I-P_n)\,x^\dagger\|_{\ell^1}$$

and

$$\|P_n\,x^\dagger\|_{\ell^1} = \|P_n\,(x-x^\dagger-x)\|_{\ell^1}\leq\|P_n\,(x-x^\dagger)\|_{\ell^1} + \|P_n\,x\|_{\ell^1}$$

shows

$$\begin{aligned}&\beta\,\|x-x^\dagger\|_{\ell^1}-\|x\|_{\ell^1}+\|x^\dagger\|_{\ell^1}\\&\qquad\le 2\,\|(I-P_n)\,x^\dagger\|_{\ell^1}+(1+\beta)\,\|P_n\,(x-x^\dagger)\|_{\ell^1}\\&\qquad\quad-(1-\beta)\,\bigl(\|(I-P_n)\,x\|_{\ell^1}+\|(I-P_n)\,x^\dagger\|_{\ell^1}\bigr).\end{aligned}$$

Combining this estimate with the previous estimate (8.11) we obtain

$$\begin{aligned}\beta\,\|x-x^\dagger\|_{\ell^1}-\|x\|_{\ell^1}+\|x^\dagger\|_{\ell^1}&\le 2\,\|(I-P_n)\,x^\dagger\|_{\ell^1}+(1+\beta)\,\gamma_n\,\|A\,x-A\,x^\dagger\|_Y\\&\le 2\,\|(I-P_n)\,x^\dagger\|_{\ell^1}+2\,\gamma_n\,\|A\,x-A\,x^\dagger\|_Y.\end{aligned}$$

Taking the infimum over all n in $\mathbb{N}$ proves the structure of φ.

That all γ_n are indeed finite numbers, is a consequence of the fact that the set $\{P_n\,\sigma : \sigma\in\{-1,1\}^{\mathbb{N}}\}$ is finite for each n. As in the proof of Theorem 8.9 one shows that φ is a concave index function. □

As in the case of smooth bases, for sparse solutions $x^\dagger$ the index function φ in Theorem 8.17 is linear. With few modifications to the proof of Proposition 8.11 we see $\gamma_n\to\infty$ if $n\to\infty$ in the theorem. Also Remark 8.14 carries over from the previous to the present section.

The major difference between Theorems 8.9 and 8.17 is that in the latter the numbers γ_n depend on β. This dependence is closely connected to Assumption 8.4.

Proposition 8.18. *Let A be weak*-to-weak continuous, let Assumption 8.15 be true and let $(\gamma_n(\beta))_{n\in\mathbb{N}}$ for β in $(0,1)$ be as in Theorem 8.17. Then Assumption 8.4 is satisfied if and only if*

$$\sup_{\beta\in(0,1)}\gamma_n(\beta)<\infty\quad\textit{for all } n \textit{ in } \mathbb{N}.$$

More precisely, for each n in $\mathbb{N}$ we have

$$\sup_{\beta\in(0,1)}\gamma_n(\beta)<\infty\qquad\Leftrightarrow\qquad e^{(n)}\in\mathcal{R}(A^*).$$

Proof. The non-trivial part is to show the implication '⇒'. Fix n in $\mathbb{N}$ and assume that $e^{(n)}\notin\mathcal{R}(A^*)$. Let $\sigma=(1,1,\ldots)$ and let $(\eta_k^{(n)})_{k\in\mathbb{N}}$ and $(\eta_k^{(n-1)})_{k\in\mathbb{N}}$ be corresponding sequences as in Assumption 8.15 for n and $n-1$. Then

$$\begin{aligned}&\left\|A^*\left(\eta_k^{(n)}-\eta_k^{(n-1)}\right)-e^{(n)}\right\|_{\ell^\infty}\\&\quad\le\left\|P_n\,A^*\,\eta_k^{(n)}-P_{n-1}\,A^*\,\eta_k^{(n-1)}-e^{(n)}\right\|_{\ell^\infty}\\&\qquad+\left\|(I-P_n)\,A^*\,\eta_k^{(n)}-(I-P_{n-1})\,A^*\,\eta_k^{(n-1)}\right\|_{\ell^\infty}\\&\quad\le\left\|(I-P_n)\,A^*\,\eta_k^{(n)}\right\|_{\ell^\infty}+\left\|(I-P_{n-1})\,A^*\,\eta_k^{(n-1)}\right\|_{\ell^\infty}\end{aligned}$$

for all k. Both summands tend to zero if $k\to\infty$.

If $(\beta_l)_{l\in\mathbb{N}}$ is a sequence in $(0,1)$ with limit one, we find subsequences $(\eta_{k_l}^{(n)})_{l\in\mathbb{N}}$ and $(\eta_{k_l}^{(n-1)})_{l\in\mathbb{N}}$ such that

$$\left\|(I-P_n)\,A^*\,\eta_k^{(n)}\right\|_{\ell^\infty}\leq\beta_l\qquad\text{and}\qquad\left\|(I-P_{n-1})\,A^*\,\eta_k^{(n-1)}\right\|_{\ell^\infty}\leq\beta_l$$

for all l in $\mathbb{N}$. Thus, with

$$\eta_l:=\eta_{k_l}^{(n)}-\eta_{k_l}^{(n-1)}$$

we obtain from the convexity of the Y^*-norm that $\|\eta_l\|_{Y^*}\leq\gamma_n(\beta_l)$.

Now assume that $(\eta_l)_{l\in\mathbb{N}}$ is bounded. Then there is a weakly* convergent subsequence, again denoted by $(\eta_l)_{l\in\mathbb{N}}$, with limit η and weak*-to-weak continuity of A implies $A^*\,\eta_l\to A^*\,\eta$. But we already saw $A^*\,\eta_l\to e^{(n)}$, which yields the contradiction $e^{(n)}=A^*\,\eta$. Thus, $(\eta_l)_{l\in\mathbb{N}}$ is not bounded, resulting in $\gamma_n(\beta_l)\to\infty$. □

Example 8.19 (bidiagonal operator)**.** We consider the example from Subsection 6.4.2 again. In Example 8.6 we saw that Assumption 8.4 is violated. Here we show that Assumption 8.15 is true.

Fix n in $\mathbb{N}$ and σ in $\{-1,1\}^{\mathbb{N}}$. We construct a sequence $(\eta_k)_{k\in\mathbb{N}}$ as in Assumption 8.15, where we omit the superscript (n) used there. Note that $Y=Y^*=\ell^2$ and that $[\eta_k]_l$ denotes th l-th component of the ℓ^2-element η_k. Given some η in ℓ^2, a formula for $A^*\,\eta$ can be found in Subsection 6.4.2.

To ensure property (i) in Assumption 8.15, that is, $P_n\,A^*\,\eta_k=P_n\,\sigma$ for all k, we have to choose

$$[\eta_k]_l:=l\sum_{m=1}^{l}\sigma_m$$

for $l=1,\ldots,n$ and for all k in $\mathbb{N}$. For property *(ii)* we fix some sequence $(\mu_k)_{k\in\mathbb{N}}$ of positive numbers converging to zero and require that $\|(I-P_n)\,A^*\,\eta_k\|_{\ell^\infty}\leq\mu_k$ for all k. The latter is equivalent to

$$l\left(\frac{[\eta_k]_{l-1}}{l-1}-\mu_k\right)\leq[\eta_k]_l\leq l\left(\frac{[\eta_k]_{l-1}}{l-1}+\mu_k\right)$$

for $l>n$. Thus, we have some freedom in choosing $[\eta_k]_l$ for $l>n$ and one easily sees that Assumption 8.15 is true, that is, there is some η_k satisfying the bounds.

We go on and calculate γ_n in Theorem 8.17. Here we only have to consider one fixed k with $\mu_k\leq\frac{1-\beta}{1+\beta}$. On the one hand we can make $\|\eta_k\|_{\ell^2}$ as large as we want be setting $[\eta_k]_l=[\eta_k]_n$ for finitely many l and then decay the $[\eta_k]_l$ to zero such that they belong to ℓ^2. This would lead to arbitrarily large γ_n in the theorem. On the other hand we could choose $[\eta_k]_l$ for $l>n$ in a way which minimizes $\|\eta_k\|_{\ell^2}$, leading to the smallest possible γ_n in the theorem. This is what we are going to do now.

The norm of η_k is the smaller the faster $|[\eta_k]_l|$ decays to zero with respect to l. For $l=1,\ldots,n$ these values are fixed and for $l>n$ they can be decreased at most by $l\,\mu_k$ in each component to stay between the above bounds. Thus, the minimum number of non-zero components with $l>n$ is

$$a_k(n):=\left\lfloor\frac{1}{\mu_k}\left|\sum_{m=1}^{n}\sigma_m\right|\right\rfloor,\quad k\in\mathbb{N}.$$

With

$$s := \operatorname{sgn} \sum_{m=1}^{n} \sigma_m$$

one easily verifies that the norm minimizing η_k satisfying the above bounds is given by

$$[\eta_k]_l = \begin{cases} l \sum\limits_{m=1}^{l} \sigma_m, & \text{if } l \in \{1, \ldots, n\}, \\ l \left(\dfrac{[\eta_k]_{l-1}}{l-1} - s\, \mu_k \right), & \text{if } l \in \{n+1, \ldots, n + a_k(n)\}, \\ 0, & \text{if } l > a_k(n). \end{cases}$$

The norm of this element is maximal with respect to σ if $\sum_{m=1}^{n} \sigma_m$ is maximal. The latter is the case for $\sigma_1 = \ldots = \sigma_n = 1$ and leads to

$$\|\eta_k\|_{\ell^\infty}^2 = \sum_{l=1}^{n} l^4 + \sum_{l=n+1}^{n+\left\lfloor \frac{n}{\mu_k} \right\rfloor} l^2 \left(n - (l-n)\, \mu_k\right)^2 = \sum_{l=1}^{n} l^4 + \sum_{l=1}^{\left\lfloor \frac{n}{\mu_k} \right\rfloor} (n+l)^2 \left(n - l\, \mu_k\right)^2.$$

Both sums are of order n^5 and therefore

$$\gamma_n = \|\eta_k\|_{\ell^2} \leq c\, n^{\frac{5}{2}}$$

with some positive constant c. Closer inspection of the constant c shows that $\beta \to 1$ implies $c \to \infty$ as already predicted by Proposition 8.18. □

Example 8.20 (simple integration and Haar wavelets)**.** Now we come back to the example introduced in Subsection 6.4.3 and show that Assumption 8.15 is satisfied. The derivation is quite elementary but longish and will be provided in detail in Appendix B. Here we only present the results, that is, the elements $\eta_k^{(n)}$ and γ_n.

Fix m in $\mathbb{N}_0$ and σ in $\{-1, 0, 1\}^{\mathbb{N}}$. In contrast to Assumption 8.15 we explicitly allow vanishing components in σ, which will turn out to be a useful feature. For the moment set $n = 2^m$. Define the $L^2(0,1)$-functions $h^{(j)}$ for j in $\mathbb{N}$ by

$$h^{(j)}(t) := \begin{cases} \frac{2^j}{1-2^{-j}}\, t, & \text{if } t \in (0, 2^{-j}), \\ \frac{1}{1-2^{-j}}, & \text{if } t \in (2^{-j}, 1-2^{-j}), \\ \frac{-2^j}{1-2^{-j}}\, (t-1), & \text{if } t \in (1-2^{-j}, 1), \end{cases}$$

and set

$$h_{l,k}^{(j)} := 2^{\frac{l}{2}}\, h^{(j)}(2^l \,\cdot\, - k)$$

for l in $\mathbb{N}_0$ and $k = 0, \ldots, 2^l - 1$. These functions are continuous and piecewise differentiable and their derivatives belong to $L^2(0,1)$.

The sequence $(\eta_j^{(n)})_{j \in \mathbb{N}}$ defined by

$$\eta_j^{(n)} := -2^{-\frac{m}{2}} \sum_{r=0}^{2^m-1} \left(\sigma_1 + \sum_{l=0}^{m-1} \sum_{k=0}^{2^l-1} \sigma_{1+2^l+k}\, \tilde{e}_{1+2^l+k} \left(\frac{r - \frac{1}{2}}{2^m} \right) \right) \left(h_{m,r}^{(j)}\right)'$$

satisfies Assumption 8.15 and the corresponding number γ_n in Theorem 8.17 can be estimated above by

$$\gamma_n \leq \frac{2\sqrt{2}}{\sqrt{2}-1}\, 2^{\frac{j(\beta)}{2}} \left(2^m\right)^{\frac{3}{2}}$$

with

$$j(\beta) := \left\lceil \frac{2}{\ln 2} \ln \frac{1+\beta}{(2-\sqrt{2})\,(1-\beta)} \right\rceil.$$

Now we drop the restriction of n to powers of two and choose m such that we have $2^{m-1} < n \leq 2^m$. Only consider σ in $\{-1,0,1\}^{\mathbb{N}}$ with $\sigma_{n+1} = \ldots = \sigma_{2^m} = 0$, which does not restrict $P_n\,\sigma$. Then $(\eta_j^{(n)})_{j\in\mathbb{N}}$ can be chosen as above and the same bound for γ_n is valid. □

8.6. Convergence rates without source-type assumptions

In the previous two chapters we discussed sufficient conditions for convergence rates in ℓ^1-regularization. Now we show that Assumption 8.15 is always satisfied and, thus, weak*-to-weak continuity and injectivity of the operator A in (6.4) are the only requirements needed to proof convergence rates. Results of this section have been published in [FG17].

Lemma 8.21. *We have*

$$\overline{\mathcal{R}(A^*)} = \mathcal{N}(A^{**})_\perp.$$

Proof. See, e. g., [Meg98, Lemma 3.1.16 and Proposition 1.10.15(c)]. □

Lemma 8.22. *The operator A is weak*-to-weak continuous if and only if*

$$\overline{\mathcal{R}(A^*)} = \mathcal{N}(A)^\perp \cap c_0.$$

Proof. Let A be weak*-to-weak continuous. Then Lemma 8.21 states

$$\overline{\mathcal{R}(A^*)} = \{\xi \in \ell^\infty : \langle u, \xi\rangle_{(\ell^\infty)^*\times\ell^\infty} \text{ for all } u \text{ in } \mathcal{N}(A^{**})\}$$

and from Lemma 7.1 we know that the elements of $\mathcal{N}(A^{**})$ may be written as $x+u$ with x in ℓ^1 and u in $c_0^\perp$. Inspecting the proof of Lemma 7.2 we see

$$\mathcal{N}(A^{**}) = \mathcal{N}(A) \oplus c_0^\perp.$$

Thus,

$$\begin{aligned}\overline{\mathcal{R}(A^*)} &= \{\xi \in \ell^\infty : \langle x+u, \xi\rangle_{(\ell^\infty)^*\times\ell^\infty} \text{ for all } x \text{ in } \mathcal{N}(A) \text{ and all } u \text{ in } c_0^\perp\}\\ &= \{\xi \in \ell^\infty : \langle x, \xi\rangle_{(\ell^\infty)^*\times\ell^\infty} \text{ for all } x \text{ in } \mathcal{N}(A)\}\\ &\quad \cap \{\xi \in \ell^\infty : \langle u, \xi\rangle_{(\ell^\infty)^*\times\ell^\infty} \text{ for all } u \text{ in } c_0^\perp\}\\ &= \mathcal{N}(A)^\perp \cap (c_0^\perp)_\perp.\end{aligned}$$

Noting $(c_0^\perp)_\perp = c_0$, see [Meg98, Proposition 1.10.15(b)], completes the proof's first part.

