Johann Radon Institute for
Computational and Applied Mathematics
Austrian Academy of Sciences (ÖAW)

On the minimization of a Tikhonov functional with a non-convex sparsity constraint

R. Ramlau, C. Zarzer

RICAM-Report 2009-05

On the minimization of a Tikhonov functional with a non-convex sparsity constraint

Ronny Ramlau[a], Clemens A Zarzer[b,*]

[a] *Johannes Kepler University Linz (JKU), Institute of Industrial Mathematics, Altenbergerstrasse 69, A-4040 Linz, Austria*
[b] *Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria*

Abstract

In this paper we present a numerical algorithm for the optimization of a Tikhonov functional with ℓ_p sparsity constraints and $p < 1$. Recently it was proven that the minimization of this functional provides a regularization method. We show that the idea used to obtain these theoretical results can also be utilized for a numerical approach. Particularly we exploit the technique of transforming the Tikhonov functional to a more viable one. In this regard we consider a surrogate functional approach and show that this technique can be applied straightforward. It is proven that at least a critical point of the transformed functional is obtained, which directly translates to the original functional. For a special case it is shown that a gradient based algorithm can be used to reconstruct the global minimizer of the transformed and the original functional, respectively. At the end we present numerical examples and provide numerical evidence for the theoretical results and desired sparsity promoting features of this method.

Keywords: sparsity, surrogate functional, inverse problem, regularization

[*] corresponding author
Email addresses: ronny.ramlau@jku.at (Ronny Ramlau), clemens.zarzer@ricam.oeaw.ac.at (Clemens A Zarzer)

1. Introduction

In this paper we consider a Tikhonov type regularization method for solving a (generally non-linear) ill-posed operator equation

$$\mathcal{F}(x) = y \tag{1}$$

with noisy measurements y^δ with $\|y^\delta - y\| \leq \delta$. Throughout the paper we assume that $\mathcal{F}$ maps between sequence spaces, i.e.

$$\mathcal{F} : \ell_p \to \ell_2 \ . \tag{2}$$

Please note that operator equations between suitable separable function spaces such as L^p, Sobolev and Besov spaces, i.e.

$$F : D(F) \subset X \to Y \ , \tag{3}$$

can be transformed to a sequence setting by using suitable basis or frames for $D(F)$ and $R(F)$: Indeed, if we assume that we are given some preassigned frames $\{\Phi_\lambda^i\}_{\lambda\in\Lambda_i, i=1,2}$, ($\Lambda_i$ countable index sets) for $D(F) \subset X$, $R(F) \subset Y$, with the associated frame operators T_1 and T_2 then the operator $\mathcal{F} := T_2 F T_1^*$ maps between sequence spaces.

We are particularly interested in *sparse* reconstructions, i.e. the reconstruction of sequences with only few non-zero elements. To this end, we want to minimize the Tikhonov functional

$$\begin{aligned} J_\alpha \ : \ \ell_p \ &\to \ \mathbb{R} \\ x \ &\mapsto \ \left\|\mathcal{F}(x) - y^\delta\right\|_2^2 + \alpha \ \|x\|_p^p \ , \end{aligned} \tag{4}$$

where $\alpha > 0$, $p \in (0, 1]$ and

$$\|x\|_p^p = \sum_k |x_k|^p \ , \tag{5}$$

is the (quasi-) norm of ℓ_p. The main aim of our paper is the development of an iterative algorithm for the minimization of (4), which is due to the non-convexity of the quasi-norm and the non-linearity of $\mathcal{F}$ a non-trivial task.

The reconstruction of the sparsest solution of an underdetermined system has already a long history, in particular in signal processing, and more recently, in compressive sensing. Usually the problem is formulated as

$$\tilde{x} := \arg\min_{y=\Phi x} \|x\|_1 \tag{6}$$

where $y \in \mathbb{R}^m$ is given and $\Phi \in \mathbb{R}^{m,n}$ is a rank deficient matrix (i.e. $m < n$), see [1, 2]. Please note that here the minimization of the ℓ_1-norm is used for the reconstruction of the sparsest solution of the equation $\Phi x = y$. Indeed, under certain assumptions on the matrix Φ, it can be shown that, if there is a sparse solution, (6) really recovers it [3, 4, 5, 6]. Moreover, Gribonval and Nielsen [7] showed that for special cases the minimization of (6) also recovers ℓ_p minimizers with $0 < p < 1$. In this sense it might seem that nothing is gained by considering ℓ_p minimization with $0 < p < 1$ instead of ℓ_1 minimization, or equivalently, using an ℓ_p penalty with $0 < p < 1$ in (4). However, we have to keep in mind that we are considering a different setting as the above cited papers. First of all, we are working in an infinite dimensional setting, whereas the above mentioned Φ is a finite dimensional matrix. Additionally, properties that guarantee the above cited results as the so-called Restricted Isometry Property, which was introduced by Candes and Tao [8, 4], or the Null Space Property [9, 10] are not likely to hold even for linear infinite dimensional ill-posed problems where, e.g., the eigenvalues of the operator converge to zero, not to speak of non-linear operators. Recently, there has also been numerical evidence from a non-linear parameter identification problem for chemical reaction systems that an ℓ_1 penalty in (4) failed to reconstruct a desired sparse parameter whereas stronger ℓ_p penalties with $0 < p < 1$ achieved sparse reconstructions [11]. In the mentioned paper, the intention of the authors was the reconstruction of reduced chemical networks (represented by a sparse parameter) from chemical measurements. Therefore, we conclude that the use of the stronger ℓ_p penalties might be necessary in infinite dimensional ill-posed problems if one wants a sparse reconstruction. In particular, algorithms for the minimization of (4) are needed.

There has been an increased interest in the investigation of the Tikhonov functional with sparsity constraints. First results on that matter were presented by Daubechies, Defriese and De Mol [12]. The authors were in particular interested in solving linear operator equations. As constraint in (4) they used a Besov semi-norm, which can be equivalently expressed by a weighted ℓ_p-norm of the wavelet coefficients of the functions with $p \geq 1$. In particular the paper focuses on the analysis of a surrogate functional approach for the minimization of (4) with $p \geq 1$. It was shown that the proposed iterative method converges towards a minimizer of the Tikhonov functional under consideration. Additionally, the authors proposed a rule for the choice of the regularization parameter that guarantees the convergence of the minimizer x_α^δ of the Tikhonov functional to the solution as the data error δ converges

to zero. Subsequently, many results on the regularization properties of the Tikhonov functional with sparsity constraints and $p \geq 1$ as well as on its minimization were published. In [13, 14] the surrogate functional approach for the minimization of the Tikhonov functional was generalized to non-linear operator equations and in [15, 16] to multi-channel data, whereas in [17, 18] a conditional gradient method and in [19] a semi-smooth Newton method were proposed for the minimization. Further results on the topic of minimization and the respective algorithms can be found in [20, 21, 22]. The regularization properties with respect to different topologies and parameter choice rules were considered in [14, 15, 23, 24, 25, 26]. Please note again that the above cited results only consider the case $p \geq 1$. For the case $p < 1$, a first regularization result for some types of linear operators was presented in [26]. In [27] and [28] the authors recently presented general results on the regularization properties of the Tikhonov functional with a non-linear operator and $0 < p < 1$. Concerning the minimization of (4) with $0 < p < 1$, to our knowledge no results are available in the infinite dimensional setting. In the finite dimensional setting, Daubechies et. al [10] presented an iteratively re-weighted least squares method for the solution of (6) that achieved local super-linear convergence. However, these results do not carry over to the minimization of (4), as the assumptions made in [10] (e.g., finite dimension, null space property) do not hold for general inverse problems. Other closely related results for the finite dimensional case can be found in [29, 30]. For a more general overview on sparse recovery we refer to [31].

In this paper, we present two algorithms for the minimization of (4) which are based on the surrogate functional algorithm [12, 13, 14, 23] and the TIGRA algorithm [32, 33]. Based on a technique presented in [28] and based on methods initially developed in [34], the functional (4) is non-linearly transformed by an operator $\mathcal{N}_{p,q}$ to a new Tikhonov functional, now with an ℓ_q-norm as penalty and $1 < q \leq 2$. Due to the non-linear transformation, the new Tikhonov functional involves a non-linear operator, even when the original problem is linear. Provided the operator $\mathcal{F}$ fulfills some properties, it is shown that the surrogate functional approach at least reconstructs a critical point of the transformed functional. Moreover, the minimizers of the original and the transformed functional are connected by the transformation $\mathcal{N}_{p,q}$, and thus we can obtain a minimizer for the original functional. For the special case $q = 2$ we show that the TIGRA algorithm reconstructs a global minimizer if the solution fulfills a smoothness condition. For the case $\mathcal{F} = \mathcal{I}$, where $\mathcal{I}$ denotes the identity, we show that the smoothness

condition is always fulfilled for sparse solutions, whereas for $\mathcal{F} = \mathcal{A}$ with linear $\mathcal{A}$ the finite basis injectivity (FBI) property is needed additionally. The paper is organized as follows: In Section 2 we recall some results from [28] and introduce the transformation operator $\mathcal{N}_{p,q}$. Section 3 is concerned with some analytical properties of $\mathcal{N}_{p,q}$, whereas Section 4 investigates the operator $\mathcal{F} \circ \mathcal{N}_{p,q}$. In Section 5 we use the surrogate functional approach for the minimization of the transformed functional, and in Section 6 we introduce the TIGRA method for the reconstruction of a global minimizer. Finally we present in Section 7 numerical results for the reconstruction of a function from its convolution data and present an application from Physical Chemistry with a highly non-linear operator. Both examples confirm our analytical findings and support the proposed enhanced sparsity promoting feature of the considered regularization technique.

Whenever it is appropriate, we omit the subscripts for norms, sequences, dual pairings and so on. If not denoted otherwise, we consider the particular notions in terms of Hilbert space ℓ_2 and the respective topology $\|\cdot\|_2$. Furthermore we would like to mention that the subscript k shall indicate the individual components of an element of a sequence. The subscripts l and n are used for sequences of elements in the respective space or their components, e.g. $x_n = \{x_{n,k}\}_{k\in\mathbb{N}}$. Whenever unclear or referring to an entire sequence we use $\{\cdot\}$ to denote the component-wise view. Iterates in terms of the considered algorithms are denoted with superscript l and n.

2. A transformation of the Tikhonov functional

In [28] it was shown that (4) provides a regularization method under classical assumptions on the operator. The key idea was to transform the Tikhonov type functional by means of a superposition operator into a standard formulation. Below we give a brief summary on some results presented in [28] and consequently show additional properties of the transformation operator.

Definition 2.1. We denote by $\eta_{p,q}$ the function given by

$$\begin{aligned} \eta_{p,q} \;:\; \mathbb{R} &\to \mathbb{R} \\ r &\mapsto \operatorname{sign}(r)\,|r|^{\frac{q}{p}}\,, \end{aligned} \tag{7}$$

for $0 < p \leq 1$ and $1 \leq q \leq 2$.

Definition 2.2. We denote by $\mathcal{N}_{p,q}$ the superposition operator given by

$$\mathcal{N}_{p,q} \ : \ x \ \mapsto \ \{\eta_{p,q}(x_k)\}_{k\in\mathbb{N}} \ , \tag{8}$$

where $x \in \ell_q$, $0 < p \leq 1$ and $1 \leq q \leq 2$.

Proposition 2.3. *For all $0 < p \leq 1$, $1 \leq q \leq 2$, $x \in \ell_q$ and $\mathcal{N}_{p,q}$ as in Definition 2.2 holds $\mathcal{N}_{p,q}(x) \in \ell_p$, and the operator $\mathcal{N}_{p,q} : \ell_q \to \ell_p$ is bounded, continuous and bijective.*

Using the concatenation operator:

$$\begin{aligned} \mathcal{G} \ : \ \ell_q \ &\to \ \ell_2 \\ x \ &\mapsto \ \mathcal{F} \circ \mathcal{N}_{p,q}(x) \ , \end{aligned} \tag{9}$$

one obtains the following two equivalent minimization problems.

Problem 1. *Let y^δ be an approximation of the right hand side of (1) with $\left\|y - y^\delta\right\| \leq \delta$ and $\alpha > 0$, then minimize:*

$$\left\|\mathcal{F}(x_s) - y^\delta\right\|_2^2 + \alpha \ \|x_s\|_p^p \ , \tag{10}$$

subject to $x_s \in \ell_p$, for $0 < p \leq 1$.

Problem 2. *Let y^δ be an approximation of the right hand side of (1) with $\left\|y - y^\delta\right\| \leq \delta$ and $\alpha > 0$. Determine $x_s = \mathcal{N}_{p,q}(x)$, where x minimizes*

$$\left\|\mathcal{G}(x) - y^\delta\right\|_2^2 + \alpha \ \|x\|_q^q \ , \tag{11}$$

subject to $x \in \ell_q$ and $0 < p \leq 1$, $1 \leq q \leq 2$.

Proposition 2.4. *Problem 1 and Problem 2 are equivalent.*

[28] provides classical results on the existence of minimizers, stability and convergence for the particular Tikhonov approach considered here. These results are obtained via the observation of weak (sequential) continuity of the transformation operator.

3. Properties of the operator $\mathcal{N}_{p,q}$

Let us start with an analysis of the operator $\mathcal{N}_{p,q}$. The following proposition was given in [28]. We restate the proof as it is used afterwards.

Proposition 3.1. *The operator $\mathcal{N}_{p,q} : \ell_q \to \ell_q$ is weakly (sequentially) continuous for $0 < p \leq 1$ and $1 < q \leq 2$, i.e.*

$$x_n \overset{\ell_q}{\rightharpoonup} x \Longrightarrow \mathcal{N}_{p,q}(x_n) \overset{\ell_q}{\rightharpoonup} \mathcal{N}_{p,q}(x) \ . \tag{12}$$

Here $\overset{X}{\rightharpoonup}$ denotes weak convergence wrt. to the space X.

Proof. We set $r = q/p + 1$ and observe $r \geq 2$. A sequence in ℓ_q is weakly convergent if and only if the coefficients converge and the sequence is bounded in norm. Thus we conclude from the weak convergence of x_n that $\|x_n\|_q \leq C$ and $x_{n,k} \to x_k$. As $r \geq q$, we have a continuous embedding of ℓ_r into ℓ_q, i.e.

$$\|x_n\|_r \leq \|x_n\|_q \leq C \ ,$$

which shows that also

$$x_n \overset{\ell_r}{\rightharpoonup} x$$

holds. The operator $(\mathcal{N}_{p,q}(x))_k = \operatorname{sgn}(x_k)|x_k|^{r-1}$ is the derivative of the function

$$f(x) = r^{-1} \cdot \|x\|_r^r \ ,$$

or, in other words, $\mathcal{N}_{p,q}(x)$ is the *duality mapping* on ℓ_r with respect to the weight function

$$\varphi(t) = t^{r-1}$$

(for more details on duality mappings we refer to [35]). Now it is a well known result that every duality mapping on ℓ_r is weakly (sequentially) continuous, see, e.g. [35], Prop. 4.14. Thus we obtain

$$x_n \overset{\ell_r}{\rightharpoonup} x \Longrightarrow \mathcal{N}_{p,q}(x_n) \overset{\ell_r}{\rightharpoonup} \mathcal{N}_{p,q}(x) \ .$$

Again, as $\mathcal{N}_{p,q}(x_n)$ is weakly convergent, we have $\{\mathcal{N}_{p,q}(x_n)\}_k \to \{\mathcal{N}_{p,q}(x)\}_k$. For for $p \leq 1,\ q \geq 1$ holds $q \leq q^2/p$ and thus we have $\|x\|_{q^2/p} \leq \|x\|_q$. It follows

$$\|\mathcal{N}_{p,q}(x_n)\|_q^q = \sum_k |x_{n,k}|^{q^2/p} = \|x_n\|_{q^2/p}^{q^2/p} \leq \|x_n\|_q^{q^2/p} \leq C^{q^2/p} \ ,$$

i.e. $\mathcal{N}_{p,q}(x_n)$ is also uniformly bounded with respect to ℓ_q and thus also weakly convergent. □

In the following proposition we show that the same result holds with respect to weak ℓ_2-convergence.

Proposition 3.2. *The operator $\mathcal{N}_{p,q} : \ell_2 \to \ell_2$ is weakly (sequentially) continuous w.r.t. ℓ_2 for $0 < p \leq 1$ and $1 < q \leq 2$, i.e.*

$$x_n \xrightarrow{\ell_2} x \Longrightarrow \mathcal{N}_{p,q}(x_n) \xrightarrow{\ell_2} \mathcal{N}_{p,q}(x) \ . \tag{13}$$

Proof. First we have for $x \in \ell_2$ with $2q/p \geq 2$

$$\|\mathcal{N}_{p,q}(x)\|_2^2 = \sum_k |x_k|^{2q/p} = \|x\|_{2q/p}^{2q/p} \leq \|x\|_2^{2q/p} < \infty \ ,$$

i.e. $\mathcal{N}_{p,q}(x) \in \ell_2$ for $x \in \ell_2$. Setting again $r = q/p + 1$, the remainder of the proof follows the lines of the previous one, with $\|\cdot\|_q$ replaced by $\|\cdot\|_2$. □

Next, we want to investigate the Fréchet derivative of $\mathcal{N}_{p,q}$. Beforehand we need the following Lemma.

Lemma 3.3. *The map $x \mapsto \operatorname{sgn}(x)\,|x|^\alpha$, $x \in \mathbb{R}$, is Hölder continuous with exponent α, for $\alpha \in (0, 1]$. Moreover we have locally for $\alpha > 1$ and globally for $\alpha \in (0, 1]$:*

$$|\operatorname{sgn}(x)\,|x|^\alpha - \operatorname{sgn}(y)\,|y|^\alpha| \leq \kappa\ |x - y|^\beta \ , \tag{14}$$

where $\beta = \min(\alpha, 1)$.

Proof. As the problem is symmetric with respect to x and y, we assume w.l.o.g. $|x| \geq |y|$ and $|y| > 0$ as (14) immediately holds for $y = 0$. Let $\gamma \in \mathbb{R}_0^+$ s.t.: $\gamma\,|y| = |x|$. For $\gamma \in [1, \infty)$ and $\alpha \in (0, 1]$ we have

$$(\gamma^\alpha - 1) \leq (\gamma - 1)^\alpha \ , \tag{15}$$

which can be obtained by comparing the derivatives of $(\gamma^\alpha - 1)$ and $(\gamma - 1)^\alpha$ for $\gamma > 1$, and by the fact that we have equality for $\gamma = 1$. Moreover we have for $\gamma \in [0, \infty)$ and $\alpha \in (0, 1]$

$$(\gamma^\alpha + 1) \leq 2\,(\gamma + 1)^\alpha \ . \tag{16}$$

As it is crucial that the constant in Inequality (16) is independent of γ, we now give a proof of the factor 2. The ratio

$$\frac{(\gamma^\alpha + 1)}{(\gamma + 1)^\alpha}$$

is monotonously increasing for $\gamma \in (0,1]$ and monotonously decreasing for $\gamma \in (1,\infty)$, which can be easily seen from its derivative. Hence the maximum is attained at $\gamma = 1$ and given by $2^{1-\alpha}$, which yields

$$\frac{(\gamma^\alpha + 1)}{(\gamma+1)^\alpha} \leq 2^{1-\alpha} \leq 2 \; .$$

Consequently we can conclude in the case of $x \cdot y > 0$ (i.e. $\operatorname{sgn}(x) = \operatorname{sgn}(y)$) that

$$\begin{aligned} |\operatorname{sgn}(x)\,|x|^\alpha - \operatorname{sgn}(y)\,|y|^\alpha| &= |\gamma^\alpha\,|y|^\alpha - |y|^\alpha| = |(\gamma^\alpha - 1)|y|^\alpha| \\ &\overset{(15)}{\leq} |(\gamma-1)^\alpha\,|y|^\alpha| = |x-y|^\alpha \; , \end{aligned}$$

and for $x \cdot y < 0$ we have:

$$\begin{aligned} |\operatorname{sgn}(x)\,|x|^\alpha - \operatorname{sgn}(y)\,|y|^\alpha| &= |\gamma^\alpha\,|y|^\alpha + |y|^\alpha| = |(\gamma^\alpha + 1)|y|^\alpha| \\ &\overset{(16)}{\leq} 2\;|(\gamma+1)^\alpha\,|y|^\alpha| = 2\;|x-y|^\alpha \; . \end{aligned}$$

In the case of $\alpha > 1$ (14) holds w.r.t. $\beta = 1$, which can be proven by the mean value theorem. For $\alpha > 1$ the function $f : x \mapsto \operatorname{sgn}(x)\,|x|^\alpha$ is differentiable and its derivative is bounded on any interval I. Hence, (14) holds for $|f'(\xi)| \leq \kappa$, $\xi \in I$, proving the local Lipschitz continuity. □

Remark 3.4. In the following Lemma 3.3 is used to uniformly estimate the remainder of a Taylor series. As shown in the proof, this immediately holds for $\alpha \in (0,1]$. In the case of the Lipschitz estimate this is valid only locally. However as all sequences in Proposition 3.5 are bounded and we are only interested in a local estimate, Lemma 3.3 can be applied directly.