If $\overline{\mathcal{R}(A^*)} = \mathcal{N}(A)^\perp \cap c_0$, then $\mathcal{R}(A^*) \subseteq c_0$ and Lemma 6.3 yields weak*-to-weak continuity of A. □

Theorem 8.23. *Let A be injective and weak*-to-weak continuous and let ε be positive and n be in $\mathbb{N}$. Then for each ξ in c_0 there exists $\tilde{\xi}$ in $\mathcal{R}(A^*)$ such that*

$$\tilde{\xi}_k = \xi_k \quad \text{for } k \leq n \qquad \text{and} \qquad |\tilde{\xi}_k - \xi_k| \leq \varepsilon \quad \text{for } k > n.$$

Proof. We proof the proposition by induction with respect to n. For ξ in c_0 set

$$\xi^+ := (\xi_1 + \varepsilon, \xi_2, \xi_3, \ldots) \qquad \text{and} \qquad \xi^- := (\xi_1 - \varepsilon, \xi_2, \xi_3, \ldots).$$

By Lemma 8.22 we find $\tilde{\xi}^+$ in $\mathcal{R}(A^*)$ and $\tilde{\xi}^-$ in $\mathcal{R}(A^*)$ with

$$\|\tilde{\xi}^+ - \xi^+\|_{\ell^\infty} \leq \varepsilon \qquad \text{and} \qquad \|\tilde{\xi}^- - \xi^-\|_{\ell^\infty} \leq \varepsilon.$$

Consequently, $\tilde{\xi}_1^+ \geq \xi_1 \geq \tilde{\xi}_1^-$ and $|\tilde{\xi}_k^+ - \xi_k| \leq \varepsilon$ as well as $|\tilde{\xi}_k^- - \xi_k| \leq \varepsilon$ for $k > 1$. Thus we find a convex combination $\tilde{\xi}$ of $\tilde{\xi}^+$ and $\tilde{\xi}^-$ such that $\tilde{\xi}_1 = \xi_1$. This $\tilde{\xi}$ obviously also satisfies $|\tilde{\xi}_k - \xi_k| \leq \varepsilon$ for $k > 1$, which proves the proposition for $n = 1$.

Now let the proposition be true for $n = m$. We prove it for $n = m + 1$. Let ξ in c_0 and set

$$\xi^+ := (\xi_1, \ldots, \xi_m, \xi_{m+1} + \varepsilon, \xi_{m+2}, \xi_{m+3}, \ldots),$$
$$\xi^- := (\xi_1, \ldots, \xi_m, \xi_{m+1} - \varepsilon, \xi_{m+2}, \xi_{m+3}, \ldots).$$

By the induction hypothesis we find $\tilde{\xi}^+$ in $\mathcal{R}(A^*)$ and $\tilde{\xi}^-$ in $\mathcal{R}(A^*)$ with

$$\tilde{\xi}_k^+ = \xi_k = \tilde{\xi}_k^- \quad \text{for } k \leq m$$

and

$$|\tilde{\xi}_k^+ - \xi_k^+| \leq \varepsilon \quad \text{and} \quad |\tilde{\xi}_k^- - \xi_k^-| \leq \varepsilon \quad \text{for } k > m.$$

Consequently, $\tilde{\xi}_{m+1}^+ \geq \xi_{m+1} \geq \tilde{\xi}_{m+1}^-$ and $|\tilde{\xi}_k^+ - \xi_k| \leq \varepsilon$ as well as $|\tilde{\xi}_k^- - \xi_k| \leq \varepsilon$ for $k > m+1$. Thus we find a convex combination $\tilde{\xi}$ of $\tilde{\xi}^+$ and $\tilde{\xi}^-$ such that $\tilde{\xi}_{m+1} = \xi_{m+1}$. This $\tilde{\xi}$ obviously also satisfies $\tilde{\xi}_k = \xi_k$ for $k < m + 1$ and $|\tilde{\xi}_k - \xi_k| \leq \varepsilon$ for $k > m + 1$, which proves the proposition for $n = m + 1$. □

Corollary 8.24. *Let A be injective and weak*-to-weak continuous. Then Assumption 8.15 is satisfied.*

Proof. Fix σ in $\{-1, 1\}^{\mathbb{N}}$ and n in $\mathbb{N}$. With $\xi := P_n\, \sigma$ Theorem 8.23 yields for arbitrarily small ε an element $A^*\, \eta$ ($\tilde{\xi}$ in the proposition) such that $P_n\, A^*\, \eta = P_n\, \sigma$ and $\|(I - P_n)\, A^*\, \eta\|_{\ell^\infty} \leq \varepsilon$. □

Complementing Corollary 8.24 one can show that Assumption 8.15 implies injectivity of A if A is known to be weak*-to-weak continuous.

Proposition 8.25. *If A is weak*-to-weak continuous, the following statements are equivalent:*

(i) Assumption 8.15 is true,

(ii) $e^{(k)} \in \overline{\mathcal{R}(A^*)}$ for all k in $\mathbb{N}$,

(iii) $\overline{\mathcal{R}(A^*)} = c_0$,

(iv) A is injective.

Proof. We show (i)⇒(ii)⇒(iii)⇒(iv)⇒(i).

(i)⇒(ii): Fix n in $\mathbb{N}$ and set $\xi^{(n)} := P_n\,(1,1.\ldots)$ as well as $\xi^{(n-1)} := P_{n-1}\,(1,1.\ldots)$. Let $\bigl(\eta_k^{(n)}\bigr)_{k\in\mathbb{N}}$ and $\bigl(\eta_k^{(n-1)}\bigr)_{k\in\mathbb{N}}$ be as in Assumption 8.15 with $\sigma = (1,1,\ldots)$. Then

$$\begin{aligned}\bigl\|e^{(k)} - A^*\bigl(\eta_k^{(n)} - \eta_k^{(n-1)}\bigr)\bigr\|_{\ell^\infty} &= \bigl\|\xi^{(n)} - \xi^{(n-1)} - A^*\bigl(\eta_k^{(n)} - \eta_k^{(n-1)}\bigr)\bigr\|_{\ell^\infty}\\ &\leq \bigl\|\xi^{(n)} - A^*\eta_k^{(n)}\bigr\|_{\ell^\infty} + \bigl\|\xi^{(n-1)} - A^*\eta_k^{(n-1)}\bigr\|_{\ell^\infty}\\ &= \bigl\|(I-P_n)\,A^*\eta_k^{(n)}\bigr\|_{\ell^\infty} + \bigl\|(I-P_{n-1})\,A^*\eta_k^{(n-1)}\bigr\|_{\ell^\infty}\end{aligned}$$

and both summands in the last line converge to zero if $k\to\infty$.

(ii)⇒(iii): $(e^{(k)})_{k\in\mathbb{N}}$ is a Schauder basis in c_0. Thus, $c_0 \subseteq \overline{\mathcal{R}(A^*)}$. In Lemma 6.3 we find that weak*-to-weak continuity implies $\mathcal{R}(A^*)\subseteq c_0$ and hence also $\overline{\mathcal{R}(A^*)}\subseteq c_0$.

(iii)⇒(iv): By Lemma 8.22 we have $c_0 = \mathcal{N}(A)^\perp \cap c_0$. Thus, $c_0 \subseteq \mathcal{N}(A)^\perp$. If we have some x in ℓ^1 with $A\,x = 0$, then for each u in c_0 we obtain

$$\langle x,u\rangle_{\ell^1\times c_0} = \langle u,x\rangle_{\ell^\infty\times\ell^1} = 0,$$

because $x\in\mathcal{N}(A)$ and $u\in\mathcal{N}(A)^\perp$. This is equivalent to $x=0$.

(iv)⇒(i): See Corollary 8.24. □

With the help of Corollary 8.24 it is easy to obtain convergence rates for ℓ^1-regularization with linear operators: weak*-to-weak continuity is almost always satisfied by construction of A from $\tilde{A}$ (cf. discussion in Section 6.2) and injectivity, or one of the equivalent conditions (ii) or (iii) in Proposition 8.25, is not hard to verify.

8.7. Convergence rates without injectivity-type assumptions

In this chapter we prove convergence rates for ℓ^1-regularization without assuming injectivity of the operator A. The basic idea is to take a suitable variational source condition as sufficient condition for convergence rates and to equivalently reformulate this variational source condition as a source-type condition similar to Assumption 8.15. The reformulated condition can be verified more easily than the original one, as will be shown for several examples at the end of the chapter.

Up to minor improvements, the contents of this chapter have been published in [Fle16].

Note that the results in this section also apply to injective operators and thus should contain the results of previous sections as special cases. Although we use a slightly different form of presentation than before to simplify notation where possible, the careful reader will see the close connections between the non-injective and the injective world. In particular, the careful reader will observe that Assumption 8.15 is not only sufficient for a variational source condition with linear φ, but that Assumption 8.15 holds if and only if variational source conditions with linear φ hold for all $x^\dagger$ in ℓ^1 and all β with $\beta < 1$.

8.7.1. Distance to norm minimizing solutions

We do not assume injectivity of A. Thus, there might by many solutions to (6.4). We denote the set of all solutions by

$$L := \{x \in \ell^1 : A\,x = y^\dagger\}.$$

Even restricting our attention to norm minimizing solutions does not guarantee uniqueness, because the norm of ℓ^1 is not strictly convex. The set of all norm minimizing solutions will be denoted by

$$S := \{x \in L : \|x\|_{\ell^1} \leq \|\tilde{x}\|_{\ell^1} \text{ for all } \tilde{x} \text{ in } L\}.$$

Obviously, all elements in S have the same norm and we denote this value by $\|S\|_{\ell^1}$. In addition we immediately see that S is bounded, closed and convex.

For x in ℓ^1 we denote by

$$\operatorname{dist}(x, S) := \inf_{x^\dagger \in S} \|x - x^\dagger\|_{\ell^1}$$

the distance of x to the set S of norm minimizing solutions.

Proposition 8.26. *Let A be weak*-to-weak continuous. Then for each x in ℓ^1 there is some $x^\dagger$ in S such that* $\operatorname{dist}(x, S) = \|x - x^\dagger\|_{\ell^1}$.

Proof. Set $c := \inf_{x^\dagger \in S} \|x - x^\dagger\|_{\ell^1}$ and let $(x^{(n)})_{n\in\mathbb{N}}$ be a sequence in S such that $\|x - x^{(n)}\|_{\ell^1} \to c$ if $n \to \infty$. This sequence is bounded and therefore contains a subsequence converging weakly* to some $x^\dagger$. Since A is weak*-to-weak continuous, $x^\dagger$ is in L. From the weak* lower semi-continuity of the norm and from the definition of c we immediately derive $\|x - x^\dagger\|_{\ell^1} = c$. Thus, $x^\dagger$ is in S. □

The next proposition states that all norm minimizing solutions lie in the same orthant.

Proposition 8.27. *For each k in $\mathbb{N}$ we have either $x_k^\dagger \geq 0$ for all $x^\dagger$ in S or $x_k^\dagger \leq 0$ for all $x^\dagger$ in S.*

Proof. Assume that there are $x^\dagger$ and $\tilde{x}^\dagger$ in S with $x_k^\dagger < 0$ and $\tilde{x}_k^\dagger > 0$ for some k. Set

$$t := \frac{\tilde{x}_k^\dagger}{\tilde{x}_k^\dagger - x_k^\dagger}.$$

Then $t \in (0,1)$ and the convex combination $t\,x^\dagger + (1-t)\,\tilde{x}^\dagger$ belongs to S. We now have

$$\begin{aligned}\|t\,x^\dagger + (1-t)\,\tilde{x}^\dagger\| &= \sum_{l\neq k} |t\,x_l^\dagger + (1-t)\,\tilde{x}_l^\dagger| \leq t \sum_{l\neq k} |x_l^\dagger| + (1-t) \sum_{l\neq k} |\tilde{x}_l^\dagger| \\ &= \|S\|_{\ell^1} - \big(t\,|x_k^\dagger| + (1-t)\,|\tilde{x}_k^\dagger|\big) < \|S\|_{\ell^1},\end{aligned}$$

which is not possible for an element in S. □

Justified by the proposition we define a sequence $\sigma^S = (\sigma_k^S)_{k\in\mathbb{N}}$ by

$$\sigma_k^S := \begin{cases} 1, & \text{if there are } x^\dagger \text{ in } S \text{ with } x_k^\dagger > 0,\\ -1, & \text{if there are } x^\dagger \text{ in } S \text{ with } x_k^\dagger < 0,\\ 0, & \text{if } x_k^\dagger = 0 \text{ for all } x^\dagger \text{ in } S. \end{cases}$$

Further we introduce the set

$$\mathbb{1}^S := \bigl\{(\sigma_k)_{k\in\mathbb{N}} : \sigma_k \in \{-1,0,1\} \text{ for all } k \text{ and } \sigma_k = 0 \text{ if } \sigma_k^S = 0\bigr\}$$

and for each σ in $\mathbb{1}^S$ subsets $S(\sigma)$ of S by

$$\begin{aligned} S(\sigma) := \{x^\dagger \in S : {} & \text{there is some } \xi \in N_S(x^\dagger) \text{ with}\\ & \xi_k = \sigma_k \text{ if } \sigma_k \neq 0, \quad \xi_k \in (-1,1) \text{ if } \sigma_k = 0,\, \sigma_k^S \neq 0\}. \end{aligned}$$

Here, $N_{x^\dagger}(S)$ denotes the normal cone of S at $x^\dagger$. We can regard $S(\sigma)$ as the face of S visible from direction σ.

Lemma 8.28. *Let A be weak*-to-weak continuous. If σ in $\mathbb{1}^S$ has only finitely many non-zero components, we have $S(\sigma) \neq \emptyset$.*

Proof. Setting $\xi := \sigma$ we show that there is some $x^\dagger$ in S with $\xi \in N_S(x^\dagger)$, that is, $x^\dagger$ maximizes $\langle \xi, x\rangle_{\ell^\infty\times\ell^1}$ over all x in S. Let $(x^{(n)})_{n\in\mathbb{N}}$ be a sequence in S with

$$\langle \xi, x^{(n)}\rangle_{\ell^\infty\times\ell^1} \to c := \sup_{x\in S} \langle \xi, x\rangle_{\ell^\infty\times\ell^1}.$$

This sequence is bounded and thus contains a subsequence converging weakly* to some $x^\dagger$ in ℓ^1. The weak*-to-weak continuity of A guarantees $x^\dagger \in L$ and the weak* lower semi-continuity of the ℓ^1-norm yields $x^\dagger \in S$. Denoting the subsequence again by $(x^{(n)})_{n\in\mathbb{N}}$ and noting that $\xi \in c_0$ we further obtain

$$c = \lim_{n\to\infty} \langle \xi, x^{(n)}\rangle_{\ell^\infty\times\ell^1} = \lim_{n\to\infty} \langle x^{(n)}, \xi\rangle_{\ell^1\times c_0} = \langle x^\dagger, \xi\rangle_{\ell^1\times c_0} = \langle \xi, x^\dagger\rangle_{\ell^\infty\times\ell^1}. \tag{8.12}$$

Thus, $x^\dagger$ indeed maximizes $\langle \xi, \cdot\,\rangle_{\ell^\infty\times\ell^1}$ over S. □

Now we restrict our attention to subsets of ℓ^1 on which $\operatorname{dist}(x,S)$ is almost affine. For σ in $\mathbb{1}^S$ and $x^\dagger$ in S we define

$$\begin{aligned} M_{x^\dagger}(\sigma) := \{x \in \ell^1 : {} & x_k \geq x_k^\dagger \text{ if } \sigma_k = 1,\\ & x_k \leq x_k^\dagger \text{ if } \sigma_k = -1,\\ & x_k = x_k^\dagger \text{ if } \sigma_k = 0,\, \sigma_k^S \neq 0\}. \end{aligned}$$

The sets $M_{x^\dagger}(\sigma)$ are obviously closed and convex and we always have $x^\dagger \in M_{x^\dagger}(\sigma)$.

Proposition 8.29. *Let $\sigma \in \mathbb{1}^S$ and let $x^\dagger \in S(\sigma)$. Then*

$$\operatorname{dist}(x,S) = \langle \sigma, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} + \sum_{k:\,\sigma_k^S=0} |x_k| \qquad \textit{for all } x \textit{ in } M_{x^\dagger}(\sigma).$$

Proof. As a standard result of convex analysis we have $\operatorname{dist}(x,S) = \|x - x^\dagger\|_{\ell^1}$ if and only if there is some ξ in the normal cone $N_S(x^\dagger)$ such that $\xi \in -\partial\|x - \cdot\|_{\ell^1}(x^\dagger)$. On the one hand we have

$$-\partial\|x - \cdot\|_{\ell^1}(x^\dagger) = \partial\|\cdot\|_{\ell^1}(x - x^\dagger) = \{\tilde{\xi} \in \ell^\infty : \tilde{\xi}_k = 1 \text{ if } x_k > x_k^\dagger, \\ \tilde{\xi}_k = -1 \text{ if } x_k < x_k^\dagger, \\ \tilde{\xi}_k \in [-1,1] \text{ if } x_k = x_k^\dagger\}.$$

On the other hand, $x^\dagger \in S(\sigma)$ and $x \in M_{x^\dagger}(\sigma)$ imply that there is some ξ in $N_S(x^\dagger)$ such that

$$\xi_k \begin{cases} = 1, & \text{if } x_k > x_k^\dagger, \\ = -1, & \text{if } x_k < x_k^\dagger, \\ \in [-1,1], & \text{if } x_k = x_k^\dagger \end{cases} \quad \text{for all } k \text{ with } \sigma_k^S \neq 0.$$

If we now define $\tilde{\xi}$ by

$$\tilde{\xi}_k := \begin{cases} \xi_k, & \text{if } \sigma_k^S \neq 0, \\ 1, & \text{if } \sigma_k^S = 0,\ x_k \geq 0, \\ -1, & \text{if } \sigma_k^S = 0,\ x_k < 0, \end{cases}$$

we immediately see, that $\tilde{\xi} \in -\partial\|x - \cdot\|_{\ell^1}(x^\dagger)$ (remember $x_k^\dagger = 0$ if $\sigma_k^S = 0$). From $\xi \in N_S(x^\dagger)$ we have

$$\langle \xi, \tilde{x}^\dagger - x^\dagger \rangle_{\ell^\infty \times \ell^1} \leq 0 \quad \text{for all } \tilde{x}^\dagger \in S,$$

which together with

$$\langle \tilde{\xi}, \tilde{x}^\dagger - x^\dagger \rangle_{\ell^\infty \times \ell^1} = \sum_{k\colon \sigma_k^S \neq 0} \tilde{\xi}_k\,(\tilde{x}_k^\dagger - x_k^\dagger) = \sum_{k\colon \sigma_k^S \neq 0} \xi_k\,(\tilde{x}_k^\dagger - x_k^\dagger) = \langle \xi, \tilde{x}^\dagger - x^\dagger \rangle_{\ell^\infty \times \ell^1}$$

yields that $\tilde{\xi}$ is in $N_S(x^\dagger)$, too. This proves $\operatorname{dist}(x,S) = \|x - x^\dagger\|_{\ell^1}$.

As the second step we observe that $x \in M_{x^\dagger}(\sigma)$ yields

$$|x_k - x_k^\dagger| = \sigma_k\,(x_k - x_k^\dagger) \quad \text{if } \sigma_k^S \neq 0.$$

Thus,

$$\|x - x^\dagger\|_{\ell^1} = \sum_{k\colon \sigma_k^S \neq 0} |x_k - x_k^\dagger| + \sum_{k\colon \sigma_k^S = 0} |x_k| = \langle \sigma, x - x^\dagger \rangle_{\ell^\infty \times \ell^1} + \sum_{k\colon \sigma_k^S = 0} |x_k|.$$

□

Corollary 8.30. *For each σ in $\mathbb{1}^S$ and each $x^\dagger$ in $S(\sigma)$ we have*

$$S \cap M_{x^\dagger}(\sigma) = \{x^\dagger\}.$$

Proof. Assume that there is a second solution $\tilde{x}^\dagger$ in $S \cap M_{x^\dagger}(\sigma)$. Then from Proposition 8.29 (and even more easily from its proof) we obtain

$$0 = \operatorname{dist}(\tilde{x}^\dagger, S) = \|\tilde{x}^\dagger - x^\dagger\|_{\ell^1}.$$

Thus, $\tilde{x}^\dagger = x^\dagger$. □

We close this subsection with the following important observation.

Proposition 8.31. *The sets $M_{x^\dagger}(\sigma)$ cover the whole space ℓ^1, that is,*

$$\ell^1 = \bigcup_{\sigma\in\mathbb{1}^S} \bigcup_{x^\dagger\in S(\sigma)} M_{x^\dagger}(\sigma). \tag{8.13}$$

Proof. For fixed x in ℓ^1 let $x^\dagger$ be a minimizer of $\|x - \cdot\,\|_{\ell^1}$ over S. Then there is some ξ in the normal cone $N_S(x^\dagger)$ such that $\xi \in -\partial\|x - \cdot\,\|_{\ell^1}(x^\dagger)$. Thus, we know

$$\xi_k = 1 \text{ if } x_k > x_k^\dagger, \quad \xi_k = -1 \text{ if } x_k < x_k^\dagger, \quad \xi_k \in [-1,1] \text{ if } x_k = x_k^\dagger.$$

If we now define σ by

$$\sigma_k := \begin{cases} 1, & \text{if } \xi_k = 1,\ \sigma_k^S \neq 0,\\ -1, & \text{if } \xi_k = -1,\ \sigma_k^S \neq 0,\\ 0, & \text{if } \xi_k \in (-1,1) \text{ or } \sigma_k^S = 0,\end{cases}$$

then $\sigma \in \mathbb{1}^S$, $x^\dagger \in S(\sigma)$ and $x \in M_{x^\dagger}(\sigma)$. □

8.7.2. Sparse solutions

Having finished the study of the distance $\operatorname{dist}(x,S)$ between an element x in ℓ^1 and the set S of norm minimizing solutions we now want to establish a variational source condition (8.3) with a linear index function $\varphi(t) = \gamma\, t$, $\gamma > 0$. It suffices to consider β in $(0,1]$, because a variational source condition with $\beta > 1$ always implies a variational source condition with $\beta \leq 1$.

At first we split the variational source condition into 'smaller' ones. Here and in the sequel we use the notation introduced in Subsection 8.7.1.

Lemma 8.32. *The variational source condition* (8.3) *on ℓ^1 is satisfied if and only if for each σ in $\mathbb{1}^S$ and each $x^\dagger$ in $S(\sigma)$ we have*

$$\beta\,\langle\sigma, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} + \beta \sum_{k:\,\sigma_k^S=0} |x_k| \leq \|x\|_{\ell^1} - \|x^\dagger\|_{\ell^1} + \gamma\,\|A\,x - A\,x^\dagger\|_Y \tag{8.14}$$

for all x in $M_{x^\dagger}(\sigma)$.

Proof. This is a direct consequence of Propositions 8.29 and 8.31. □

Lemma 8.33. *For σ in $\mathbb{1}^S$ and $x^\dagger$ in $S(\sigma)$ the variational source condition* (8.14) *on $M_{x^\dagger}(\sigma)$ is satisfied if and only if there is some η in Y^* with $\|\eta\|_{Y^*} \leq \frac{\gamma}{1+\beta}$ such that*

$$\begin{cases} [A^*\,\eta]_k \in [-\mu,\mu], & \text{if } \sigma_k^S = 0,\\ \sigma_k\,[A^*\,\eta]_k \leq \mu, & \text{if } \sigma_k^S \neq 0,\ x_k^\dagger = 0,\ \sigma_k \neq 0,\\ \sigma_k^S\,[A^*\,\eta]_k \leq \mu, & \text{if } x_k^\dagger \neq 0,\ \sigma_k = \sigma_k^S,\\ \sigma_k^S\,[A^*\,\eta]_k \geq 1, & \text{if } x_k^\dagger \neq 0,\ \sigma_k = -\sigma_k^S\end{cases}$$

for all k, where $\mu := \frac{1-\beta}{1+\beta} \in [0,1)$.

Proof. We rewrite (8.14) as

$$\|x^\dagger\|_{\ell^1} \leq -\beta\,\langle\sigma, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} + (1-\beta)\sum_{k:\,\sigma_k^S=0}|x_k| + \sum_{k:\,\sigma_k^S\neq 0}|x_k| + \gamma\,\|A\,x - A\,x^\dagger\|_Y$$

and, taking into account that $x_k^\dagger = 0$ if $\sigma_k^S = 0$, see that $x^\dagger$ is a minimizer of the convex functional on the right-hand side with respect to x in $M_{x^\dagger}(\sigma)$. Thus, there is some ξ in the normal cone $N_{M_{x^\dagger}(\sigma)}(x^\dagger)$ such that $-\xi$ belongs to the subdifferential of the functional at $x^\dagger$. This subdifferential is the sum of the subdifferentials for each summand. We have

$$\begin{aligned} N_{M_{x^\dagger}(\sigma)}(x^\dagger) = \{\xi\in\ell^\infty : \;& \xi_k = 0 \text{ if } \sigma_k^S = 0,\\ & \xi_k \leq 0 \text{ if } \sigma_k = 1,\\ & \xi_k \geq 0 \text{ if } \sigma_k = -1\},\end{aligned}$$

$$\partial(-\beta\,\langle\sigma, \cdot - x^\dagger\rangle_{\ell^\infty\times\ell^1})(x^\dagger) = -\beta\,\sigma,$$

$$\begin{aligned}\partial\left(x\mapsto(1-\beta)\sum_{k:\,\sigma_k^S=0}|x_k|\right)(x^\dagger) = \{\tilde{\xi}\in\ell^\infty : \;& \tilde{\xi}_k\in[-(1-\beta),1-\beta] \text{ if } \sigma_k^S = 0,\\ & \tilde{\xi}_k = 0 \text{ if } \sigma_k^S \neq 0\},\end{aligned}$$

$$\begin{aligned}\partial\left(x\mapsto\sum_{k:\,\sigma_k^S\neq 0}|x_k|\right)(x^\dagger) = \{\tilde{\xi}\in\ell^\infty : \;& \tilde{\xi}_k = 0 \text{ if } \sigma_k^S = 0,\\ & \tilde{\xi}_k = 1 \text{ if } x_k^\dagger > 0\\ & \tilde{\xi}_k = -1 \text{ if } x_k^\dagger < 0\\ & \tilde{\xi}_k\in[-1,1] \text{ if } \sigma_k^S\neq 0,\, x_k^\dagger = 0\},\end{aligned}$$

and

$$\partial(\gamma\,\|A\,\cdot\, - A\,x^\dagger\|_Y)(x^\dagger) = \{A^*\,\eta : \eta\in Y^*,\, \|\eta\|_{Y^*}\leq\gamma\}.$$

From these equations we see that there is some η in Y^* with $\|\eta\|_{Y^*}\leq\gamma$ such that

$$\begin{cases} -[A^*\,\eta]_k\in[-(1-\beta),1-\beta], & \text{if } \sigma_k^S = 0,\\ -\sigma_k\,[A^*\,\eta]_k \leq 1-\beta, & \text{if } \sigma_k^S\neq 0,\, x_k^\dagger = 0,\, \sigma_k\neq 0,\\ -\sigma_k^S\,[A^*\,\eta]_k\leq 1-\beta, & \text{if } x_k^\dagger\neq 0,\, \sigma_k = \sigma_k^S,\\ -\sigma_k^S\,[A^*\,\eta]_k\geq 1+\beta, & \text{if } x_k^\dagger\neq 0,\, \sigma_k = -\sigma_k^S.\end{cases}$$

Replacing η by $-(1+\beta)\,\eta$ completes the proof. □

Theorem 8.34. *The variational source condition* (8.3) *on ℓ^1 with $\varphi(t) = \gamma\,t$, $t>0$, is satisfied if and only if for each σ in $\mathbb{1}^S$ and each $x^\dagger$ in $S(\sigma)$ there is some η in Y^* with $\|\eta\|_{Y^*}\leq\frac{\gamma}{1+\beta}$ such that*

$$\begin{cases} [A^*\,\eta]_k\in[-\mu,\mu], & \textit{if } \sigma_k^S = 0,\\ \sigma_k\,[A^*\,\eta]_k\leq\mu, & \textit{if } \sigma_k^S\neq 0,\, x_k^\dagger = 0,\, \sigma_k\neq 0,\\ \sigma_k^S\,[A^*\,\eta]_k\leq\mu, & \textit{if } x_k^\dagger\neq 0,\, \sigma_k = \sigma_k^S,\\ \sigma_k^S\,[A^*\,\eta]_k\geq 1, & \textit{if } x_k^\dagger\neq 0,\, \sigma_k = -\sigma_k^S\end{cases} \tag{8.15}$$

for all k, where $\mu := \frac{1-\beta}{1+\beta} \in [0,1)$.

Proof. This is a direct consequence of Lemma 8.32 and Lemma 8.33. □

Remark 8.35. Let A be weak*-to-weak continuous, that is, by Lemma 6.3, $\mathcal{R}(A^*) \subseteq c_0$. Then Theorem 8.34 implies that a variational source condition (8.3) can only be satisfied if all solutions in S are sparse. To see this choose $\sigma = -\sigma^S$. Then $|[A^*\,\eta]_k| \geq 1$ on the support of $x^\dagger$, which is only possible if the support is finite.

Remark 8.36. Obviously, condition (8.15) is weaker than Assumption 8.4 (smooth basis) because each finitely supported element in $\mathcal{R}(A^*)$ is a linear combination of the standard unit sequences $e^{(k)}$. In Proposition 8.5 we have shown that Assumption 8.4 implies injectivity of A whereas the new characterization (8.15) of a variational source condition does not imply injectivity (cf. Subsection 8.7.5).

We close this section with three remarks which reduce the set of elements σ and $x^\dagger$ for which condition (8.15) has to be verified in order to obtain convergence rates.

Remark 8.37. For fixed σ in $\mathbb{1}^S$ condition (8.15) is satisfied for all $x^\dagger$ in $S(\sigma)$ if and only if it is satisfied for all $x^\dagger$ in $S(\sigma)$ having maximal support. Here we say that some $x^\dagger$ from $S(\sigma)$ has maximal support if there is no $\tilde{x}^\dagger$ in $S(\sigma)$ with $\{k \in \mathbb{N} : \tilde{x}^\dagger_k \neq 0\} \supsetneqq \{k \in \mathbb{N} : x^\dagger_k \neq 0\}$.

Remark 8.38. Let $\sigma \in \mathbb{1}^S$. If $\sigma_k = \sigma^S_k$ for all k with $\sigma_k \neq 0$ and with $x^\dagger_k \neq 0$ for at least one $x^\dagger$ in $S(\sigma)$, then condition (8.15) is satisfied with $\eta = 0$.

Remark 8.39. Let $\sigma \in \mathbb{1}^S$ and $\tilde{\sigma} \in \mathbb{1}^S$ such that $\tilde{\sigma}$ has smaller support than σ, that is, $\sigma_k \neq 0$ whenever $\tilde{\sigma}_k \neq 0$. Further, let $x^\dagger$ be in $S(\sigma)$ and also in $S(\tilde{\sigma})$. Then condition (8.15) is satisfied for $\tilde{\sigma}$ if it is satisfied for σ.

8.7.3. Sparse unique norm minimizing solution

We consider the case that the set of norm minimizing solutions contains only one element, that is,

$$S = \{x^\dagger\}.$$

Note that this does not necessarily imply injectivity of A. The variational source condition (8.3) now reads

$$\beta\,\|x - x^\dagger\|_{\ell^1} \leq \|x\|_{\ell^1} - \|x^\dagger\|_{\ell^1} + \gamma\,\|A\,x - A\,x^\dagger\|_Y \quad \text{for all } x \text{ in } \ell^1 \tag{8.16}$$

and Theorem 8.34 can be refined as follows.

Theorem 8.40. *Assume $S = \{x^\dagger\}$. Then the variational source condition* (8.16) *on ℓ^1 is satisfied if and only if for each σ in $\mathbb{1}^S$ there is some η in Y^* with $\|\eta\|_{Y^*} \leq \frac{\gamma}{1+\beta}$ such that*

$$\begin{cases} [A^*\,\eta]_k \in [-\mu, \mu], & \textit{if } x^\dagger_k = 0, \\ \sigma^S_k\,[A^*\,\eta]_k = \mu, & \textit{if } x^\dagger_k \neq 0,\ \sigma_k = \sigma^S_k, \\ \sigma^S_k\,[A^*\,\eta]_k = 1, & \textit{if } x^\dagger_k \neq 0,\ \sigma_k = -\sigma^S_k \end{cases}$$

for all k, where $\mu := \frac{1-\beta}{1+\beta} \in [0,1)$.