Proposition 3.5. *The Fréchet derivative of* $\mathcal{N}_{p,q} : \ell_q \to \ell_q$, $0 < p \leq 1$, $1 < q \leq 2$ *is given by the sequence*

$$\mathcal{N}'_{p,q}(x)h = \left\{ \frac{q}{p} |x_k|^{(q-p)/p} \cdot h_k \right\}_{k\in\mathbb{N}} \; . \tag{17}$$

Proof. Let $w := \min\left(\frac{q}{p} - 1, 1\right) > 0$. The derivative of the function $\eta_{p,q}(t) = |t|^{q/p} \operatorname{sgn}(t)$ is given by $\eta'_{p,q}(t) = \frac{q}{p}|t|^{(q-p)/p}$ and we have

$$\eta_{p,q}(t+\tau) - \eta_{p,q}(t) - \eta'_{p,q}(t)\,\tau \;\; := r(t,\tau) \; . \tag{18}$$

Integration of the following expression yields (18):

$$r(t,\tau) = \int_t^{t+\tau} \frac{q}{p}\frac{q-p}{p}\,(t+\tau-s)\,\operatorname{sgn}(s)\,|s|^{\frac{q}{p}-2}\,\mathrm{d}s\ .$$

Given the considered ranges of p and q, $\eta_{p,q}$ is not twice differentiable. On this account we derive the following estimate, using the mean value theorem:

$$\begin{aligned}\left|\int_t^{t+\tau} \frac{q}{p}\frac{q-p}{p}\,(t+\tau-s)\,\operatorname{sgn}(s)\,|s|^{\frac{q}{p}-2}\,\mathrm{d}s\right| &= \\ &= \left|\left[\frac{q}{p}\,(t+\tau-s)\,|s|^{q/p-1}\right]_t^{t+\tau} + \int_t^{t+\tau}\frac{q}{p}\,|t|^{q/p-1}\,\mathrm{d}s\right| \\ &= \left|\frac{q}{p}\,\tau\,\left(|\xi|^{q/p-1}-|t|^{q/p-1}\right)\right| \overset{(14)}{\leq}\ \kappa\,\frac{q}{p}\,|\tau|^{w+1}\ ,\end{aligned}$$

with $\xi \in (t, t+\tau)$ and by using Lemma 3.3 with $\alpha = q/p-1$, where κ is independent of τ (see Remark 3.4). Hence we may write for $\|h\| = \|\{h_k\}\|$ sufficiently small

$$\begin{aligned}\left\|\mathcal{N}_{p,q}(x+h)-\mathcal{N}_{p,q}(x)-\mathcal{N}'_{p,q}(x)h\right\|_q^q = \|\{r(x_k,h_k)\}\|_q^q &= \sum_k |r(x_k,h_k)|^q \\ &\leq \sum_k \left(\frac{\kappa\,q}{p}\right)^q |h_k|^{q(w+1)} \\ &\leq \left(\frac{\kappa\,q}{p}\right)^q \max\left(\{|h_k|^{qw}\}\right)\,\sum_k |h_k|^q\ .\end{aligned}$$

Hence we conclude $\|\{r(x_k,h_k)\}\|_q \,/\, \|h\|_q \to 0$ for $\|h\|_q \to 0$ and obtain for the derivative $\mathcal{N}'_{p,q}(x)h = \left\{\eta'_{p,q}(x_k)h_k\right\}_{k\in\mathbb{N}}$. □

Remark 3.6. Please note that the result of Proposition 3.5 also holds in the case of the operator $\mathcal{N}_{p,q} : \ell_2 \to \ell_2$, as one can immediately see from the proof.

Lemma 3.7. *The operator $\mathcal{N}'_{p,q}(x)$ is self-adjoint with respect to ℓ_2.*

Proof. We have $\langle \mathcal{N}'_{p,q}(x)h, z\rangle = \frac{q}{p}\sum |x_k|^{(q-p)/p} h_k z_k = \langle h, \mathcal{N}'_{p,q}(x)z\rangle$. □

Please note that the Fréchet derivative of the operator $\mathcal{N}_{p,q}$ and its adjoint can be understood as (infinite dimensional) diagonal matrices, that is

$$\mathcal{N}'_{p,q}(x) = \operatorname{diag}\left(\left\{\frac{q}{p}|x_k|^{(q-p)/p}\right\}_{k\in\mathbb{N}}\right) ,$$

and $\mathcal{N}'_{p,q}(x)h$ is then a matrix-vector multiplication.

4. Properties of the concatenation operator $\mathcal{G}$

The convergence of the surrogate functional approach, which will be applied to the transformed Tikhonov functional (11), relies mainly on some mapping properties of the operator $\mathcal{G} = \mathcal{F}\circ\mathcal{N}_{p,q}$. In the following, we assume that the operator $\mathcal{F}$ is Fréchet differentiable and $\mathcal{F}, \mathcal{F}'$ fulfill the following conditions:

$$x_n \rightharpoonup x \Longrightarrow \quad \mathcal{F}(x_n) \to \mathcal{F}(x) \text{ for } n \to \infty \tag{19}$$

$$x_n \rightharpoonup x \Longrightarrow \quad \mathcal{F}'(x_n)^* z \to \mathcal{F}'(x)^* z \text{ for } n \to \infty \text{ and all } z \tag{20}$$

$$\|\mathcal{F}'(x) - \mathcal{F}'(x')\| \le L\|x - x'\| \ \text{ locally .} \tag{21}$$

Convergence and weak convergence in (19),(20) has to be understood with respect to ℓ_2. The main goal of this section is to show that the concatenation operator $\mathcal{G}$ is Fréchet differentiable and that this operator also fulfills the conditions given above. First we obtain

Proposition 4.1. *Let $\mathcal{F} : \ell_q \to \ell_2$ be strongly continuous w.r.t. ℓ_q, i.e.*

$$x_n \overset{\ell_q}{\rightharpoonup} x \Longrightarrow \mathcal{F}(x_n) \overset{\ell_q}{\to} \mathcal{F}(x). \tag{22}$$

Then $\mathcal{F}\circ\mathcal{N}_{p,q}$ is also strongly continuous w.r.t. ℓ_q. If $\mathcal{F} : \ell_2 \to \ell_2$ is strongly continuous w.r.t. ℓ_2, then $\mathcal{F}\circ\mathcal{N}_{p,q}$ is also strongly continuous w.r.t. ℓ_2.

Proof. If $x_n \overset{\ell_q}{\rightharpoonup} x$, then, by Proposition 3.1, also $\mathcal{N}_{p,q}(x_n) \overset{\ell_q}{\rightharpoonup} \mathcal{N}_{p,q}(x)$, and due to the strong continuity of $\mathcal{F}$ follows $\mathcal{F}(\mathcal{N}_{p,q}(x_n)) \to \mathcal{F}(\mathcal{N}_{p,q}(x))$. The second part of the proposition follows in the same way by Proposition 3.2. □

By the chain rule we immediately obtain the following result.

Lemma 4.2. *Let $\mathcal{F} : \ell_q \to \ell_2$ be Fréchet differentiable. Then*

$$(\mathcal{F} \circ \mathcal{N}_{p,q})'(x) = \mathcal{F}'(\mathcal{N}_{p,q}(x)) \cdot \mathcal{N}'_{p,q}(x) \ , \tag{23}$$

where the multiplication has to be understood as a matrix product. The adjoint (with respect to ℓ_2) of the Fréchet derivative is given by

$$((\mathcal{F} \circ \mathcal{N}_{p,q})'(x))^* = \mathcal{N}'_{p,q}(x) \cdot \mathcal{F}'(\mathcal{N}_{p,q}(x))^* \ . \tag{24}$$

Proof. Equation (23) is simply the chain rule. For the adjoint of the Fréchet derivative we get:

$$\begin{aligned} \langle ((\mathcal{F} \circ \mathcal{N}_{p,q})'(x))\, u, z\rangle &= \langle \mathcal{F}'(\mathcal{N}_{p,q}(x)) \cdot \mathcal{N}'_{p,q}(x) \cdot u, z\rangle \\ &= \langle \mathcal{N}'_{p,q}(x) \cdot u, \mathcal{F}'(\mathcal{N}_{p,q}(x))^* \cdot z\rangle \\ &= \langle u, \mathcal{N}'_{p,q}(x) \cdot \mathcal{F}'(\mathcal{N}_{p,q}(x))^* z\rangle \ , \end{aligned}$$

as $\mathcal{N}'_{p,q}(x)$ is self-adjoint. □

We further need the following result.

Lemma 4.3. *Let $\mathcal{B} : \ell_q \to \ell_q$ be a (infinite dimensional) diagonal matrix with diagonal elements $b = \{b_k\}$. Then*

$$\|\mathcal{B}\|_\infty \le \|b\|_q \tag{25}$$

Proof. The assertion follows by

$$\|\mathcal{B}\|_\infty^q = \sup_{\|u\|\le 1} \|\mathcal{B}u\|_q^q = \sup_{\|u\|\le 1} \sum_k |b_k \cdot u_k|^q \le \sum_k |b_k|^q \ .$$

□

Hence we may identify the operator $\mathcal{N}'_{p,q}(x_n)$ with its sequence and vice versa. Now we can conclude the first required property.

Proposition 4.4. *Let $x_n \rightharpoonup x$ with respect to ℓ_2, $z \in \ell_2$ and let q and p be such that $q \ge 2p$. Assume that*

$$(\mathcal{F}'(x_n))^* z \to (\mathcal{F}'(x))^* z \tag{26}$$

w.r.t. ℓ_2 holds for any weakly convergent sequence $x_n \to x$. Then we have as well

$$((\mathcal{F} \circ \mathcal{N}_{p,q})'(x_n))^*\, z \to ((\mathcal{F} \circ \mathcal{N}_{p,q})'(x))^*\, z \ , \tag{27}$$

w.r.t. ℓ_2.