Proof. We apply Theorem 8.34 to the case $S = \{x^\dagger\}$. Note that σ belongs to $\mathbb{1}^S$ if and only if its support coincides with the support of $x^\dagger$, and that σ_k^S provides the sign of $x_k^\dagger$ for each k. Further, the normal cone in the definition of $S(\sigma)$ is $N_S(x^\dagger) = l^\infty$, which allows to choose $\xi = \sigma$ in that definition. We immediately obtain $S(\sigma) = \{x^\dagger\}$ for each σ in $\mathbb{1}^S$ and therefore Theorem 8.34 states the the variational source condition (8.16) holds if and only if for each σ in $\mathbb{1}^S$ there is some η with $\|\eta\|_{Y^*} \leq \frac{\gamma}{1-\beta}$ such that

$$\begin{cases} [A^*\,\eta]_k \in [-\mu,\mu], & \text{if } x_k^\dagger = 0,\\ \sigma_k^S\,[A^*\,\eta]_k \leq \mu, & \text{if } x_k^\dagger \neq 0,\ \sigma_k = \sigma_k^S,\\ \sigma_k^S\,[A^*\,\eta]_k \geq 1, & \text{if } x_k^\dagger \neq 0,\ \sigma_k = -\sigma_k^S \end{cases} \tag{8.17}$$

for each k.

Now fix σ in $\mathbb{1}^S$ and let $k_1, k_2, \ldots$ be an enumeration (finite or infinite) of all indices k satisfying $\sigma_k \neq 0$. Note that $x_{k_n}^\dagger \neq 0$ for all n. We prove the theorem by induction over n.

Let $\bar\sigma$ in $\mathbb{1}^S$ satisfy $\bar\sigma_{k_1} = \sigma_{k_1}^S$ and let $\tilde\sigma$ be the same except for $\tilde\sigma_{k_1} = -\sigma_{k_1}^S$. Then there are $\bar\eta$ and $\tilde\eta$ such that (8.17) holds with σ replaced by $\bar\sigma$ and $\tilde\sigma$, respectively. At index k_1 we have $\sigma_{k_1}^S[A^*\,\bar\eta]_{k_1} \leq \mu$ and $\sigma_{k_1}^S[A^*\,\tilde\eta]_{k_1} \geq 1$. Thus, there exists a convex combination $\eta^{(1)}(\bar\sigma)$ of $\bar\eta$ and $\tilde\eta$ which satisfies $\sigma_{k_1}^S[A^*\,\eta^{(1)}(\bar\sigma)]_{k_1} = \mu$ (if $\sigma_{k_1} = \sigma_{k_1}^S$) or $\sigma_{k_1}^S[A^*\,\eta^{(1)(\bar\sigma)}]_{k_1} = 1$ (if $\sigma_{k_1} = -\sigma_{k_1}^S$). In addition, such an element $\eta^{(1)}(\bar\sigma)$ satisfies (8.17) with σ replaced by $\bar\sigma$ for all other indices k not equalling k_1.

Now let $\bar\sigma$ in $\mathbb{1}^S$ satisfy $\bar\sigma_{k_l} = \sigma_{k_l}$ for $l = 1, \ldots, n-1$ and $\bar\sigma_{k_n} = \sigma_{k_n}^S$. Further, let $\tilde\sigma$ be the same except for $\tilde\sigma_{k_n} = -\sigma_{k_n}^S$. Assume that there is $\eta^{(n-1)}(\bar\sigma)$ such that (8.17) holds for all k and such that for $k_1, \ldots, k_{n-1}$ it holds with equality signs. The existence of such an $\eta^{(n-1)}(\bar\sigma)$ has been shown above for $n = 2$. Again there are $\bar\eta$ and $\tilde\eta$ such that (8.17) holds with σ replaced by $\bar\sigma$ and $\tilde\sigma$, respectively. At index k_n we have $\sigma_{k_n}^S[A^*\,\bar\eta]_{k_n} \leq \mu$ and $\sigma_{k_n}^S[A^*\,\tilde\eta]_{k_n} \geq 1$. Thus, there exists a convex combination $\eta^{(n)}(\bar\sigma)$ of $\bar\eta$ and $\tilde\eta$ which satisfies $\sigma_{k_n}^S[A^*\,\eta^{(n)}(\bar\sigma)]_{k_n} = \mu$ (if $\sigma_{k_n} = \sigma_{k_n}^S$) or $\sigma_{k_n}^S[A^*\,\eta^{(n)(\bar\sigma)}]_{k_n} = 1$ (if $\sigma_{k_n} = -\sigma_{k_n}^S$). In addition, such an element $\eta^{(n)}(\bar\sigma)$ satisfies (8.17) with σ replaced by $\bar\sigma$ for all other indices k not equaling k_n.

So far we have shown that for each σ in $\mathbb{1}^S$ and each n we can construct $\eta^{(n)}(\sigma)$ which satisfies (8.17), where we can replace inequality by equality signs at indices $k_1, \ldots, k_n$. Consequently we find η such that equality holds at all indices k at which $\sigma_k \neq 0$. □

Remark 8.41. Analogously to Remark 8.39 we can replace $\mathbb{1}^S$ in Theorem 8.40 by the set of all σ which satisfy $\sigma_k = \pm 1$ if $\sigma_k^S \neq 0$ and $\sigma_k = 0$ else.

Corollary 8.42. *Assume $S = \{x^\dagger\}$. Then the variational source condition* (8.16) *on ℓ^1 is satisfied if and only if for each σ in $\mathbb{1}^S$ with $\sigma_k \neq 0$ if $\sigma_k^S \neq 0$ there is some η in Y^* with $\|\eta\|_{Y^*} \leq \frac{\gamma}{1+\beta}$ such that*

$$\begin{cases} [A^*\,\eta]_k = \sigma_k^S, & \text{if } x_k^\dagger \neq 0,\ \sigma_k = -\sigma_k^S,\\ [A^*\,\eta]_k \in [-\mu,\mu], & \text{if } x_k^\dagger = 0 \ \text{ or } \ x_k^\dagger \neq 0,\ \sigma_k = \sigma_k^S \end{cases} \tag{8.18}$$

for all k, where $\mu := \frac{1-\beta}{1+\beta} \in [0,1)$.

Proof. This is a direct consequence of Theorem 8.40 (necessity) and Theorem 8.34 (sufficiency). □

Corollary 8.43. *Assume $S = \{x^\dagger\}$ and denote by $\operatorname{supp} x^\dagger := \{k \in \mathbb{N} : x^\dagger_k \neq 0\}$ the support of $x^\dagger$. Then the variational source condition* (8.16) *on ℓ^1 is satisfied if and only if for each subset K of $\operatorname{supp} x^\dagger$ there is some η in Y^* with $\|\eta\|_{Y^*} \leq \frac{\gamma}{1+\beta}$ such that*

$$\begin{cases} [A^*\,\eta]_k = \operatorname{sgn} x^\dagger_k, & \text{if } k \in K, \\ [A^*\,\eta]_k \in [-\mu, \mu], & \text{if } k \notin K \end{cases}$$

for all k, where $\mu := \frac{1-\beta}{1+\beta} \in [0, 1)$.

Proof. This corollary is a simple consequence of Corollary 8.42, because with

$$K = \{k \in \mathbb{N} : x^\dagger_k \neq 0,\, \sigma_k = -\sigma^S_k\}$$

we have a one-to-one correspondence between the subsets of $\operatorname{supp} x^\dagger$ and the restrictions of the sequences σ in Corollary 8.42 to $\operatorname{supp} x^\dagger$. □

Note that the condition $[A^*\,\eta]_k \in [-\mu, \mu]$ if $x^\dagger_k = 0$ in (8.18) and corresponding conditions in Theorems 8.34 and 8.40 are closely related to a property called *strict sparsity pattern* in [BL08, Definition 2] and *strong source condition* in [GHS11a, Condition 4.3].

8.7.4. Non-sparse solutions

We now extend Theorem 8.34 to solution sets S which may contain non-sparse solutions (cf. Remark 8.35). The aim is to obtain a variational source condition (4.2) with concave index function φ, which depends on the decay of the solutions' components. Here, again, $\|S\|_{\ell^1}$ denotes the norm of the norm minimizing solutions.

A sufficient condition for such a variational source condition can be deduced from the characterization (8.15) in Theorem 1.1.

Theorem 8.44. *Assume that*

$$\lim_{n\to\infty} \sup_{x^\dagger \in S} \sum_{k>n} |x^\dagger_k| = 0 \tag{8.19}$$

and let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers. Then the variational source condition (8.3) *on ℓ^1 is satisfied with*

$$\varphi(t) = \inf_{n\in\mathbb{N}} \left(2 \sup_{x^\dagger \in S} \sum_{k>n} |x^\dagger_k| + \gamma_n\, t \right)$$

if for each n in $\mathbb{N}$, each σ in $\mathbb{1}^S$ and each $x^\dagger$ in $S(\sigma)$ there is some η in Y^* with $\|\eta\|_{Y^*} \leq \frac{\gamma_n}{1+\beta}$ such that

$$\begin{cases} [A^*\,\eta]_k \in [-\mu, \mu], & \text{if } \sigma^S_k = 0 \text{ or } k > n, \\ \sigma_k\,[A^*\,\eta]_k \leq \mu, & \text{if } \sigma^S_k \neq 0,\ x^\dagger_k = 0,\ \sigma_k \neq 0,\ k \leq n, \\ \sigma^S_k\,[A^*\,\eta]_k \leq \mu, & \text{if } x^\dagger_k \neq 0,\ \sigma_k = \sigma^S_k,\ k \leq n, \\ \sigma^S_k\,[A^*\,\eta]_k \geq 1, & \text{if } x^\dagger_k \neq 0,\ \sigma_k = -\sigma^S_k,\ k \leq n \end{cases}$$

for all k, where $\mu := \frac{1-\beta}{1+\beta} \in [0,1)$.

Proof. Fix x in ℓ^1. By Proposition 8.31 there are σ in $\mathbb{1}^S$ and $x^\dagger$ in $S(\sigma)$ such that x is in $M_{x^\dagger}(\sigma)$. Proposition 8.29 yields

$$\beta \operatorname{dist}(x,S) - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1} = \beta \, \langle \sigma, x - x^\dagger \rangle_{\ell^\infty \times \ell^1} + \beta \sum_{k\colon \sigma_k^S = 0} |x_k| - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1}.$$

This can be written as a sum

$$\beta \operatorname{dist}(x,S) - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1} = \sum_{k\in\mathbb{N}} a_k$$

with a_k depending only on x_k and $x_k^\dagger$ and we have

$$a_k = \begin{cases} -(1-\beta)\,|x_k|, & \text{if } \sigma_k^S = 0,\\ -(1-\beta)\,|x_k - x_k^\dagger|, & \text{if } \sigma_k^S \neq 0,\ x_k^\dagger = 0,\ \sigma_k \neq 0\\ & \text{or if } x_k^\dagger \neq 0,\ \sigma_k = \sigma_k^S,\\ -\beta\,\sigma_k^S\,x_k - |x_k| + (1+\beta)\,|x_k^\dagger|, & \text{if } x_k^\dagger \neq 0,\ \sigma_k = -\sigma_k^S,\\ 0 & \text{if } \sigma_k^S \neq 0,\ \sigma_k = 0. \end{cases}$$

Now let η be as in the theorem. Then

$$\gamma_n \, \|A\,x - A\,x^\dagger\|_Y \geq -(1+\beta)\,\langle \eta, A\,x - A\,x^\dagger\rangle_{Y^* \times Y} = -(1+\beta)\,\langle A^*\,\eta, x - x^\dagger\rangle_{\ell^\infty \times \ell^1}$$

and, because $x \in M_{x^\dagger}(\sigma)$, we see

$$\gamma_n \, \|A\,x - A\,x^\dagger\|_Y \geq -(1+\beta) \sum_{k\colon \sigma_k^S \neq 0} [A^*\,\eta]_k\,\sigma_k\,|x_k - x_k^\dagger| - (1+\beta) \sum_{k\colon \sigma_k^S = 0} [A^*\,\eta]_k\,x_k.$$

Using the properties of $A^*\,\eta$ we obtain

$$2 \sum_{k>n} |x_k^\dagger| - \gamma_n \, \|A\,x - A\,x^\dagger\|_Y \geq \sum_{n\in\mathbb{N}} b_n$$

with

$$b_k \geq \begin{cases} -(1-\beta)\,|x_k|, & \text{if } \sigma_k^S = 0,\\ -(1-\beta)\,|x_k - x_k^\dagger|, & \text{if } \sigma_k^S \neq 0,\ x_k^\dagger = 0,\ \sigma_k \neq 0,\ k \leq n,\\ & \text{or if } x_k^\dagger \neq 0,\ \sigma_k = \sigma_k^S,\ k \leq n,\\ (1+\beta)\,|x_k - x_k^\dagger|, & \text{if } x_k^\dagger \neq 0,\ \sigma_k = -\sigma_k^S,\ k \leq n,\\ 2\,|x_k^\dagger| - (1-\beta)\,|x_k - x_k^\dagger|, & \text{if } \sigma_k^S \neq 0,\ \sigma_k \neq 0,\ k > n,\\ 2\,|x_k^\dagger| & \text{if } \sigma_k^S \neq 0,\ \sigma_k = 0. \end{cases}$$

It is not hard to show that $a_k \leq b_k$ for all k. Thus,

$$\beta \operatorname{dist}(x,S) - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1} \leq 2 \sum_{k>n} |x_k^\dagger| - \gamma_n \, \|A\,x - A\,x^\dagger\|_Y.$$

Taking the supremum over all $x^\dagger$ and the infimum over all n the variational source condition (8.3) is proven and it remains to show that the function φ is a concave index function.

Obviously, φ is non-negative. As an infimum of affine functions it further is concave and upper semi-continuous. Thus, φ is continuous on the interior $(0,\infty)$ of its domain. Together with

$$\varphi(0) = \inf_{n\in\mathbb{N}} \left(2 \sup_{x^\dagger \in S} \sum_{k>n} |x_k^\dagger| \right) = 0$$

we obtain continuity on $[0,\infty)$. Monotonicity of φ follows from $\gamma_n > 0$ for all n. That φ is strictly increasing in a neighborhood of the origin follows from $\varphi(0) = 0$ and $\varphi(t) > 0$ for all positive t, where $\varphi(t) > 0$ is again a consequence of $\gamma_n > 0$ for all n. □

Note that condition (8.19) may be violated in some cases. For example if S is the convex hull of the standard unit sequences $\{e^{(1)}, e^{(2)}, \ldots\}$ in ℓ^1, then

$$\sup_{x^\dagger\in S} \sum_{k>n} |x_k^\dagger| \geq \sum_{k>n} |e_k^{(n+1)}| = 1 \tag{8.20}$$

for all n.

8.7.5. Examples

We provide two very simple examples and a more realistic one to show how the developed results can be applied to non-injective operators. The first example considers multiple norm minimizing solutions. The second and the third one have only one norm minimizing solution and they show, by the way, that the constant β in a variational source condition cannot be chosen arbitrarily close to one.

Example 8.45. For the first example take $Y := \mathbb{R}$, $y^\dagger := 1$ and

$$A\,x := x_1 + x_2.$$

Then the set of solutions is $L = \{x \in \ell^1 : x_2 = 1 - x_1\}$ and the set of norm minimizing solutions is

$$S = \{x \in \ell^1 : x_2 = 1 - x_1,\ x_1 \in [0,1],\ x_k = 0 \text{ for } k > 2\}.$$

Further,

$$A^*\,\eta := (\eta, \eta, 0, \ldots).$$

Figure 8.1 provides a sketch of the geometric situation.

We now verify condition (8.15) in Theorem 8.34 with $\beta = 1$. First note that $\sigma^S = (1, 1, 0, \ldots)$ and that by Remark 8.39 we only have to consider

$$\begin{aligned}
\sigma^{(1)} &= (1, 1, 0, \ldots),\\
\sigma^{(2)} &= (1, -1, 0, \ldots),\\
\sigma^{(3)} &= (-1, 1, 0, \ldots),\\
\sigma^{(4)} &= (-1, -1, 0, \ldots).
\end{aligned}$$

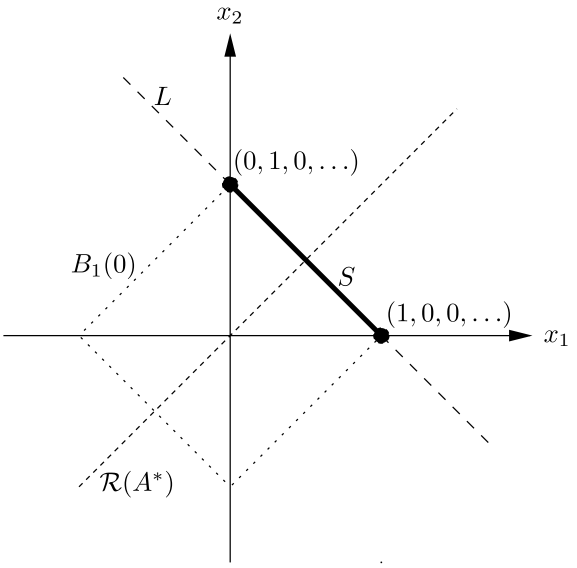

Figure 8.1.: Sketch for the first example of the x_1-x_2-plane with set S of norm minimizing solutions, set L of all solutions, unit ball $B_1(0)$ and 'subspace' $\mathcal{R}(A^*)$.

The corresponding subsets $S(\sigma^{(i)})$ of S are the faces of S looking in direction $\sigma^{(i)}$, that is,

$$\begin{aligned} S(\sigma^{(1)}) &= S, \\ S(\sigma^{(2)}) &= \{(1,0,0,\dots)\}, \\ S(\sigma^{(3)}) &= \{(0,1,0,\dots)\}, \\ S(\sigma^{(4)}) &= S. \end{aligned}$$

Taking into account Remark 8.38, only $\sigma^{(4)}$ remains to be considered. Here condition (8.15) is equivalent to $\eta \geq 1$, which is obviously satisfied when choosing $\eta = 1$ (by Remark 8.37 we only have to check the condition for $x^\dagger = (\frac{1}{2}, \frac{1}{2}, 0, \dots)$ for example). Consequently, Theorem (8.34) applies to our first example and yields convergence rates although the operator A is not injective. □

Example 8.46. For the second example take $Y := \mathbb{R}$, $y^\dagger := 1$ and

$$A\,x := x_1 + \frac{x_2}{2}.$$

Then the set of solutions is $L = \{x \in \ell^1 : x_2 = 2 - 2\,x_1\}$ and there is only one norm minimizing solution

$$S = \{(1,0,\dots)\}.$$

Further,

$$A^*\,\eta := (\eta, \frac{\eta}{2}, 0, \dots).$$

Figure 8.2 provides a sketch of the geometric situation.

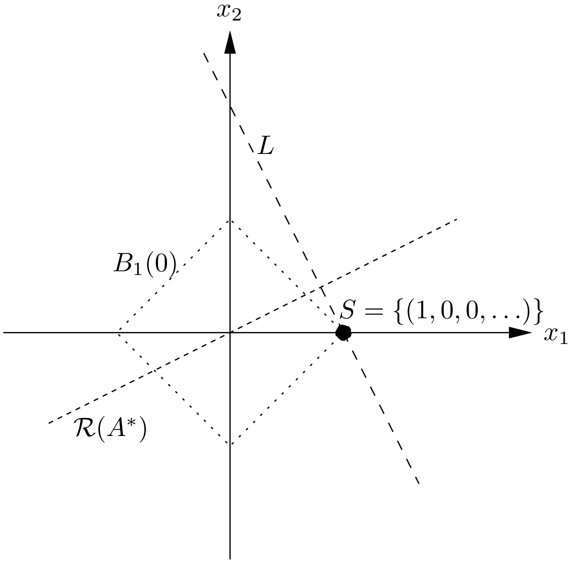

Figure 8.2.: Sketch for the second example of the x_1-x_2-plane with set S of norm minimizing solutions, set L of all solutions, unit ball $B_1(0)$ and 'subspace' $\mathcal{R}(A^*)$.

We now verify condition (8.18) in Corollary 8.42. First note that $\sigma^S = (1,0,0,\ldots)$ and so we only have to consider

$$\sigma^{(1)} = (1,0,\ldots) \quad \text{and} \quad \sigma^{(2)} = (-1,0,\ldots).$$

For $\sigma^{(1)}$ condition (8.18) is satisfied by $\eta = 0$. For $\sigma^{(2)}$ the condition is equivalent to

$$\eta = 1 \quad \text{and} \quad -\frac{1-\beta}{1+\beta} \leq \frac{\eta}{2} \leq \frac{1-\beta}{1+\beta},$$

which is only possible if $\beta \leq \frac{1}{3}$. Consequently, Corollary 8.42 yields a variational source condition with $\beta \leq \frac{1}{3}$ and corresponding convergence rates for our second non-injective example.

If we had chosen the solution set L to be parallel to the x_2-axis, then $\beta = 1$ would be possible. On the other hand, the more slanting the set L in Figure 8.2 is, the closer β has to be to zero. The limit case where only $\beta = 0$ would be possible then coincides with the situation discussed in Example 8.45. Generalizing this observation we may say that the constant β in a variational source condition is a 'measure' for paraxiality of the nullspace of A or the range of A^*. □

Example 8.47. Consider the example introduced in Subsection 6.4.4. We look at the exact right-hand side $y^\dagger := (0,0,-\frac{1}{2\pi},0,-\frac{1}{4\pi},0,0,\ldots)$. One easily sees that $x^\dagger = (0,1,0,1,0,0,\ldots)$ is a corresponding solution and it turns out that this is the only 1-norm minimizing solution, that is, $S = \{x^\dagger\}$ (here some very basic but longish calculations are necessary).

To verify the assumptions of Corollary 8.42 we have to show that the elements

$$\begin{aligned} \sigma^{(1)} &= (0,1,0,1,0,0,\ldots), \\ \sigma^{(2)} &= (0,1,0,-1,0,0,\ldots), \\ \sigma^{(3)} &= (0,-1,0,1,0,0,\ldots), \\ \sigma^{(4)} &= (0,-1,0,-1,0,0,\ldots) \end{aligned}$$

satisfy condition (8.18) for some η. We only mention how to choose η in each case and do not provide all details of the (basic but longish) calculations. Since $\sigma^S = (0,1,0,1,0,0,\ldots)$ we may choose $\eta = 0$ in case of $\sigma^{(1)}$. For $\sigma^{(2)}$ one possible choice is

$$\eta = \left(2\sqrt{2} + \frac{4\pi - 4}{2\pi + \sqrt{2}}, 0, 0, 0, 4\pi, 0, 0, \ldots\right) \qquad \text{if} \quad \mu \geq \frac{2\pi - 2}{2\pi + \sqrt{2}}.$$

Note that for smaller μ there is no η satisfying (8.18) if $\sigma = \sigma^{(2)}$. For $\sigma^{(3)}$ one possible choice is

$$\eta = \left(2\sqrt{2} + \frac{2\pi - 4}{\pi + \sqrt{2}}, 0, 2\pi, 0, 0, \ldots\right) \qquad \text{if} \quad \mu \geq \frac{\pi - 2}{\pi + \sqrt{2}}.$$

Again for smaller μ there is no η satisfying (8.18) if $\sigma = \sigma^{(3)}$. Finally, for $\sigma^{(4)}$ we may choose

$$\eta = \left(\frac{12}{5}\pi, 0, 3\pi, 0, 4\pi, 0, 0, \ldots\right) \qquad \text{if} \quad \mu \geq \frac{2}{5}$$

and for smaller μ there is no η.

Thus, if

$$\mu \geq \frac{2\pi - 2}{2\pi + \sqrt{2}} \approx 0.5564$$

we obtain a variational source condition with

$$\beta \leq \frac{2 + \sqrt{2}}{4\pi - 2 + \sqrt{2}} \approx 0.2850$$

and corresponding convergence rates.

Playing around with this example one also sees that the more non-zero components $x^\dagger$ has the smaller is the best possible β in the variational source condition. Since β enters the $\mathcal{O}$-constant c in the convergence rate result (4.1) as a factor $\frac{1}{\beta}$ (cf. Propositions 8.2 and 8.3), the $\mathcal{O}$-constant becomes greater if $x^\dagger$ is 'less sparse'. If the number of non-zero components in $x^\dagger$ goes to infinity, then β goes to zero and consequently the $\mathcal{O}$-constant blows up to infinity. Such situations then can be handled by Theorem 8.44, resulting in slower convergence rates. □

9. Open Problems

To the author's regret two problems closely related to this thesis remain unsolved. Here we briefly describe both.

The first problem is the question whether source-type conditions (8.15) are always satisfied in the non-injective case. For injective operators A we gave a positive answer in Section 8.6. For non-injective operators the author conjectures that a similar result holds true: if A is weak*-to-weak continuous and if $\mathcal{N}(A)$ does not 'approximate' a face of the unit ball in ℓ^1 arbitrarily well, then the source-type condition (8.15) is always satisfied. More precisely, the number constructed as follows has to be strictly smaller than one: intersect $\mathcal{N}(A)$ with the unit ball of ℓ^∞, for each element in this intersection take the maximal absolute value of the components not being 1 or -1, take the supremum over all these maxima. This number then plays the role of μ in (8.15).

The second open problem is related to variational source conditions in general. For linear ill-posed problems in Hilbert spaces it is known that variational source conditions are not only sufficient but also necessary for corresponding convergence rates. In more general settings, for example in our ℓ^1-setting, it is completely unclear whether an error estimate with rate function φ implies a variational source condition with φ, that is, whether variational source conditions yield the optimal rate. The author conjectures that up to few special cases such converse results in Banach spaces can be shown.

Appendix

A. Topology, functional analysis, convex analysis

We briefly summarize some definitions and results from set theoretic topology, functional analysis and convex analysis used in the thesis. Everything can be found in standard literature on the subjects. The focus is on ℓ^1 and related spaces. We recommend the books [Die84, Meg98] to the reader interested in the precipices of non-reflexive Banach spaces.

A.1. Topological spaces and nets

A *topological space* (X, τ) is a non-empty set X endowed with a topology τ. A *topology* on X is a family of subsets of X with the following properties:

- $\emptyset \in \tau$ and $X \in \tau$,
- intersections of finitely many sets in τ belong to τ,
- unions of arbitrarily many sets in τ belong to τ.

A topology τ_1 on X is *weaker* or *coarser* than a topology τ_2 on X if $\tau_1 \subseteq \tau_2$. In this case, τ_2 is *stronger* or *finer* than τ_1.

The sets in τ are called *open sets*, their complements *closed sets*. The *interior* $\operatorname{int} B$ of a subset B of X is the union of all open sets contained in B. The *closure* $\overline{B}$ is the intersection of all closed sets containing B.

A subset B of X is called *compact* if every covering of B with open sets contains a finite covering of B. The set B is *relatively compact* if its closure is compact.

A *neighborhood* of an element x in X is a subset B of X which contains on open set C with $x \in C$.

Given two topological spaces (X, τ_X) and (Y, τ_Y) and an element x in X, a mapping $f : X \rightarrow Y$ is called *continuous at x* if the full preimage of each neighborhood of $f(x)$ is a neighborhood of x. The mapping f is *continuous* if it is continuous at every element x. One can show that f is continuous if and only if the preimages of open sets are open sets or, equivalently, if the preimages of closed sets are closed.

A mapping $f : X \rightarrow (-\infty, \infty]$ is *lower semi-continuous* if the sublevel sets $\{x \in X : f(x) \leq c\}$, $c \in \mathbb{R}$, are closed.

A non-empty set I is *directed* if there is a relation $\preceq$ on I with the following properties:

- $i \preceq i$ for all i in I,
- $i_1 \preceq_2$, $i_2 \preceq i_3$ implies $i_1 \preceq i_3$ for all i_1, i_2, i_3 in I,

- for all i_1 and i_2 in I there is some i_3 in I such that $i_1 \preceq i_3$ and $i_2 \preceq i_3$.

Given a directed set I, a *net* $(x_i)_{i\in I}$ is a mapping from I into X. A net $(x_i)_{i\in I}$ *converges to* x in X if for each neighborhood B of x there is some i_B in I with $x_i \in B$ if $i \succeq i_B$. A *sequence* is a net with index set $\mathbb{N}$, where $\preceq$ is the usual ordering in $\mathbb{N}$.

A convergent net may have more than one limit. A topological space is called *Hausdorff space* if for each pair of distinct elements there is a pair of corresponding disjoint neighborhoods. Each net has at most one limit if and only if (X, τ) is a Hausdorff space.

A set B is closed if and only if all limits of convergent nets in B belong to B. A mapping $f : X \to Y$ is continuous if and only if for each convergent net $(x_i)_{i\in I}$ with limit x_0 the net $(f(x_i))_{i\in I}$ converges to $f(x_0)$. A mapping $f : X \to (-\infty, \infty]$ is lower semi-continuous if and only if for each net $(x_i)_{i\in I}$ with limit x_0 we have $f(x_0) \leq \liminf_{i\in I} f(x_i)$. Compactness of sets can be characterized in terms of net convergence, too, but this requires introduction of subnets, which is beyond the scope of this chapter.

A set B is *sequentially closed* if and only if all limits of convergent sequences in B belong to B. Analogously, *sequential compactness*, *sequential continuity* and *sequential semi-continuity* can be introduced. The sequential and non-sequential versions coincide if (X, τ) is an A_1*-space*, that is, if for each x in X there is a sequence $(B_k)_{k\in\mathbb{N}}$ of neighborhoods of x such that for each neighborhood C of x we find k with $B_k \subseteq C$. Compactness is an exception here, but one can show that in A_1-spaces compactness implies sequential compactness.

A.2. Reflexivity, weak and weak* topologies

A topological vector space (X, τ) is a vector space X endowed with a topology τ with respect to which the vector space operations (addition, multiplication by scalars) are continuous. Typical examples are normed vector spaces, where the topology is induced by the norm. That is, the topology contains exactly the sets which are open with respect to the norm.

In a normed vector space we may introduce another topology. The weakest topology with respect to which all norm-continuous linear functionals are continuous is called *weak topology* on X. The vector space X endowed with the weak topology is a topological vector space and a Hausdorff space but not necessarily an A_1-space (cf. previous section). Denoting the set of norm-continuous linear functionals on X by X^*, one can show that a net $(x_i)_{i\in I}$ converges weakly to x_0 in X if and only if

$$\lim_{i\in I} \langle \xi, x_i \rangle_{X^*\times X} = \langle \xi, x_0 \rangle_{X^*\times X} \qquad \text{for all } \xi \text{ in } X^*.$$

The set X^* of norm-continuous linear functionals on X forms a normed vector space itself and thus has a weak topology, too. X^* is called the *dual space* of X. The dual space X^{**} of X^* is closely related to X: Define the mapping $E : X \to X^{**}$ by

$$\langle E\,x, \xi \rangle_{X^{**}\times X^*} := \langle \xi, x \rangle_{X^*\times X}, \quad \xi \in X^*,$$

for all x in X. Then E is continuous (with respect to the norm topologies), isometric, injective and continuously invertible on its range. In other words, E is an isometric isomorphism between X and $\mathcal{R}(E)$. If E is surjective, then X is said to be *reflexive*. If we write $x \in X^{**}$ for some x in X, then this has to be understood as $E\,x \in X^{**}$.