Proof. As $x_n \xrightarrow{\ell_2} x$, we have in particular $x_{n,k} \to x_k$ for fixed k. The sequence $\mathcal{N}'_{p,q}(x_n)$ is given element-wise by

$$\frac{q}{p}|x_{n,k}|^{(q-p)/p} \to \frac{q}{p}|x_k|^{(q-p)/p} ,$$

and thus the coefficients of $\mathcal{N}'_{p,q}(x_n)$ converge to the coefficients of $\mathcal{N}'_{p,q}(x)$. In order to show weak convergence of the sequences, it remains to show that $\{\frac{q}{p}|x_{n,k}|^{(q-p)/p}\}$ stays uniformly bounded: We have

$$\|\mathcal{N}'_{p,q}(x_n)\|_2^2 = \left(\frac{q}{p}\right)^2 \sum \left(|x_{n,k}|^{(q-p)/p}\right)^2 .$$

As $q \geq 2p$ and $\|x\|_r \leq \|x\|_s$ for $s \leq r$ we conclude with $r = 2(q-p)/p \geq 2$

$$\|\mathcal{N}'_{p,q}(x_n)\|_2^2 = \left(\frac{q}{p}\right)^2 \|x_n\|_r^r \leq \left(\frac{q}{p}\right)^2 \|x_n\|_2^r \leq C , \tag{28}$$

as weakly convergent sequences are uniformly bounded. Thus we conclude

$$\mathcal{N}'_{p,q}(x_n) \rightharpoonup \mathcal{N}'_{p,q}(x) .$$

With the same arguments we get for fixed z

$$\mathcal{N}'_{p,q}(x_n)z \rightharpoonup \mathcal{N}'_{p,q}(x)z .$$

The convergence of this sequence holds also in the strong sense. For this, it is sufficient to show that $\lim_{n\to\infty} \|\mathcal{N}'_{p,q}(x_n)z\| = \|\mathcal{N}'_{p,q}(x)z\|$ holds: As x_n is weakly convergent, the sequence is also uniformly bounded, i.e. $\|x_n\|_{\ell_2} \leq \tilde{C}$, thus $|x_{n,k}| \leq \tilde{C}$ and hence $|x_{n,k}|^{2(q-p)/p} \cdot z_k^2 \leq \tilde{C}^{2(q-p)/p} z_k^2$. We observe:

$$\left(\frac{q}{p}\right)^2 \sum_k |x_{n,k}|^{\frac{2(q-p)}{p}} \cdot z_k^2 \leq \left(\frac{q}{p}\right)^2 \tilde{C}^{\frac{2(q-p)}{p}} \sum_k z_k^2 = \left(\frac{q}{p}\right)^2 \tilde{C}^{\frac{2(q-p)}{p}} \|z\|_2^2 < \infty .$$

Therefore, by the dominated convergence theorem, we can interchange limit and summation, i.e.

$$\begin{aligned}
\lim_{n\to\infty} \|\mathcal{N}'_{p,q}(x_n)z\|_2^2 &= \lim_{n\to\infty} \left(\frac{q}{p}\right)^2 \sum_k |x_{n,k}|^{2(q-p)/p} \cdot z_k^2 \\
&= \left(\frac{q}{p}\right)^2 \sum_k \lim_{n\to\infty} |x_{n,k}|^{2(q-p)/p} \cdot z_k^2 \\
&= \left(\frac{q}{p}\right)^2 \sum_k |x_k|^{2(q-p)/p} \cdot z_k^2 = \left(\frac{q}{p}\right)^2 \|\mathcal{N}'_{p,q}(x)z\|_2^2 ,
\end{aligned}$$

and thus

$$\mathcal{N}'_{p,q}(x_n)z \xrightarrow{\ell_2} \mathcal{N}'_{p,q}(x)z \ . \tag{29}$$

We further conclude

$$\begin{aligned}
&\| \left((\mathcal{F}\circ\mathcal{N}_{p,q})'(x_n)\right)^* z - \left((\mathcal{F}\circ\mathcal{N}_{p,q})'(x)\right)^* z\|_2 \\
&\qquad = \|\mathcal{N}'_{p,q}(x_n)\mathcal{F}'(\mathcal{N}_{p,q}(x_n))^* z - \mathcal{N}'_{p,q}(x)\mathcal{F}'(\mathcal{N}_{p,q}(x))^* z\|_2 \\
&\qquad \le \underbrace{\|\mathcal{N}'_{p,q}(x_n)\mathcal{F}'(\mathcal{N}_{p,q}(x_n))^* z - \mathcal{N}'_{p,q}(x_n)\mathcal{F}'(\mathcal{N}_{p,q}(x))^* z\|_2}_{D_1} \\
&\qquad\qquad + \underbrace{\|\mathcal{N}'_{p,q}(x_n)\mathcal{F}'(\mathcal{N}_{p,q}(x))^* z - \mathcal{N}'_{p,q}(x)\mathcal{F}'(\mathcal{N}_{p,q}(x))^* z\|_2}_{D_2} \ ,
\end{aligned}$$

and by Proposition 3.2 we get

$$\mathcal{N}_{p,q}(x_n) \xrightarrow{\ell_2} \mathcal{N}_{p,q}(x). \tag{30}$$

Hence the two terms can be estimated as follows:

$$D_1 \le \underbrace{\|\mathcal{N}'_{p,q}(x_n)\|_2}_{\overset{(28)}{\le} C} \| \underbrace{\mathcal{F}'(\mathcal{N}_{p,q}(x_n))^* z - \mathcal{F}'(\mathcal{N}_{p,q}(x))^* z\|_2}_{\overset{(26),(30)}{\longrightarrow} 0}$$

and therefore $D_1 \to 0$. For D_2 we get with $\tilde{z} := \mathcal{F}'(\mathcal{N}_{p,q}(x))^* z$

$$D_2 = \|\mathcal{N}'_{p,q}(x_n)\tilde{z} - \mathcal{N}'_{p,q}(x)\tilde{z}\|_2 \xrightarrow{(29)} 0 \ ,$$

which concludes the proof. □

In the final step of this section we show the Lipschitz continuity of the derivative.

Proposition 4.5. *Assume that $\mathcal{F}'(x)$ is (locally) Lipschitz continuous with constant L. Then $(\mathcal{F}\circ\mathcal{N}_{p,q})'(x)$ is locally Lipschitz for $p < 1$ and $1 \le q \le 2$ such that $2p < q$.*

Proof. The function $f(t) = |t|^s$ with $s > 1$ is locally Lipschitz continuous, i.e. we have on a bounded interval $[a, b]$:

$$|f(t) - f(\tilde{t})| \le s \max_{\tau\in[a,b]} |\tau|^{s-1} |t - \tilde{t}| \ . \tag{31}$$

Assume $x \in B_\rho(x_0)$, then $\|x\|_2 \leq \|x - x_0\|_2 + \|x_0\|_2 \leq \rho + \|x_0\|_2$, and therefore

$$\sup_{x \in B_\rho(0)} \|x\|_2 \leq \rho + \|x_0\|_2 =: \tilde{\rho} \ .$$

We have $s := (q-p)/p \geq 1$, and $|t|^s$ is locally Lipschitz according to (31). $\mathcal{N}'_{p,q}(x)$ is a diagonal matrix, thus we obtain with Lemma 4.3 for $x, \tilde{x} \in B_\rho(x_0)$

$$\begin{aligned}
\|\mathcal{N}'_{p,q}(x) - \mathcal{N}'_{p,q}(\tilde{x})\|^2 &= \left(\frac{q}{p}\right)^2 \sum_k \left(|x_k|^{(q-p)/p} - |\tilde{x}_k|^{(q-p)/p}\right)^2 \\
&\overset{(31)}{\leq} \left(\frac{q}{p}\right)^2 \left(\frac{q-p}{p}\tilde{\rho}^{(q-2p)/p}\right)^2 \sum_k |x_k - \tilde{x}_k|^2 \\
&\leq \left(\frac{q}{p}\right)^2 \left(\frac{q-p}{p}\tilde{\rho}^{(q-2p)/p}\right)^2 \|x - \tilde{x}\|_2^2 \ .
\end{aligned}$$

With the same arguments we show that $\mathcal{N}_{p,q}$ is Lipschitz,

$$\|\mathcal{N}_{p,q}(x) - \mathcal{N}_{p,q}(\tilde{x})\|_2 \leq \frac{q}{p}\tilde{\rho}^{(q-p)/p}\|x - \tilde{x}\|_2 \ .$$

The assertion now follows from

$$\begin{aligned}
&\|\mathcal{F}'(\mathcal{N}_{p,q}(x))\mathcal{N}'_{p,q}(x) - \mathcal{F}'(\mathcal{N}_{p,q}(\tilde{x}))\mathcal{N}'_{p,q}(\tilde{x})\| \\
&\qquad \leq \| \left(\mathcal{F}'(\mathcal{N}_{p,q}(x)) - \mathcal{F}'(\mathcal{N}_{p,q}(\tilde{x}))\right)\mathcal{N}'_{p,q}(x)\| \\
&\qquad\qquad + \|\mathcal{F}'(\mathcal{N}_{p,q}(\tilde{x}))\left(\mathcal{N}'_{p,q}(x) - \mathcal{N}'_{p,q}(\tilde{x})\right)\| \\
&\qquad \leq L\|\mathcal{N}_{p,q}(x) - \mathcal{N}_{p,q}(\tilde{x})\|\|\mathcal{N}'_{p,q}(x)\| \\
&\qquad\qquad + \|\mathcal{F}'(\mathcal{N}'_{p,q}(\tilde{x}))\|\|\mathcal{N}'_{p,q}(x) - \mathcal{N}'_{p,q}(\tilde{x})\| \\
&\qquad \leq \tilde{L}\|x - \tilde{x}\| \ ,
\end{aligned}$$

with

$$\begin{aligned}
\tilde{L} = L \ &\max_{x \in B_\rho} \|\mathcal{N}'_{p,q}(x)\| \ \frac{q}{p} \ \tilde{\rho}^{(q-p)/p} + \\
&\max_{x \in B_\rho} \|\mathcal{F}'(\mathcal{N}_{p,q}(x))\| \ \left(\frac{q}{p}\right)^2 \ \frac{q-p}{p} \ \tilde{\rho}^{(q-2p)/p} \ .
\end{aligned}$$