The space ℓ^1 of absolutely summable sequences is not reflexive. Its dual is ℓ^∞, the space of absolutely bounded sequences. In addtion, we know that ℓ^1 is the dual of c_0, the space of sequences converging to zero. The weak topology on ℓ^1 is strictly weaker than the norm topology, but a sequence (not a general net) in ℓ^1 is weakly convergent if and only if it is norm-convergent. This remarkable property is known as *Schur's property*.

Next to the norm topology and the weak topology the dual space X^* of a normed vector space X carries another useful topology. With E as above the weak* topology is the weakest topology on X^* with respect to which all norm-continuous linear functionals on X^* belonging to $\mathcal{R}(E)$ are continuous. This topology is obviously weaker than the weak topology on X^*. A net $(\xi_i)_{i\in I}$ is weakly* convergent to ξ_0 in X^* if and only if

$$\lim_{i\in I} \langle u, \xi_i\rangle_{X^{**}\times X^*} = \langle u, \xi_0\rangle_{X^{**}\times X^*} \qquad \text{for all } u \text{ in } \mathcal{R}(E),$$

which by the definition of E is equivalent to

$$\lim_{i\in I} \langle \xi_i, x\rangle_{X^*\times X} = \langle \xi_0, x\rangle_{X^*\times X} \qquad \text{for all } x \text{ in } X.$$

Weak and weak* topology on X^* coincide if and only if X is reflexive. The use of the weak* topology comes from the fact, that closed balls in X^* are weakly* compact (Banach–Alaoglu theorem). If X is separable, then closed balls in X^* are weakly* sequentially compact (sequential Banach–Alaoglu theorem). With respect to the norm topology closed balls are compact if and only if the space is finite-dimensional (Heine–Borel property). With respect to the weak topology closed balls are compact if and only if the space is reflexive.

Because ℓ^1 is the dual of c_0, the weak* topology is available on ℓ^1. The sublevel sets of the ℓ^1-norm are weakly* compact and the ℓ^1-norm is a weakly* lower semi-continuous functional.

In Hilbert spaces each closed subspace has a complement, that is, there is a second closed subspace such that the direct sum of both subspaces equals the whole Hilbert space. In infinite-dimensional Banach spaces there always exist closed subspaces which do not have a complement, see [Meg98, page 301]. A *complemented subspace* is a closed subspace with complement.

A.3. Subdifferentials and Bregman distances

Let X be a real normed vector space and let X^* be its dual. In the following we consider functionals on X which may attain the value $+\infty$ or $-\infty$, but not both. Addition and multiplication by scalars are extended to such functionals whenever this extension is intuitive, e. g. $1 + \infty = +\infty$ or $1 \cdot (+\infty) = +\infty$.

A functional $\Omega : X \to (-\infty, \infty]$ is *convex* if

$$\Omega(\lambda\, x_1 + (1-\lambda)\, x_2) \leq \lambda\, \Omega(x_1) + (1-\lambda)\, \Omega(x_2)$$

for all x_1, x_2 in X and all λ in $[0,1]$. A functional $\Omega : X \to [-\infty, \infty)$ is *concave* if $-\Omega$ is convex. A functional is *proper* if it attains a finite value.

For a convex functional $\Omega : X \to (-\infty, \infty]$ an element ξ in X^* is called a *subgradient* of Ω at x_0 if

$$\Omega(x) \geq \Omega(x_0) + \langle \xi, x - x_0 \rangle_{X^* \times X} \qquad \text{for all } x \text{ in } X.$$

The set of all subgradients at x_0 is called *subdifferential* of Ω at x_0 and is denoted by $\partial\Omega(x_0)$. If $\Omega(x_0) = \infty$ and if there is some x in X with $\Omega(x) < \infty$, then $\partial\Omega(x_0) = \emptyset$. But also in case $\Omega(x_0) < \infty$ it may happen that $\partial\Omega(x_0) = \emptyset$.

For $X = \ell^1$ we have

$$\xi \in \partial(\|\cdot\|_{\ell^1})(x) \qquad \Leftrightarrow \qquad \xi_k \begin{cases} = 1, & \text{if } x_k > 1, \\ = -1, & \text{if } x_k < 1, \\ \in [0,1], & \text{if } x_k = 0, \end{cases} \quad k \in \mathbb{N}.$$

An element x minimizes a lower semi-continuous proper convex functional $\Omega : X \to (-\infty, \infty]$ over X if and only if $0 \in \partial\Omega(x)$. For minimization with constraints we have to introduce normal cones: The *normal cone* of a convex set C at a point x in C is the set

$$N_C(x) := \{\xi \in X^* : \langle \xi, \tilde{x} - x \rangle_{X^* \times X} \leq 0\}.$$

Let $\Omega : X \to (-\infty, \infty]$ be proper, convex and lower semi-continuous and let C be a closed convex set. Then x in C minimizes Ω over C if and only if $0 \in \partial\Omega(x) + N_C(x)$.

Based on a convex functional $\Omega : X \to (-\infty, \infty]$ one can define another convex functional which expresses the distance between Ω and one of its linearizations at a fixed point x_0: For ξ_0 in $\partial\Omega(x_0)$ the functional $B^{\Omega}_{\xi_0}(\cdot, x_0) : X \to [0, \infty]$ defined by

$$B^{\Omega}_{\xi_0}(x, x_0) := \Omega(x) - \Omega(x_0) - \langle \xi_0, x - x_0 \rangle_{X^* \times X}, \quad x \in X,$$

is called *Bregman distance* with respect to Ω, x_0 and ξ_0. The Bregman distance can only be defined for x_0 in X with $\partial\Omega(x_0) \neq \emptyset$. Since Ω is assumed to be convex, the Bregman distance is also convex. The non-negativity of $B^{\Omega}_{\xi_0}(\cdot, x_0)$ follows from $\xi_0 \in \partial\Omega(x_0)$.

If X is a Hilbert space and $\Omega = \frac{1}{2}\|\cdot\|^2$, then $\partial\Omega(x_0) = \{x_0\}$ and the corresponding Bregman distance is given by $B^{\Omega}_{x_0}(\cdot, x_0) = \frac{1}{2}\|x - x_0\|^2$. Thus, Bregman distances can be regarded as a generalization of Hilbert space norms.

B. Verification of Assumption 8.15 for Example 8.20

Here we present the detailed derivation of the results in Example 8.20. The basic setting for this example was introduced in Subsection 6.4.3. Remember $\tilde{X} = Y = L^2(0,1)$ and that $\tilde{e}_{1+2^l+k}$ denotes the k-th element of level l of the Haar basis in $L^2(0,1)$. The mapping $\tilde{A}$ assigns to a function its antiderivative and $\tilde{A}^*$ has a very similar structure. The functions in the range of $\tilde{A}^*$ are continuous, have a (generalized) derivative and are zero at the right end of the interval $(0,1)$. More precisely, for $\tilde{x}$ in $L^2(0,1)$ and η in $L^2(0,1)$ we have

$$\tilde{x} = \tilde{A}^* \eta \qquad \Leftrightarrow \qquad \eta = -\tilde{x}', \qquad \tilde{x}(1) = 0.$$

We use this observation to construct a sequence $(\eta_j^{(n)})_{j\in\mathbb{N}}$ as in Assumption 8.15. Fix m in $\mathbb{N}_0$, j in $\mathbb{N}$ and σ in $\{-1,0,1\}^{\mathbb{N}}$ and set $n := 2^m$. To simplify notation we write η instead of $\eta_j^{(n)}$. If we find some $\tilde{x}$ in $L^2(0,1)$ such that

- the Haar transform ξ in ℓ^∞ of $\tilde{x}$ satisfies (i),
- $\|(I - P_n)\,\xi\|_{\ell^\infty}$ is bounded in terms of j and the bound goes to zero if $j \to \infty$,
- $\tilde{x}$ is continuous and piecewise differentiable,
- $\tilde{x}(1) = 0$,

then we are done.

A suitable function $\tilde{x}$ can be constructed as a linear combination of 2^m hat-like functions with disjoint supports, where the upper width of each hat depends on j. For $j = 1$ we would have a triangular hat and for $j \to \infty$ we would approximate a box-shaped hat. Corresponding formulas for the initial hat-like function $h^{(j)}$ and the scaled translates $h_{l,k}^{(j)}$ were given in Example 8.20.

Denoting by ξ the Haar transform of $\tilde{x}$ we have to ensure $P_n\,\xi = P_n\,\sigma$. For this purpose we construct functions $g_i^{(m)}$ for $i = 1,\ldots,2^m$ such that P_n applied to corresponding Haar transforms yields the standard unit vectors $e^{(i)}$. Here is the formula:

$$g_i^{(m)} := 2^{-\frac{m}{2}} \sum_{r=0}^{2^m-1} \tilde{e}_i\left(\frac{r+\frac{1}{2}}{2^m}\right) h_{m,r}^{(j)}.$$

The Haar basis functions $\tilde{e}_i$ are constant on the intervals $(\frac{r}{2^m}, \frac{r+1}{2^m})$ and each such interval is the support of the hat-like function $h_{m,r}^{(j)}$. The formula for $g_i^{(m)}$ thus re-samples the function graph of $\tilde{e}_i$ with the help of the hat-like functions. Simple calculations show that indeed P_n applied to the Haar transform of $g_i^{(m)}$ yields the standard unit vector $e^{(i)}$. Now set

$$\tilde{x} := \sum_{i=1}^{2^m} \sigma_i \, g_i^{(m)} \qquad \text{and} \qquad \eta := -\tilde{x}'$$

and denote by ξ the corresponding Haar transform. Then $P_n \, \xi = P_n \, \sigma$.

Since $\tilde{x}$ is obviously continuous and piecewise differentiable and satisfies $\tilde{x}(1) = 0$ it remains to calculate a sufficiently small upper bound for $\|(I - P_n)\,\xi\|_{\ell^\infty}$. The symmetry of the hat-like functions $h_{m,r}^{(j)}$ ensures $\xi_{1+2^m+r} = 0$ for $r = 0, \ldots, 2^m - 1$. Thus, only $|\xi_{1+2^p+q}|$ for $p > m$ and $q = 0, \ldots, 2^p - 1$ have to be considered.

Denote by E^* the Haar transform operator (cf. Subsection 6.4.3). Then

$$\begin{aligned}
|\xi_{1+2^p+q}| &= 2^{-\frac{m}{2}} \left| \sum_{r=0}^{2^m-1} \left(\sigma_1 + \sum_{l=0}^{m-1} \sum_{k=0}^{2^l-1} \sigma_{1+2^l+k} \, \tilde{e}_{1+2^l+k} \left(\frac{r + \frac{1}{2}}{2^m} \right) \right) [E^* \, h_{m,r}^{(j)}]_{1+2^p+q} \right| \\
&\leq 2^{-\frac{m}{2}} \sum_{r=0}^{2^m-1} \left(1 + \sum_{l=0}^{m-1} \sum_{k=0}^{2^l-1} \left| \tilde{e}_{1+2^l+k} \left(\frac{r + \frac{1}{2}}{2^m} \right) \right| \right) \left| [E^* \, h_{m,r}^{(j)}]_{1+2^p+q} \right|.
\end{aligned}$$

From $p > m$ we see that $[E^* \, h_{m,r}^{(j)}]_{1+2^p+q} \neq 0$ for only one r in $\{0, \ldots, 2^m - 1\}$. Denote this index r by $r(p,q)$. From $l < m$ we see that the inner sum of the inner double sum has only one non-vanishing summand. Denote the corresponding index k by $k(r,l)$. Consequently

$$\begin{aligned}
|\xi_{1+2^p+q}| &\leq 2^{-\frac{m}{2}} \left(1 + \sum_{l=0}^{m-1} \left| \tilde{e}_{1+2^l+k(r(p,q),l)} \left(\frac{r(p,q) + \frac{1}{2}}{2^m} \right) \right| \right) \left| [E^* \, h_{m,r(p,q)}^{(j)}]_{1+2^p+q} \right| \\
&\leq 2^{-\frac{m}{2}} \left(1 + \sum_{l=0}^{m-1} 2^{\frac{l}{2}} \right) \left| [E^* \, h_{m,r(p,q)}^{(j)}]_{1+2^p+q} \right| \\
&= 2^{-\frac{m}{2}} \left(1 + \frac{2^{\frac{m}{2}} - 1}{\sqrt{2} - 1} \right) \left| [E^* \, h_{m,r(p,q)}^{(j)}]_{1+2^p+q} \right| \\
&= \left(\frac{\sqrt{2} - 2}{\sqrt{2} - 1} \, 2^{-\frac{m}{2}} + \frac{1}{\sqrt{2} - 1} \right) \left| [E^* \, h_{m,r(p,q)}^{(j)}]_{1+2^p+q} \right| \\
&\leq \frac{1}{\sqrt{2} - 1} \left| [E^* \, h_{m,r(p,q)}^{(j)}]_{1+2^p+q} \right|
\end{aligned}$$

and

$$\begin{aligned}
&\left| [E^* \, h_{m,r(p,q)}^{(j)}]_{1+2^p+q} \right| \\
&\qquad \leq \sup\left\{ \left| [E^* \, h_{m,r}^{(j)}]_{1+2^l+k} \right| : \, r \in \{0, \ldots, 2^m - 1\}, \, l > m, \, k \in \{0, \ldots, 2^l - 1\} \right\}.
\end{aligned}$$

The supremum does not change if we only consider $r = 0$. In addition, from simple geometric considerations we see that for $r = 0$ and fixed l the objective function is maximized at $k = 0$. Hence,

$$|\xi_{1+2^p+q}| \leq \frac{1}{\sqrt{2} - 1} \sup\left\{ \left| [E^* \, h_{m,0}^{(j)}]_{1+2^l+0} \right| : \, l > m \right\}.$$

For $m < l \le m + j - 1$ we have

$$\left|[E^* h^{(j)}_{m,0}]_{1+2^l+0}\right| = \left|\int\limits_0^1 \psi_{l,0}(t)\, h^{(j)}_{m,0}(t)\,\mathrm{d}t\right| = \left|\frac{1}{1-2^{-j}}\, 2^{\frac{l-m}{2}-j-1}\right| \le \left|\frac{1}{1-2^{-j}}\, 2^{-\frac{j}{2}-\frac{3}{2}}\right|$$

and for $l \ge m + j$ we have

$$\left|[E^* h^{(j)}_{m,0}]_{1+2^l+0}\right| = \left|\int\limits_0^1 \psi_{l,0}(t)\, h^{(j)}_{m,0}(t)\,\mathrm{d}t\right| = \left|\frac{1}{1-2^{-j}}\, 2^{\frac{3}{2}(m-l)+j-2}\right| \le \left|\frac{1}{1-2^{-j}}\, 2^{-\frac{j}{2}-2}\right|.$$

In both cases

$$|\xi_{1+2^p+q}| \le \frac{1}{\sqrt{2}-1}\,\left|\frac{1}{1-2^{-j}}\, 2^{-\frac{j}{2}-\frac{3}{2}}\right| \le \frac{1}{\sqrt{2}-1}\,\left|2\cdot 2^{-\frac{j}{2}-\frac{3}{2}}\right| = \frac{1}{2-\sqrt{2}}\, 2^{-\frac{j}{2}}.$$

This proves the bound

$$\|(I-P_n)\,\xi\|_{\ell^\infty} \le \frac{2^{-\frac{j}{2}}}{2-\sqrt{2}}$$

and the bound goes to zero if $j \to \infty$.