□

Combining the results of Lemma 4.2 and Propositions 4.1, 4.4 and 4.5, we get

Proposition 4.6. *Assume that the operator $\mathcal{F} : \ell_2 \to \ell_2$ is Fréchet differentiable and fulfills conditions* (19)-(21). *Then $\mathcal{G} = \mathcal{F} \circ \mathcal{N}_{p,q}$ is also Fréchet differentiable. If the parameters $0 < p < 1$ and $1 < q \leq 2$ fulfill the relation $2p < q$, then we have*

$$x_n \rightharpoonup x \Longrightarrow \quad \mathcal{G}(x_n) \to \mathcal{G}(x) \text{ for } n \to \infty \tag{32}$$

$$x_n \rightharpoonup x \Longrightarrow \quad \mathcal{G}'(x_n)^* z \to \mathcal{G}'(x)^* z \text{ for } n \to \infty \text{ and all } z \in \ell_2 \tag{33}$$

$$\|\mathcal{G}'(x) - \mathcal{G}'(x')\|_2 \leq \tilde{L}\|x - x'\|_2 \quad \text{locally.} \tag{34}$$

Proof. Proposition 4.1 yields (32). According to Lemma 4.2, $\mathcal{G}$ is differentiable. If $q > 2p$ then the conditions of Proposition 4.4 hold and thus (33). Moreover, the condition $q > 2p$ is equivalent to $q > 2p$, i.e. Proposition 4.5 holds and therefore (34). □

5. Minimization by surrogate functionals

In order to compute a minimizer of the Tikhonov functional (4), we can either use algorithms that minimize (4) directly or, alternatively, we can try to minimize (10). It turns out that the transformed functional, with an ℓ_q-norm and $q > 1$ as penalty, can be minimized more effectively by the proposed or other standard algorithms. The main drawback of the transformed functional is that, due to the transformation, we have to deal with a non-linear operator, even if the original operator $\mathcal{F}$ is linear.
A well investigated algorithm for the minimization of the Tikhonov functional with ℓ_q penalty that works for all $1 \leq q \leq 2$ is the minimization via surrogate functionals. The method was introduced by Daubechies, Defrise and De Mol [12] for penalties with $q \geq 1$ and linear operator $\mathcal{F}$. Later on, the method was generalized in [13, 14, 23] to non-linear operators $\mathcal{G} = \mathcal{F} \circ \mathcal{N}_{p,q}$. The method works as follows: For given iterate x^n, we consider the surrogate functional

$$J_\alpha^s(x, x^n) = \|y^\delta - \mathcal{G}(x)\|^2 + \alpha\|x\|_q^q + C\|x - x^n\|_2^2 - \|\mathcal{G}(x) - \mathcal{G}(x^n)\|_2^2 \tag{35}$$

and determine the new iterate as

$$x^{n+1} = \arg\min_x J_\alpha^s(x, x^n) \ . \tag{36}$$

The constant C in the definition of the surrogate functional has to be chosen large enough, for more details see [13, 23]. Now it turns out that the functional $J_\alpha^s(x, x^n)$ can be easily minimized by means of a fixed point iteration. For fixed x^n, the functional is minimized by the limit of the fixed point iteration

$$x^{n,l+1} = \Phi_q^{-1}\left(\frac{1}{C}\mathcal{G}'(x^{n,l})^*\left(y^\delta - \mathcal{G}(x^n)\right) + x^n\right) , \tag{37}$$

$x^{n,0} = x^n$ and $x^{n+1} = \lim_{l\to\infty} x^{n,l}$. For $q > 1$, the map Φ_q is defined point-wise on the coefficients of a sequence by

$$\Phi_q(x_k) = x_k + \frac{\alpha \cdot q}{C}|x_k|^{q-1}\operatorname{sgn}(x_k) , \tag{38}$$

i.e. in order to compute the new iterate $x^{n,l+1}$ we have to solve the equation

$$\Phi_q\left(\left\{x^{n,l+1}\right\}_k\right) = \left\{\frac{1}{C}\mathcal{G}'(x^{n,l})^*\left(y^\delta - \mathcal{G}(x^n)\right) + x^n\right\}_{k\in\mathbb{N}} \tag{39}$$

for each $k \in \mathbb{N}$. It has been shown that the fixed point iteration converges to the unique minimizer of the surrogate functional $J_\alpha^s(x, x^n)$, provided the constant C is chosen large enough and the operator fulfills the Requirements (19)–(21), for full details we refer the reader to [13, 23]. Moreover, it was also shown that the outer iteration (36) converges at least to a critical point of the Tikhonov functional

$$J_\alpha(x) = \|y^\delta - \mathcal{G}(x)\|_2^2 + \alpha\|x\|_q^q , \tag{40}$$

provided that the operator $\mathcal{G}$ fulfills the conditions (32)-(34).

Based on the results of Section 2, we can now formulate our main result.

Theorem 5.1. *Let $\mathcal{F} : \ell_2 \to \ell_2$ be a weakly (sequentially) closed operator fulfilling Conditions* (19)–(21)*, and choose $q > 1$ s.t. $2p < q$, with $0 < p < 1$. Then the operator $\mathcal{G}(x) = \mathcal{F} \circ \mathcal{N}_{p,q}$ is Fréchet differentiable and fulfills the conditions* (32)–(34)*. The iterates x_n, computed by the surrogate functional algorithm* (36)*, converge at least to a critical point of the functional*

$$J_{\alpha,q}(x) = \|y^\delta - \mathcal{G}(x)\|_2^2 + \alpha\|x\|_q^q . \tag{41}$$

If the limit of the iteration, $x_\alpha^\delta := \lim_{n\to\infty} x^n$, is a global minimizer of (41)*, then $x_{s,\alpha}^\delta := \mathcal{N}_{p,q}(x_\alpha^\delta)$ is a global minimizer of*

$$\|y^\delta - \mathcal{F}(x)\|_2^2 + \alpha\|x\|_p^p . \tag{42}$$

Proof. According to Proposition 4.6, the operator $\mathcal{G}$ fulfills the properties necessary for the convergence of the iterates to a critical point of the functional (41), see [23], Proposition 4.7. If x_α^δ is a global minimizer of (41), then, according to Proposition 2.4, $x_{s,\alpha}^\delta$ is a minimizer of (4). □

One may notice that the main result in Theorem 5.1 is stated with respect to the transformed functional. Where a global minimizer is reconstructed the result can be interpreted in terms of the original functional. In fact this can be slightly generalized. Assuming that the limit of the iteration is no saddle point, i.e. we obtain a local minimizer or a stationary point where the objective function is locally constant, we can directly translate this result to the original functional. Let x_α^δ be the limit of the iteration and assume there exists a neighborhood $U_\epsilon(x_\alpha^\delta)$ such that:

$$\forall x \in U_\epsilon(x_\alpha^\delta) \quad : \quad \|y^\delta - \mathcal{G}(x)\|_2^2 + \alpha\|x\|_q^q \geq \|y^\delta - \mathcal{G}(x_\alpha^\delta)\|_2^2 + \alpha\|x_\alpha^\delta\|_q^q \,. \quad (43)$$

Let $M := \{x_s : \mathcal{N}_{p,q}^{-1}(x_s) \in U_\epsilon(x_\alpha^\delta)\}$ and $x_{s,\alpha}^\delta := \mathcal{N}_{p,q}(x_\alpha^\delta)$, then we can derive that:

$$\forall x_s \in M \quad : \quad \|y^\delta - \mathcal{F}(x_s)\|_2^2 + \alpha\|x_s\|_p^p \geq \|y^\delta - \mathcal{F}(x_{s,\alpha}^\delta)\|_2^2 + \alpha\|x_{s,\alpha}^\delta\|_p^p \,. \quad (44)$$

Since $\mathcal{N}_{p,q}$ and $\mathcal{N}_{p,q}^{-1}$ are continuous there exists a neighborhood U_{ϵ_s} around the solution for the original functional $x_{s,\alpha}^\delta$, such that $U_{\epsilon_s}(x_{s,\alpha}^\delta) \subseteq M$. Consequently also stationary points and local minima of the transformed functional translate to the original functional.

6. A global minimization strategy for the transformed Tikhonov functional: the case $q = 2$

The minimization by surrogate functionals, presented in Section 5, guarantees the reconstruction of a critical point of the transformed functional only. If we have not found the global minimizer of the transformed functional, then this also implies that we have not reconstructed the global minimizer for the original functional. In this Section we would like to recall an algorithm that, under some restrictions, guarantees the reconstruction of a global minimizer. In contrast to the surrogate functional approach, this algorithm works in the case of $q = 2$ only, i.e. we are looking for a global minimizer of the standard Tikhonov functional

$$J_\alpha(x) = \|y^\delta - \mathcal{G}(x)\|^2 + \alpha\|x\|_2^2 \quad (45)$$

with $\mathcal{G}(x) = \mathcal{F}(\mathcal{N}_{p,2}(x))$. For the minimization of the functional, we want to use the TIGRA method [32, 33]. The main ingredient of the algorithm is a standard gradient method for the minimization of (45), i.e. the iteration is given by

$$x^{n+1} = x^n + \beta_n \left(\mathcal{G}'(x^n)^*(y^\delta - \mathcal{G}(x^n)) - \alpha x^n \right). \tag{46}$$

The following arguments are taken out of [33], where the reader finds all the proofs and further details. If the operator $\mathcal{G}$ is twice Fréchet differentiable, its first derivative is Lipschitz continuous, and a solution $x^\dagger$ of $\mathcal{G}(x) = y$ fulfills the smoothness condition

$$x^\dagger = \mathcal{G}'(x^\dagger)^*\omega \ , \tag{47}$$

then it has been shown that (45) is locally convex around a global minimizer x_α^δ. If an initial iterate x^0 within the area of convexity is known, then the scaling parameter β_n can be chosen s.t. all iterates stay within the area of convexity and $x^n \to x_\alpha^\delta$ as $n \to \infty$. However, the area of convexity shrinks to zero if $\alpha \to 0$, i.e., a very good initial iterate for smaller α is needed. For an arbitrary initial iterate x^0 this problem can be overcome by choosing a monotone decreasing sequence $\alpha_0 > \alpha_1 > \cdots > \alpha_n = \alpha$ with sufficiently large α_0 and small stepsize α_{i+1}/α_i, and iterate as follows:

Input: $x^0, \alpha_0, \cdots \alpha_n$

Iterate: For $i = 1, \cdots, n$

- If $i > 1$, set $x^0 = x_{\alpha_{i-1}}^\delta$
- Minimize $J_{\alpha_i}(x)$ by the gradient method (46) and initial value x^0.