Now, that Assumption 8.15 has been verified, we calculate the constant γ_n in Theorem 8.17. We keep all the notation introduced so far. We have

$$\|\eta\|^2_{Y^*} = 2^{-m} \sum_{r=0}^{2^m-1} \left(\sigma_1 + \sum_{l=0}^{m-1}\sum_{k=0}^{2^l-1} \sigma_{1+2^l+k}\, \tilde{e}_{1+2^l+k}\left(\frac{r+\frac{1}{2}}{2^m}\right)\right)^2 \int\limits_{\frac{r}{2^m}}^{\frac{r+1}{2^m}} \left(h^{(j)}_{m,r}\right)'(t)^2\,\mathrm{d}t$$

and

$$\int\limits_{\frac{r}{2^m}}^{\frac{r+1}{2^m}} \left(h^{(j)}_{m,r}\right)'(t)^2\,\mathrm{d}t = \int\limits_{\frac{r}{2^m}}^{\frac{r+1}{2^m}} \left(2^{\frac{3}{2}m}\left(h^{(j)}\right)'(2^m\,t-r)\right)^2\,\mathrm{d}t = 2^{2\,m}\int\limits_0^1 \left(h^{(j)}\right)'(t)^2\,\mathrm{d}t,$$

which leads to

$$\|\eta\|^2_{Y^*} = 2^m \left(\int\limits_0^1 \left(h^{(j)}\right)'(t)^2\,\mathrm{d}t\right) \sum_{r=0}^{2^m-1} \left(\sigma_1 + \sum_{l=0}^{m-1}\sum_{k=0}^{2^l-1} \sigma_{1+2^l+k}\, \tilde{e}_{1+2^l+k}\left(\frac{r+\frac{1}{2}}{2^m}\right)\right)^2.$$

Above, when estimating ξ_{1+2^p+q}, we saw that the inner double sum can be bounded above by

$$\sum_{l=0}^{m-1}\sum_{k=0}^{2^l-1} \sigma_{1+2^l+k}\, \tilde{e}_{1+2^l+k}\left(\frac{r+\frac{1}{2}}{2^m}\right) \le \frac{2^{\frac{m}{2}}-1}{\sqrt{2}-1}.$$

Thus,

$$\begin{aligned}\|\eta\|_{Y^*}^2 &\le 2^m \left(\int\limits_0^1 \big(h^{(j)}\big)'(t)^2 \,\mathrm{d}t\right) \sum_{r=0}^{2^m-1} \left(1+\frac{2^{\frac{m}{2}}-1}{\sqrt{2}-1}\right)^2 \\ &= 2^{2\,m} \left(\int\limits_0^1 \big(h^{(j)}\big)'(t)^2 \,\mathrm{d}t\right) \left(1+\frac{2^{\frac{m}{2}}-1}{\sqrt{2}-1}\right)^2 \\ &= 2^{2\,m} \left(1+\frac{2^{\frac{m}{2}}-1}{\sqrt{2}-1}\right)^2 \frac{2^{j+1}}{(1-2^{-j})^2}\end{aligned}$$

and therefore

$$\gamma_n \le \|\eta\|_{Y^*} \le 2^m \left(1+\frac{2^{\frac{m}{2}}-1}{\sqrt{2}-1}\right) \frac{2^{\frac{j+1}{2}}}{1-2^{-j}} \le 2^m \,\frac{2^{\frac{m}{2}}}{\sqrt{2}-1}\, 2^{\frac{j}{2}+\frac{3}{2}} \le \frac{2\sqrt{2}}{\sqrt{2}-1}\, 2^{\frac{j}{2}} \left(2^m\right)^{\frac{3}{2}}.$$

The index j has to be chosen large enough to ensure $\|(I-P_n)\,A^*\,\eta\|_{\ell^\infty} \le \frac{1-\beta}{1+\beta}$ for prescribed β in $(0,1)$. Since we have already seen

$$\|(I-P_n)\,A^*\,\eta\|_{\ell^\infty} \le \frac{2^{-\frac{j}{2}}}{2-\sqrt{2}}$$

a suitable choice for j is

$$j(\beta) := \left\lceil \frac{2}{\ln 2} \ln \frac{1+\beta}{(2-\sqrt{2})\,(1-\beta)} \right\rceil .$$

C. Variational source conditions for nonlinear equations in Banach spaces

Complementing Theorem 4.4 we proof an analogous result for Tikhonov regularization with convex penalties in Banach spaces, which has not been published elsewhere up to now. Let X and Y be Banach spaces and let $F : X \supseteq \mathcal{D}(F) \to Y$ be a (nonlinear) mapping. To obtain approximate but stable solutions to the possibly ill-posed equation

$$F(x) = y^\dagger, \quad x \in \mathcal{D}(F), \tag{C.1}$$

with exact right-hand side $y^\dagger$ from the range of F we search for minimizers of the Tikhonov-type functional

$$T_\alpha^\delta(x) := \|F(x) - y^\delta\|^p + \alpha\,\Omega(x), \quad x \in \mathcal{D}(F).$$

Here, y^δ in Y is the available data and is assumed to satisfy

$$\|y^\delta - y^\dagger\| \leq \delta$$

for a positive noise level δ and $\Omega : X \to (-\infty, \infty]$ is a convex functional. We assume $p > 1$ and $\alpha > 0$.

To guarantee existence, stability and convergence of the minimizers the following assumptions suffice (cf. [SKHK12, Section 4.1] or [Fle12, Chapter 3])

Assumption C.1. Let X, Y, F and Ω be as introduced above. In addition we assume that

(i) F is weakly sequentially continuous,

(ii) $\mathcal{D}(F)$ is weakly sequentially closed,

(iii) Ω is convex and proper,

(iv) the sublevel sets of Ω are weakly sequentially compact.

As sufficient condition for convergence rates one may consider variational source conditions

$$\beta\, B_{\xi^\dagger}^\Omega(x, x^\dagger) \leq \Omega(x) - \Omega(x^\dagger) + \varphi(\|F(x) - F(x^\dagger\|), \quad x \in \mathcal{D}(F), \tag{C.2}$$

where $\beta > 0$, φ is a concave index function and $B_{\xi^\dagger}^\Omega$ denotes the Bregman distance with respect to Ω and to a subgradient $\xi^\dagger$ of Ω at an Ω minimizing solution $x^\dagger$ to (C.1). Note that Ω minimizing solutions always exist under Assumption C.1.

Variational source conditions (C.2) are known to imply convergence rates

$$B^{\Omega}_{\xi^\dagger}(x^\delta_\alpha, x^\dagger) = \mathcal{O}(\varphi(\delta)), \qquad \text{if } \delta \to 0,$$

with x^δ_α denoting the Tikhonov minimizers, if α is chosen properly in dependence on δ, see [SKHK12, Section 4.2] or [Fle12, Chapter 4].

The questions is, under which conditions variational source conditions are satisfied. This question has been discussed extensively in [Fle12, Part III] and has been solved there for linear mappings in Hilbert spaces, see also [AEdHS16, HW17]. The following Theorem provides an answer in case of nonlinear mappings in Banach spaces.

Theorem C.2. *Let Assumption C.1 be true and assume that there is only one solution $x^\dagger$ to* (C.1)*. Then there are a positive constant β with $\beta < 1$ and a concave index function φ such that the variational source condition* (C.2) *holds on $\mathcal{D}(F)$.*

Proof. To obtain a variational source condition (C.2) we use the concept of approximate variational source conditions introduced in [Fle12, Section 12.1.5]. That is, for fixed β with $\beta < 1$ we define a distance function $D_\beta : [0, \infty) \to [0, \infty)$ by

$$D_\beta(r) := \sup_{x \in \mathcal{D}(F)} \bigl(\beta\, B^{\Omega}_{\xi^\dagger}(x, x^\dagger) - \Omega(x) + \Omega(x^\dagger) - r\, \|F(x) - F(x^\dagger)\|\bigr).$$

We immediately see $0 \leq D_\beta(r) < \infty$ for all r. Further, D_β is convex, monotonically decreasing and continuous. The distance function D_β expresses the violation of a variational source condition with linear φ and allows to derive a variational source condition with some (nonlinear) φ if $D_\beta(r) \to 0$ for $r \to \infty$. To see this we estimate

$$\begin{aligned}
&\beta\, B^{\Omega}_{\xi^\dagger}(x, x^\dagger) - \Omega(x) + \Omega(x^\dagger) \\
&\quad = \inf_{r \geq 0} \bigl(\beta\, B^{\Omega}_{\xi^\dagger}(x, x^\dagger) - \Omega(x) + \Omega(x^\dagger) - r\, \|F(x) - F(x^\dagger)\| + r\, \|F(x) - F(x^\dagger)\|\bigr) \\
&\quad \leq \inf_{r \geq 0} \bigl(D_\beta(r) + r\, \|F(x) - F(x^\dagger)\|\bigr)
\end{aligned}$$

and show that

$$\varphi(t) := \inf_{r \geq 0} \bigl(D_\beta(r) + r\, t\bigr), \qquad t \geq 0,$$

defines a concave index function. Obviously, $0 \leq \varphi(t) < \infty$ and φ is monotonically increasing. Since φ is an infimum of affine functions, it is concave, upper semi-continuous and continuous on $(0, \infty)$. Monotonicity and upper semi-continuity imply continuity on $[0, \infty)$. The decay of D_β to zero yields $\varphi(0) = 0$ and by $D_\beta(0) > 0$ we see $\varphi(t) > 0$ for $t > 0$, that is, φ is strictly increasing in a neighborhood of zero.

Next, we show that for each r the supremum in the definition of $D_\beta(r)$ is attained at some x. Rearranging the terms in the supremum and flipping the sign we obtain a functional

$$x \mapsto (1 - \beta)\, \bigl(\Omega(x) - \Omega(x^\dagger)\bigr) + \beta\, \langle \xi^\dagger, x - x^\dagger \rangle + r\, \|F(x) - F(x^\dagger)\|. \tag{C.3}$$

Let $(x_n)_{n \in \mathbb{N}}$ be a sequence for which the values of the functional become arbitrarily close to the functional's infimum. Then this functional is bounded by some constant c

on this sequence. Weak compactness of the sublevel sets of Ω implies that there are a positive constant c_1 and a constant c_2 such that

$$\Omega(x) \geq c_1 \|x\| + c_2 \qquad \text{for all } x \text{ in } X,$$

see [Zăl02, Exercise 2.41 and pages 324–326]. With this observation we obtain

$$\begin{aligned} c &\geq (1-\beta)\big(\Omega(x_n) - \Omega(x^\dagger)\big) + \beta\,\langle \xi^\dagger, x_n - x^\dagger\rangle + r\,\|F(x_n) - F(x^\dagger)\| \\ &\geq (1-\beta)\big(\Omega(x_n) - \Omega(x^\dagger)\big) - \beta\,\|\xi^\dagger\|\,\big(\|x_n\| + \|x^\dagger\|\big) \\ &\geq (1-\beta)\big(\Omega(x_n) - \Omega(x^\dagger)\big) + \beta\,\|\xi^\dagger\|\,\frac{c_2 - \Omega(x_n)}{c_1} - \beta\,\|\xi^\dagger\|\,\|x^\dagger\| \\ &\geq \left(1 - \beta\left(1 + \frac{\|\xi^\dagger\|}{c_1}\right)\right)\Omega(x_n) - (1-\beta)\,\Omega(x^\dagger) + \beta\,\|\xi^\dagger\|\,\frac{c_2}{c_1} - \beta\,\|\xi^\dagger\|\,\|x^\dagger\|. \end{aligned}$$

Thus, if β is small enough, the sequence $(\Omega(x_n))_{n\in\mathbb{N}}$ is bounded. Now weak compactness of the sublevel sets of Ω implies that there is a weakly convergent subsequence and weak lower semi-continuity of (C.3) implies that the corresponding limit minimizes (C.3). Consequently, the supremum in the definition of D_β is attained for each r.

Now let $(r_n)_{n\in\mathbb{N}}$ be a sequence in $[0,\infty)$ with $r_n \to \infty$ and let $(x_n)_{n\in\mathbb{N}}$ be a sequence of corresponding maximizers in the definition of D_β. To complete the proof we have to show $D_\beta(r_n) \to 0$. With the same arguments as above we see

$$\begin{aligned} 0 &\leq D_\beta(r_n) \\ &\leq -\left(1 - \beta\left(1 + \frac{\|\xi^\dagger\|}{c_1}\right)\right)\Omega(x_n) + (1-\beta)\,\Omega(x^\dagger) - \beta\,\|\xi^\dagger\|\,\frac{c_2}{c_1} + \beta\,\|\xi^\dagger\|\,\|x^\dagger\|, \end{aligned}$$

which implies that $(\Omega(x_n))_{n\in\mathbb{N}}$ is bounded. Therefore $(x_n)_{\mathbb{N}}$ contains a weakly convergent subsequence with limit $\tilde{x}$. The subsequence again will be denoted by $(x_n)_{n\in\mathbb{N}}$. Since x_n realizes the supremum in the definition of $D_\beta(r_n)$ and because we have $D_\beta(r_n) \geq 0$, we see

$$r_n\,\|F(x_n) - F(x^\dagger)\| \leq \beta\,B_{\xi^\dagger}^{\Omega}(x_n, x^\dagger) - \Omega(x_n) + \Omega(x^\dagger)$$

and the right-hand side is bounded, again same arguments as above. This implies $F(x_n) \to F(x^\dagger)$ and together with the weak continuity of F we obtain $F(\tilde{x}) = F(x^\dagger)$, that is, $\tilde{x} = x^\dagger$ by assumption. Eventually,

$$\begin{aligned} 0 &\leq \liminf_{n\to\infty} D_\beta(r_n) \leq \limsup_{n\to\infty} D_\beta(r_n) = -\liminf_{n\to\infty}\big(-D_\beta(r_n)\big) \\ &= -\liminf_{n\to\infty}\Big((1-\beta)\big(\Omega(x_n) - \Omega(x^\dagger)\big) + \beta\,\langle \xi^\dagger, x_n - x^\dagger\rangle + r_n\,\|F(x_n) - F(x^\dagger)\|\Big) \\ &\leq -\liminf_{n\to\infty}\Big((1-\beta)\big(\Omega(x_n) - \Omega(x^\dagger)\big) + \beta\,\langle \xi^\dagger, x_n - x^\dagger\rangle\Big) \\ &\leq -\Big((1-\beta)\big(\Omega(x^\dagger) - \Omega(x^\dagger)\big) - \beta\,\langle \xi^\dagger, x^\dagger - x^\dagger\rangle\Big) = 0, \end{aligned}$$

which proves $D_\beta(r_n) \to 0$. □

The theorem shows that variational source conditions are widely applicable. Assumption C.1 is typically used to prove existence, stability and convergence of Tikhonov minimizers. Without additional assumptions, except for uniqueness of the solution, convergence rates can be obtained. The function φ is not given explicitly here, but with the above result we now know that there is a function φ. That is, variational source conditions are the right tool for convergence rate analysis in Banach spaces.

An analogous result has been obtained in [MH08] for general source conditions in Hilbert spaces: there is always an index function such that the corresponding general source condition is satisfied.

Bibliography

[ABHS16] S. W. Anzengruber, S. Bürger, B. Hofmann, and G. Steinmeyer. Variational regularization of complex deautoconvolution and phase retrieval in ultrashort laser pulse characterization. *Inverse Problems*, 32(3):035002 (27pp), 2016.

[AEdHS16] V. Albani, P. Elbau, M. V. de Hoop, and O. Scherzer. Optimal convergence rates results for linear inverse problems in Hilbert spaces. *Numerical Functional Analysis and Optimization*, 37(5):521–540, 2016.

[AHR13] S. W. Anzengruber, B. Hofmann, and R. Ramlau. On the interplay of basis smoothness and specific range conditions occurring in sparsity regularization. *Inverse Problems*, 29:125002 (21pp), 2013.

[Bau91] J. Baumeister. Deconvolution of appearance potential spectra. In R. Kleinman, R. Kress, and E. Martensen, editors, *Direct and Inverse Boundary Value Problems*, volume 37 of *Methoden und Verfahren der mathematischen Physik*, pages 1–13. Peter Lang Frankfurt am Main, 1991.

[BF15] S. Bürger and J. Flemming. Deautoconvolution: A new decomposition approach versus TIGRA and local regularization. *Journal of Inverse and III-posed Problems*, 23(3):231–243, 2015.

[BFH13] M. Burger, J. Flemming, and B. Hofmann. Convergence rates in ℓ^1-regularization if the sparsity assumption fails. *Inverse Problems*, 29:025013 (16pp), 2013.