End

We wish to remark that the iteratively regularized Landweber iteration, introduced by Scherzer [36], is close to TIGRA. Its iteration is similar to (46), but requires the use of a summable sequence α_k (instead of a fixed α). In contrast to TIGRA, the iteratively regularized Landweber iteration aims at the solution of a nonlinear equation but not on the minimization of a Tikhonov functional. Additionally, iteratively regularized Landweber iteration requires more restrictive conditions on the nonlinear operator.

In a numerical realization, the iteration (46) has to be stopped after finitely many steps. Therefore the final iterate is taken as starting value for

the minimization of the Tikhonov functional with the next regularization parameter. As mentioned above this procedure reconstructs a global minimizer of J_α if the operator $\mathcal{G}$ is twice Fréchet differentiable, its first derivative is Lipschitz continuous and (47) holds [33]. We will show these conditions for two important cases, namely where $\mathcal{F}$ is the identity (i.e. the problem of data denoising), and when $\mathcal{F}$ is a linear operator, $\mathcal{F} = \mathcal{A}$.

Proposition 6.1. *The operator $\mathcal{N}_{p,2}(x)$, $0 < p < 1$, is twice continuous differentiable, and therefore also the operator $\mathcal{A}\mathcal{N}_{p,2}(x)$ with continuous and linear $\mathcal{A}$.*

Proof. The proof is completely analogous to the one of Proposition 3.5, when considering the fact that $\frac{2}{p} \geq 2$. Using the Taylor expansion of the function $\eta_{p,2}(t) = |t|^{2/p}\operatorname{sgn}(t)$:

$$\eta_{p,2}(t+\tau) - \eta_{p,2}(t) - \eta'_{p,2}(t)\,\tau - \frac{1}{2}\,\eta''_{p,2}(t)\,\tau^2 \;:= r(t,\tau)\;,$$

with

$$\eta''_{p,2}(t) = \frac{2(2-p)}{p^2}\operatorname{sgn}(t)|t|^{2(1-p)/p}\;,$$

one obtains the following representation of the remainder:

$$r(t,\tau) = \int_t^{t+\tau} \frac{1}{2}\frac{2}{p}\frac{2-p}{p}\frac{2-2p}{p}\,(t+\tau-s)^2\,|s|^{\frac{2}{p}-3}\,\mathrm{d}s\;,$$

and again by the mean value theorem:

$$\begin{aligned}
&\left|\int_t^{t+\tau} \frac{1}{2}\frac{2}{p}\frac{2-p}{p}\frac{2-2p}{p}\,(t+\tau-s)^2\,|s|^{\frac{2}{p}-3}\,\mathrm{d}s\right| = \\
&= \left|\left[\frac{1}{2}\frac{2}{p}\frac{2-p}{p}\,(t+\tau-s)^2\,|s|^{\frac{2}{p}-2}\right]_t^{t+\tau} + \int_t^{t+\tau}\frac{2}{p}\frac{2-p}{p}\,(t+\tau-s)\,|s|^{\frac{2}{p}-2}\,\mathrm{d}s\right| \\
&= \left|\tau\frac{2}{p}\frac{2-p}{p}\,\left((t+\tau-\xi)\operatorname{sgn}(\xi)|\xi|^{2/p-2} - \frac{1}{2}\tau\operatorname{sgn}(t)|t|^{2/p-2}\right)\right| \\
&\overset{(14)}{\leq} \tilde{\kappa}\,\frac{2}{p}\frac{2-p}{p}\,|\tau|^{w+2}\;,
\end{aligned}$$

where $\xi \in (t, t+\tau)$, $w := \min\left(\frac{2}{p}-2, 1\right) > 0$ and by using Lemma 3.3 with $\alpha = \frac{2}{p} - 2$. One may note that the scaling factor $1/2$ requires a redefinition

of κ in Lemma 3.3, leading to $\tilde{\kappa}$. Eventually we conclude for $\|h\|_2 \to 0$

$$\left\|\mathcal{N}'_{p,2}(x+h)\bar{h} - \mathcal{N}'_{p,2}(x)\bar{h} - \mathcal{N}''_{p,2}(x)(\bar{h},h)\right\|_2 / \|h\|_2 \to 0$$

analogously to the proof of Proposition 3.5. Thus we have

$$\mathcal{N}''_{p,2}(x)(\bar{h},h) = \left\{\eta''_{p,q}(x_k)\bar{h}_k h_k\right\}_{k\in\mathbb{N}} .$$

The twice differentiability of $\mathcal{A}\mathcal{N}_{p,2}(x)$ follows from the linearity of $\mathcal{A}$. □

Now let us turn to the source condition (47).

Proposition 6.2. *Let $\mathcal{F} = \mathcal{I}$. Then $x^\dagger \in \ell_2$ fulfills the source condition* (47) *iff it is sparse.*

Proof. As $\mathcal{I} = \mathcal{I}^*$ in ℓ_2, we have $\mathcal{F}'(\mathcal{N}_{p,2}(x^\dagger))^* = \mathcal{I}$, and it follows from (24) that

$$(\mathcal{F}(\mathcal{N}_{p,2}(x))')^* = \mathcal{N}'_{p,2}(x) .$$

Therefore, the source condition (47) reads coefficient-wise as

$$\frac{2}{p}|x_k^\dagger|^{(2-p)/p}\omega_k = x_k^\dagger$$

or

$$\omega_k = \frac{2}{p}\operatorname{sgn}(x_k^\dagger)|x_k^\dagger|^{(2p-2)/p} ,$$

for $x_k \neq 0$, for $x_k = 0$ we can set $w_k = 0$, too. As $\omega_k,\ x^\dagger \in \ell_2$ and $2p-2<0$ this can only hold if $x^\dagger$ has only a finite number of non-zero elements. □

The case of $\mathcal{F} = \mathcal{A}$ is a little bit more complicated. In particular, we need the operator $\mathcal{A}$ to fulfill the finite basis injectivity (FBI) property which was introduced by Bredies and Lorenz [37]. Let $\mathcal{T}$ be a finite index set, and let $\#\mathcal{T}$ be the number of elements in $\mathcal{T}$. We say that $u \in \ell_2(\mathcal{T})$ iff $u_k = 0$ for all $k \in \mathbb{N} \setminus \mathcal{T}$. The FBI property states that whenever $u, v \in \ell_2(\mathcal{T})$ with $\mathcal{A}u = \mathcal{A}v$ it follows $u = v$. This is equivalent to

$$\mathcal{A}_{|\ell_2(\mathcal{T})}u = 0 \Longrightarrow u = 0 , \tag{48}$$

where $\mathcal{A}_{|\ell_2(\mathcal{T})}$ is the restriction of $\mathcal{A}$ to $\ell_2(\mathcal{T})$. For simplicity, we set $\mathcal{A}_{|\ell_2(\mathcal{T})} = \mathcal{A}_{\mathcal{T}}$.

Proposition 6.3. *Assume that $x^\dagger$ is sparse, $\mathcal{T} = \{k : x_k^\dagger \neq 0\}$, and that $\mathcal{A} : \ell_2 \to \ell_2$ is bounded. If $\mathcal{A}$ admits the FBI property, then $x^\dagger$ fulfills the source condition* (47)*.*

Proof. As $x^\dagger$ is sparse, $\mathcal{T}$ is finite. By $x_\mathcal{T}$ we denote the (finite) vector that contains only those elements of x with indices out of $\mathcal{T}$. As $\mathcal{A}$ is considered as an operator between ℓ_2 we have $\mathcal{A}^* = \mathcal{A}^T$ and $\mathcal{A}_\mathcal{T}^* = \mathcal{A}_\mathcal{T}^T$. Due to the sparse structure of $x^\dagger$ we observe

$$\mathcal{N}'_{p,2}(x^\dagger) : \ell_2 \to \ell_2(\mathcal{T})$$

and therefore also

$$\begin{aligned} \mathcal{A}\mathcal{N}'_{p,2}(x^\dagger) &= \mathcal{A}_\mathcal{T}\mathcal{N}'_{p,2}(x^\dagger) && (49) \\ \left(\mathcal{A}\mathcal{N}'_{p,2}(x^\dagger)\right)^* &= \mathcal{N}'_{p,2}(x^\dagger)\mathcal{A}_\mathcal{T}^* = \mathcal{N}'_{p,2}(x^\dagger)\mathcal{A}_\mathcal{T}^T \ , && (50) \end{aligned}$$

where we use the fact that $\mathcal{N}'_{p,2}(x^\dagger)$ is self-adjoint.

With $\mathcal{F} = \mathcal{A}$, (47) reads as

$$x^\dagger = \mathcal{N}'_{p,2}(x^\dagger)\mathcal{A}_\mathcal{T}^T\omega \ . \tag{51}$$

The operator $\mathcal{N}'_{p,2}(x^\dagger)^{-1}$ is well defined on $\ell_2(\mathcal{T})$, and as $\ell_2(\mathcal{T}) = \mathcal{D}(\mathcal{A}_\mathcal{T}) = \mathcal{R}(\mathcal{A}_\mathcal{T}^T)$, we get

$$\mathcal{A}_\mathcal{T}^T\omega = \mathcal{N}'_{p,2}(x^\dagger)^{-1}x^\dagger \ .$$

Now we have by the FBI property $\mathcal{N}(\mathcal{A}_\mathcal{T}) = \{0\}$, and therefore

$$\ell_2(\mathcal{T}) = \mathcal{N}(\mathcal{A}_\mathcal{T})^\perp = \overline{\mathcal{R}(\mathcal{A}_\mathcal{T}^*)} = \overline{\mathcal{R}(\mathcal{A}_\mathcal{T}^T)} \ .$$

As $dim(\ell_2(\mathcal{T})) = \#\mathcal{T} < \infty$, $\mathcal{R}(\mathcal{A}_\mathcal{T}^T) = \ell_2(\mathcal{T})$ and therefore the generalized inverse of $\mathcal{A}_\mathcal{T}^T$ exists and is bounded. We finally get

$$\omega = \left(\mathcal{A}_\mathcal{T}^T\right)^\dagger \mathcal{N}'_{p,2}(x^\dagger)^{-1}x^\dagger \tag{52}$$

and

$$\|\omega\|_2 \leq \| \left(\mathcal{A}_\mathcal{T}^T\right)^\dagger \|_2 \|\mathcal{N}'_{p,2}(x^\dagger)^{-1}\|_2 \|x^\dagger\|_2 \ . \tag{53}$$

□

Please note that a similar result can be obtained for twice continuous differentiable non-linear operators $\mathcal{F}$ if we additionally assume that $\mathcal{F}'(\mathcal{N}_{p,2}(x^\dagger))$ admits the FBI condition. Propositions 6.1–6.3 show that the TIGRA algorithm can be applied in principle to the minimization of the transformed Tikhonov functional for the case $q = 2$ and reconstructs a global minimizer. Please note that the surrogate functional approach can also be applied to the case $q < 2$. This is in particular important for the numerical realization, as we show in the following section.

7. Numerical Results

In this section we exemplify the utilization of the proposed algorithm for two classical Inverse Problems. We examine a deconvolution problem in Fourier Spaces and a parameter identification problem from physical chemistry with a highly non-linear operator. Considering the proposed non-standard approach, the impacts of a numerical realization are hardly predictable even though the analytic properties of the non-linear transformation are well understood and the surrogate approach has been tested extensively.

7.1. Deconvolution on Sequence Spaces

Subsequently we present some numerical results on the reconstruction of a function from convolution data. We define the convolution operator A by

$$y(\tau) = (Ax)(\tau) = \int_{-\pi}^{\pi} r(\tau - t)x(t)\,dt =: (r * x)(\tau)\,, \tag{54}$$

where x, r and Au are 2π-periodic functions belonging to $L_2((-\pi, \pi))$. In the above formulation the operator A is defined between function spaces. In order to obtain a numerical realization in accordance with the present notation we have to transform this operator to sequence spaces (cf. Section 1). For this purpose we interpret all quantities in terms of the Fourier basis or their Fourier coefficients, respectively. A periodic function on $[-\pi, \pi]$ can be either expressed via the orthonormal bases formed by $\{\frac{1}{\sqrt{2\pi}} e^{ikt}\}_{k\in\mathbb{Z}}$ or $\{\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos(kt), \frac{1}{\sqrt{\pi}}\sin(kt)\}_{k\in\mathbb{N}}$. Naturally these representations provide also the appropriate discretization of the (linear) operator. By using the Fourier convolution theorem for the exponential basis and transformation formulas between the exponential and trigonometrical bases, we obtain a

formulation in terms of the considered real sequence spaces. We refer to [24] for details on the deconvolution problem and the operator.

For the numerical implementation we divide the interval $[-\pi, \pi]$ into 2^{12} equidistant intervals, leading to a discretization of the convolution operator as a $2^{12} \times 2^{12}$ matrix. We define the convolution kernel r by its Fourier coefficients with

$$\begin{aligned} a_0^r &= 0 \\ a_k^r &= (-1)^k \cdot k^{-2} \\ b_k^r &= (-1)^{k+1} \cdot k^{-2} \,, \end{aligned} \tag{55}$$

where

$$r(t) = a_0^r + \sum_{k \in \mathbb{N}} a_k^r \cos(kt) + b_k^r \sin(kt) \,. \tag{56}$$

For the numerical tests a convolution data set has been generated based on a (sparse-representable) solution $x^\dagger$ with 14 non-zero components (see Table 1).

index	$\left(x_1^\dagger, \ldots, x_{2041}^\dagger\right)$	$x_{2042}^\dagger$	$x_{2043}^\dagger$	$x_{2044}^\dagger$	$x_{2045}^\dagger$	$x_{2046}^\dagger$	$x_{2047}^\dagger$	$x_{2048}^\dagger$	$x_{2049}^\dagger$
value	$(0, \ldots, 0)$	$\frac{1}{70}$	$\frac{1}{60}$	$\frac{1}{50}$	$\frac{1}{40}$	$\frac{1}{30}$	$\frac{1}{20}$	$\frac{1}{10}$	0

index	$x_{2050}^\dagger$	$x_{2051}^\dagger$	$x_{2052}^\dagger$	$x_{2053}^\dagger$	$x_{2054}^\dagger$	$x_{2055}^\dagger$	$x_{2056}^\dagger$	$\left(x_{2057}^\dagger, \ldots, x_{4097}^\dagger\right)$
value	$-\frac{1}{10}$	$-\frac{1}{20}$	$-\frac{1}{30}$	$-\frac{1}{40}$	$-\frac{1}{50}$	$-\frac{1}{60}$	$-\frac{1}{70}$	$(0, \ldots, 0)$

Table 1: The coefficients of the true solution $x^\dagger$ and in particular the 14 non-zero components are shown, where $a^{x^\dagger} = \left(x_1^\dagger, \ldots, x_{2048}^\dagger\right)$, $a_0^{x^\dagger} = x_{2049}^\dagger$ and $b^{x^\dagger} = \left(x_{2050}^\dagger, \ldots, x_{4097}^\dagger\right)$ in accordance to (56).

The added noise is normally distributed and scaled with respect to the relative noise level. If not stated otherwise we assume an approximate noise level of five percent relative Gaussian noise. In all numerical tests the regularization parameter is chosen based on the quasi optimality principle (cf. [38, 39]) or by the discrepancy principle (cf. [40]). The quasi optimality principle provides an easy heuristic method of estimating the regularization parameter in the considered setting. Both methods provide rather good estimates of the regularization parameter. However, since the exact solution was known prior to the numerical experiments, we can compute the optimal values for the

regularization parameter a posteriori by using fine grids of different regularization parameters. We can confirm the stable convergence for the algorithm, which we found in all our experiments. In [28] it was shown that (4) provides a regularization method and moreover a result on convergence rates was given, stating that the convergence of the accuracy error is at least in the order of $\sqrt{\delta}$,

$$\left\| x^* - x_\alpha^\delta \right\|_2 = \mathcal{O}(\sqrt{\delta}) \ ,$$

under standard (source) conditions (see [28] for further details). Using the proposed algorithm, we can observe the theoretical result on the convergence rates also in our numerical tests. Figure 1 shows the rates of convergence

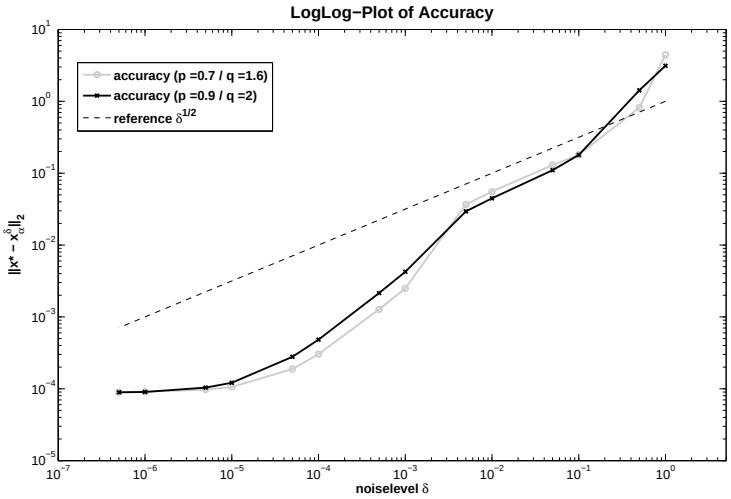

Figure 1: The plot shows the numerically observed rate of convergence for $p = 0.9$ / $q = 2$ and $p = 0.7$ / $q = 1.6$, compared to the reference rate $\delta^{1/2}$.

for decreasing noise levels δ. One may notice that the convergence slows down for very small values of δ. This behavior is well known and can be caused by local minima or too coarse accuracy goals. Another difficulty could be a numerical inaccuracy of the inner iterations (37), which might cause a stagnation of the iteration. In fact, the observed loss of accuracy is very pronounced if the error criteria in the inner iteration are not chosen appropriately. This observation seems comprehensible when considering the transformation operator and the fact that all unknowns are potentiated by a factor of q/p. This is also in accordance with our findings of an increased computational effort (i.e. higher iteration numbers) and the necessity of

a sufficiently high numerical precision for decreasingly smaller values of p. These requirements are met by an efficient implementation with stringent error criteria in all internal routines or iterations, respectively. For the inner iteration we control the absolute error per coefficient, i.e.

$$\max_k |\{x^{n,l+1} - x^{n,l}\}_k| = \mathcal{O}(10^{-14}) \,, \tag{57}$$

whereas the outer iteration is stopped after sufficient convergence with respect to the data error and point-wise changes for the parameters, i.e.

$$\left\|y^\delta - \mathcal{G}(x^{n+1})\right\|_2^2 = \mathcal{O}(10^{-8}) \quad \text{and} \quad \max_k \left|\{x^{n+1} - x^n\}_k\right| = \mathcal{O}(10^{-6}) \,. \tag{58}$$

The inner loop usually converges within 2 to 10 iterations, as frequently observed for the surrogate approach. The number of outer iterations strongly depends on the respective settings and chosen parameters.