[BFH16] S. Bürger, J. Flemming, and B. Hofmann. On complex-valued deautoconvolution of compactly supported functions with sparse Fourier representation. *Inverse Problems*, 32(10):104006 (12pp), 2016.

[BH10] R. I. Boţ and B. Hofmann. An extension of the variational inequality approach for nonlinear ill-posed problems. *Journal of Integral Equations and Applications*, 22(3):369–392, 2010.

[BH13] R. I. Boţ and B. Hofmann. The impact of a curious type of smoothness conditions on convergence rates in ℓ^1-regularization. *Eurasian Journal of Mathematical and Computer Applications*, 1(1):29–40, 2013.

[BH15] S. Bürger and B. Hofmann. About a deficit in low-order convergence rates on the example of autoconvolution. *Applicable Analysis*, 94(3):477–493, 2015.

[BL08] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14(5–6):813–837, 2008.

[BL09] K. Bredies and D. A. Lorenz. Regularization with non-convex separable constraints. *Inverse Problems*, 25(8):085011 (14pp), 2009.

[BM16] S. Bürger and P. Mathé. Discretized Lavrenti'ev regularization for the autoconvolution equation. *arXiv.org*, arXiv:1604.03275v1 [math.NA], 2016. http://arxiv.org/abs/1604.03275v1.

[BO04] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, 2004.

[BSK$^+$15] S. Birkholz, G. Steinmeyer, S. Koke, D. Gerth, S. Bürger, and B. Hofmann. Phase retrieval via regularization in self-diffraction based spectral interferometry. *Journal of the Optical Society of America B*, 32(5):983–992, 2015.

[Bür14] S. Bürger. About an autoconvolution problem arising in ultrashort laser pulse characterization. Preprint series of the Faculty of Mathematics 2014-16, Chemnitz University of Technology, Chemnitz, Germany, 2014.

[Bür16] S. Bürger. *Inverse Autoconvolution Problems with an Application in Laser Physics*. PhD thesis, Chemnitz University of Technology, Chemnitz, Germany, September 2016.

[CHZ17] D.-H. Chen, B. Hofmann, and J. Zou. Elastic-net regularization versus ℓ^1-regularization for linear inverse problems with quasi-sparse solutions. *Inverse Problems*, 33(1):015004 (17pp), 2017.

[CRT06] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[DDDM04] 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(11):1413–1457, 2004.

[Die84] J. Diestel. *Sequences and Series in Banach Spaces*, volume 92 of *Graduate Texts in Mathematics*. Springer, New York, 1984.

[Din41] L. L. Dines. On the mapping of quadratic forms. *Bulletin of the American Mathematical Society*, 47(6):494–498, 1941.

[DL08] Z. Dai and P. Lamm. Local regularization for the nonlinear inverse autoconvolution problem. *SIAM Journal on Numerical Analysis*, 46(2):832–868, 2008.

[EHN96] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Kluwer Academic Publishers, Dordrecht, 1996.

[EKN89] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5:523–540, 1989.

[FG17] J. Flemming and D. Gerth. Injectivity and weak*-to-weak continuity suffice for convergence rates in ℓ^1-regularization. 2017. Submitted to Journal of Inverse and Ill-Posed Problems.

[FH96] G. Fleischer and B. Hofmann. On inversion rates for the autoconvolution equation. *Inverse Problems*, 12(4):419–435, 1996.

[FH10] J. Flemming and B. Hofmann. A new approach to source conditions in regularization with general residual term. *Numerical Functional Analysis and Optimization*, 31(3):254–284, 2010.

[FH15] J. Flemming and M. Hegland. Convergence rates in ℓ^1-regularization when the basis is not smooth enough. *Applicable Analysis*, 94:464–476, 2015.

[FHV15] J. Flemming, B. Hofmann, and I. Veselić. On ℓ^1-regularization in light of nashed's ill-posedness concept. *Computational Methods in Applied Mathematics*, 15:279–289, 2015.

[FHV16] J. Flemming, B. Hofmann, and I. Veselić. A unified approach to convergence rates for ℓ^1-regularization and lacking sparsity. *Journal of Inverse and Ill-Posed Problems*, 24:139–148, 2016.

[Fle10] J. Flemming. Theory and examples of variational regularization with non-metric fitting functionals. *Journal of Inverse and Ill-posed Problems*, 18(6):677–699, 2010.

[Fle11] J. Flemming. *Generalized Tikhonov regularization. Basic theory and comprehensive results on convergence rates.* PhD thesis, Chemnitz University of Technology, Chemnitz, Germany, October 2011.

[Fle12] J. Flemming. *Generalized Tikhonov regularization and modern convergence rate theory in Banach spaces.* Shaker Verlag, Aachen, 2012.

[Fle14] J. Flemming. Regularization of autoconvolution and other ill-posed quadratic equations by decomposition. *Journal of Inverse and Ill-posed Problems*, 22(4):551–567, 2014.

[Fle16] J. Flemming. Convergence rates for ℓ^1-regularization without injectivity-type assumptions. *Inverse Problems*, 32(9):095001 (19pp), 2016.

[Gei09] J. Geißler. Studies on Convergence Rates for the Tikhonov Regularization with General Residual Functionals. Diploma thesis, Chemnitz University of Technology, Chemnitz, Germany, July 2009. In German.

[Ger11a] D. Gerth. Regularization of an autoconvolution problem occuring in measurements of ultra-short laser pulses. Diploma thesis, Chemnitz University of Technology, Chemnitz, Germany, 2011.

[GER11b] E. K. Gnang, A. Elgammal, and V. Retakh. A spectral theory for tensors. *arXiv.org*, arXiv:1008.2923 [math.SP], 2011. http://arxiv.org/abs/1008.2923.

[GH94] R. Gorenflo and B. Hofmann. On autoconvolution and regularization. *Inverse Problems*, 10(2):353–373, 1994.

[GHB$^+$14] D. Gerth, B. Hofmann, S. Birkholz, S. Koke, and G. Steinmeyer. Regularization of an autoconvolution problem in ultrashort laser pulse characterization. *Inverse Problems in Science and Engineering*, 22(2):245–266, 2014.

[GHS08] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with ℓ^q penalty term. *Inverse Problems*, 24:055020 (13pp), 2008.

[GHS11a] M. Grasmair, M. Haltmeier, and O. Scherzer. Necessary and sufficient conditions for linear convergence of ℓ^1-regularization. *Communications on Pure and Applied Mathematics*, 64:161–182, 2011.

[GHS11b] M. Grasmair, M. Haltmeier, and O. Scherzer. The residual method for regularizing ill-posed problems. *Applied Mathematics and Computation*, 218(6):2693–2710, 2011.

[Gra08] M. Grasmair. Well-posedness and convergence rates for sparse regularization with sublinear ℓ^q penalty term. Industrial Geometry report 74, University of Innsbruck, Innsbruck, Austria, August 2008.

[Gra09] M. Grasmair. Well-posedness and convergence rates for sparse regularization with sublinear ℓ^q penalty term. *Inverse Problems and Imaging*, 33:383–387, 2009.

[Gra10a] M. Grasmair. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Problems*, 26(11):115014 (16pp), 2010.

[Gra10b] M. Grasmair. Non-convex sparse regularisation. *Journal of Mathematical Analysis and Applications*, 365(1):19–28, 2010.

[GT63] S. Goldberg and E Thorp. On some open questions concerning strictly singular operators. *Proc. Amer. Math. Soc.*, 14:334–336, 1963.

[Hau19] F. Hausdorff. Der Wertvorrat einer Bilinearform. *Mathematische Zeitschrift*, 3:314–316, 1919.

[HH09] T. Hein and B. Hofmann. Approximate source conditions for nonlinear ill-posed problems—chances and limitations. *Inverse Problems*, 25(3):035033 (16pp), 2009.

[HKPS07] 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(3):987–1010, 2007.

[HM12] B. Hofmann and P. Mathé. Parameter choice in Banach space regularization under variational inequalities. *Inverse Problems*, 28:104006 (17pp), 2012.

[HMvW09] B. Hofmann, P. Mathé, and H. von Weizsäcker. Regularization in Hilbert space under unbounded operators and general source conditions. *Inverse Problems*, 25(11):115013 (15pp), 2009.

[Hof06] B. Hofmann. Approximate source conditions in Tikhonov–Phillips regularization and consequences for inverse problems with multiplication operators. *Mathematical Methods in the Applied Sciences*, 29(3):351–371, 2006.

[HS98] B. Hofmann and O. Scherzer. Local ill-posedness and source conditions of operator equations in Hilbert spaces. *Inverse Problems*, 14:1189–1206, 1998.

[HW13] T. Hohage and F. Werner. Iteratively regularized Newton methods with general data misfit functionals and applications to Poisson data. *Numerische Mathematik*, 123(4):745–779, 2013.

[HW17] T Hohage and F. Weidling. Characterizations of variational source conditions, converse results, and maxisets of spectral regularization methods. *SIAM Journal on Numerical Analysis*, 55(2):598–620, 2017.

[Jan97] J. Janno. On a regularization method for the autoconvolution equation. *Journal of Applied Mathematics and Mechanics*, 77(5):393–394, 1997.

[Jan00] J. Janno. Lavrent'ev regularization of ill-posed problems containing nonlinear near-to-monotone operators with application to autoconvolution equation. *Inverse Problems*, 16(2):333–348, 2000.

[KB09] T. G. Kolda and W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[KR12] E. Kopecká and S. Reich. A note on alternating projections in Hilbert space. *Journal of Fixed Point Theory and Applications*, 12(1):41–47, 2012.

[LF12] S. Lu and J. Flemming. Convergence rate analysis of tikhonov regularization for nonlinear ill-posed problems with noisy operators. *Inverse Problems*, 28(10):104003 (20 pp), 2012.

[Lor08] D. A. Lorenz. Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. *Journal of Inverse and Ill-Posed Problems*, 16:463–478, 2008.

[Meg98] R. E. Megginson. *An Introduction to Banach Space Theory*, volume 183 of *Graduate Texts in Mathematics*. Springer, New York, 1998.

[MH08] P. Mathé and B. Hofmann. How general are general source conditions? *Inverse Problems*, 24(1):015009 (5pp), 2008.

[Nas87] M. Z. Nashed. A new approach to classification and regularization of ill-posed operator equations. In *Inverse and Ill-posed Problems (Sankt Wolfgang, 1986), volume 4 of Notes and Reports in Mathematics in Science and Engineering*, pages 53–75. Academic Press, Boston, MA, 1987.

[NV76] M. Z. Nashed and G. F. Votruba. A unified operator theory of generalized inverses. In *Generalized Inverses and Applications*, pages 1–109. Academic Press, New York, 1976.

[Pös08] C. Pöschl. *Tikhonov Regularization with General Residual Term.* PhD thesis, University of Innsbruck, Innsbruck, Austria, October 2008. Corrected version.

[Ram03] R. Ramlau. TIGRA–an iterative algorithm for regularizing nonlinear ill-posed problems. *Inverse Problems*, 19(2):433–465, 2003.

[RR10] T. Ramlau and E. Resmerita. Convergence rates for regularization with sparsity constraints. *Electronic Transactions on Numerical Analysis*, 37:87–104, 2010.

[She13] J. L. Sheriff. *The convexity of quadratic maps and the controllability of coupled systems.* PhD thesis, Harvard University, Cambridge, Massachusetts, USA, February 2013.

[SKHK12] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski. *Regularization Methods in Banach Spaces*, volume 10 of *Radon Series on Computational and Applied Mathematics.* Walter de Gruyter, Berlin/Boston, 2012.

[Tak02] M. Takesaki. *Theory of Operator Algebra I*, volume 124 of *Encyclopaedia of Mathematical Sciences.* Springer, Berlin Heidelberg New York, 2002.

[Tit26] E. C. Titchmarsh. The zeros of certain integral functions. *Proceedings of the London Mathematical Society*, s2-25(1):283–302, 1926.

[Toe18] O. Toeplitz. Das algebraische Analogon zu einem Satze von Fejr. *Mathematische Zeitschrift*, 2:187–197, 1918.

[Wer15] F. Werner. On convergence rates for iteratively regularized Newton-type methods under a Lipschitz-type nonlinearity condition. *Journal of Inverse and Ill-Posed Problems*, 23(1):75–84, 2015.

[Xia14] Y. Xia. On local convexity of quadratic transformations. *Journal of the Operations Research Society of China*, 2(3):341–350, 2014.

[Zăl02] C. Zălinescu. *Convex Analysis in General Vector Spaces.* World Scientific, River Edge, 2002.

[Zei85] E. Zeidler. *Nonlinear Functional Analysis and its Applications III. Variational Methods and Optimization.* Springer, New York, Berlin, Heidelberg, 1985.

Theses

1. Nonlinear ill-posed equations $F(x) = y^\dagger$, $x \in X$, are hard to handle without additional knowledge about the type of nonlinearity. In several applications, especially in laser optics, nonlinear mappings with quadratic structure appear naturally. Thus, there is a need for deeper investigation of quadratic inverse problems.

2. Standard Tikhonov regularization can be applied to quadratic mappings, but numerical calculation of corresponding regularized solutions is difficult. Under strong assumptions the TIGRA method is able to find the regularized solutions.

3. The set of quadratic mappings between two Hilbert spaces forms a normed vector space. To some extent notions from the world of linear operators can be carried over to quadratic mappings. For example, the notion of quadratic isometries can be justified and utilized.

4. Every continous quadratic mapping F can be decomposed into a quadratic isometry Q and a linear operator A, that is, $F = A\,Q$. Based on such decompositions, stable solution methods for quadratic equations can be developed. Applying classical regularization techniques to the linear part A and then inverting the well-posed isometry Q yields competitive numerical results.

5. Classical convergence rates theory for nonlinear equations bases on source conditions. For quadratic mappings it turns out that this technique is not applicable. Variational source conditions are a recourse. In case of quadratic mappings a variational source condition is always satisfied if the involved index function is chosen properly.

6. If the solution of a quadratic equation is sparse with respect to some basis and the basis satisfies additional assumptions, then variational source conditions and, thus, convergence rates for regularized solutions can be proven. The rates then depend on the number of non-vanishing coefficients or, if this number is infinite, on the decay of the coefficients.

7. Sparsity promoting regularization methods, especially ℓ^1-regularization, are widely used in signal processing. They have a strong theoretic backing, but some questions had not been answered until recently. For example error estimates for ℓ^1-regularization with linear operators in case of non-sparse solutions or non-injective operators were lacking.

8. Operator equations with ℓ^1 as preimage space are always ill-posed and thus require regularization.

9. Source-type conditions, that is, Banach space source conditions and approximate source conditions, never are satisfied for operators defined on ℓ^1. Therefore, error estimates or convergence rates cannot be obtained this way.

10. Convergence rates in case of non-sparse solutions can be shown if the canonical basis of ℓ^1 is smooth with respect to the operator, that is, the basis elements belong to the range of the adjoint. This is the case in many applications, but there exist operators, with respect to which the canonical basis is not smooth.

11. In case of non-smooth bases weakened assumptions also lead to convergence rates. If the linear operator is injective and weak*-to-weak continuous, then such assumptions are always satisfied.

12. For non-injective operators source-type conditions can be formulated which imply convergence rates for sparse and non-sparse solutions. The error measure then is the distance between a point and the set of solutions. The developed source-type condition is quite technical, but can be verified for different concrete operators.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig angefertigt und keine anderen als die angegebenen Hilfsmittel verwendet habe. Sämtliche wissentlich verwendete Textausschnitte, Zitate oder Inhalte anderer Verfasser wurden ausdrücklich als solche gekennzeichnet.

Desweiteren erkläre ich hiermit, dass die vorliegende Arbeit in dieser oder ähnlicher Form an keiner anderen Stelle zum Zwecke eines Habilitationsverfahrens vorgelegt wurde. Weder früher noch zum jetzigen Zeitpunkt wurden Habilitationsverfahren bei anderen Stellen durch mich beantragt.

Chemnitz, den 15. Dezember 2017

Dr. Jens Flemming