Figure 2 shows the exact data curve (without noise) and the obtained reconstructions of the data for various values of p and $q = 2$ and approximatly 5% Gaussian noise. The right hand axis refers to the difference of these curves, which is plotted below. For increasing values of p the data fit improves. Correspondingly the number of non-zero coefficients in the reconstructed Fourier coefficients increases as well. For $p = 0.4$ we reconstruct seven non-zero coefficients, for $p = 0.5$ the obtained solution consists of seven non-zero components, for $p = 0.7$ we get ten non-zero coefficients and for $p = 0.9$ the solution has twelve non-zero coefficients, compared to 14 non-zero entries in the true solution $x^\dagger$. The decreasing number of non-zero coefficients indicates an increased promotion of sparse coefficient vectors for smaller values of p. This is especially worth mentioning since already for $p = 0.9$ the number of non-zero coefficients is underestimated. Further one may note that the zero components of these solutions are really zero with respect to the machine precision (ca. 10^{-16}). Only in the solution for $p = 0.9$, $q = 2$ several "outliers" in the order of 10^{-5}–10^{-10} are found. Further increasing the error criteria would provide a remedy. Additionally we find that the proposed regularization method is sensitive w.r.t. the choice of the regularization parameter α, which we account for by using fine grids of values for α ($0.5^{\{0,1,2,\ldots,\}}$). Moreover, we would like to emphasize that all the obtained non-zero coefficients are within the support of the true solution $x^\dagger$. Only for $p = 0.9$ and $q = 2$, some of the addressed very small outliers lie outside the support of the original solution. Eventually we obtain good data fits for

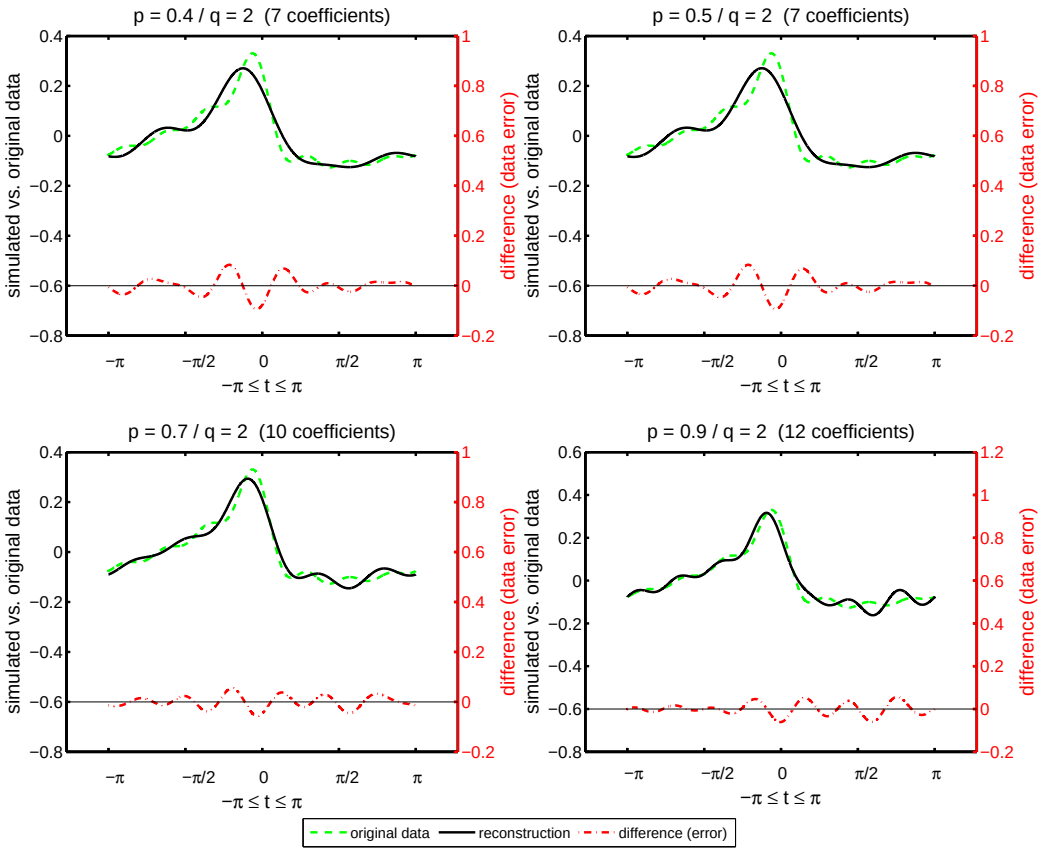

Figure 2: The exact data of the deconvolution problem and the obtained simulated data from the reconstructions for $p = \{0.4, 0.5, 0.7, 0.9\}$ are plotted. Below the difference between the (noise free) data curve and the simulated data from the reconstruction is given (see right y-axis).

all values of p and q, even in those cases where the reconstructed solution consists only of four non-zero components, compared to 14 coefficients in the true solution.

Figure 3 shows the progress of the iteration routine in the case of $p = 0.7$ and $q = 2$. Due to the high number of unknown coefficients (4097) we consider the progress of the iterates for a sub sample of the coefficients only. After 1500 iterations all coefficients greater than 10^{-6} lie within the index 2032 to 2068. The coefficients depicted in the first row are normalized w.r.t. the iterate after 1500 iterations, i.e. the iterate x_{1500} is taken as a reference and the subsequent iterates are scaled with respect to x_{1500}. As some coefficients

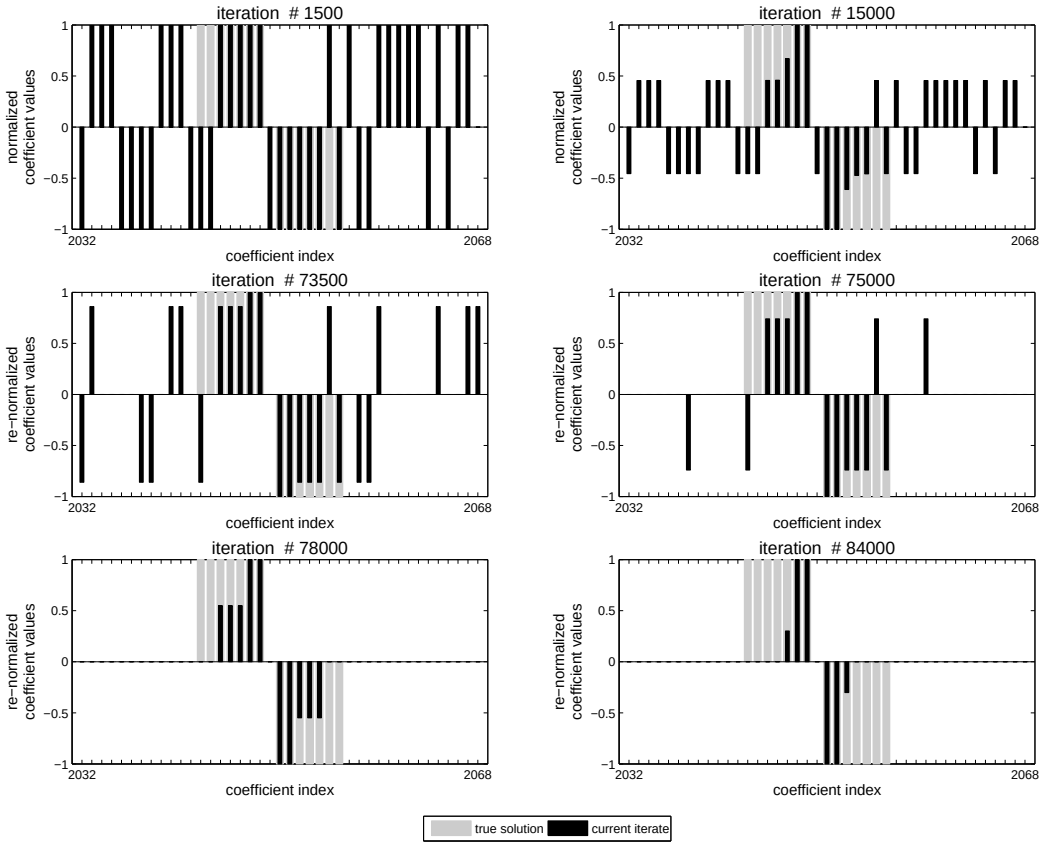

Figure 3: The six graphics show the progress of the support for the iterations 1.500, 15.000, 73.500, 75.000, 78.000 and 84.000 in the above test example for $p = 0.4$ and $q = 2$. Only normalized values of coefficients greater than 10^{-6} are plotted. Moreover the coefficients of the original coefficients $x^\dagger$ are highlighted. Bars of the current iterate are additionally scaled w.r.t. the intermediate iterate after 1.500 for the first row (iteration 1.500 and 15.000) and w.r.t. to the intermediate iterate after 30.000 iterations for the remaining plots for better visualization.

become very small, we "re-normalize" the values of the coefficients w.r.t. the iterate after 3000 iterations for the remaining four graphics. We observe that the support of the final iterate is contained in the support of the true solution. Moreover, the individual coefficients outside the support decrease monotonously.

We now address the choice of q, as it directly affects the algorithm. In [13] it was shown that the solution to (39) can be calculated analytically for $q = 2$. Consequently the computational effort is reduced significantly at the expense

of numerical artifacts in the case of $q = 2$. Figure 4 shows the number of

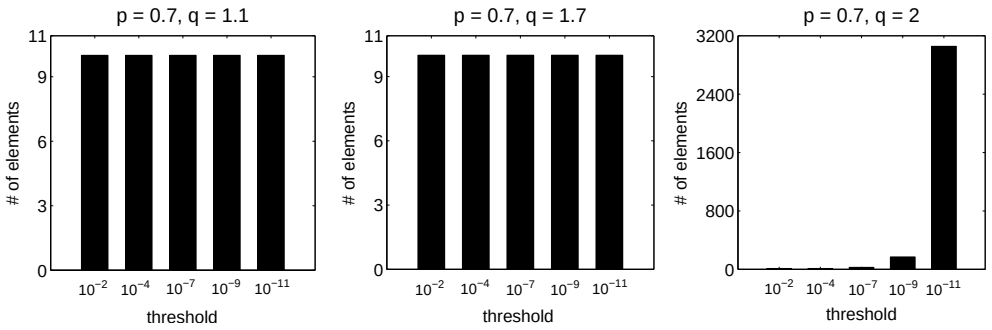

Figure 4: The bar charts show the number of (non-zero) entries above the thresholds $10^{-2}, 10^{-4}, 10^{-7}, 10^{-9}$ and 10^{-11}, for $q = 1.1, 1.4, 2$ and $p = 0.7$.

(non-zero) entries above a certain threshold in the cases of $q = 1.1, 1.4$ and $q = 2$. As one expects from the theory (cf. Sections 2 to 5) the choice of $q \in (1, 2)$ has no effect on the solution. Particularly no small non-zero entries occur. For $q = 2$ structurally the same solution is obtained, however due to numerical artifacts there is an increasing number of small coefficients w.r.t. the thresholds of 10^{-1}, 10^{-2}, 10^{-7}, 10^{-9} and 10^{-11}. These effects can be controlled by stringent error criteria, which were relaxed by a factor of 10^{-3} for the results in Figure 4. Eventually one can conclude that the choice of $q = 2$ reduces the computational effort but requires more stringent error criteria if small artifacts shall be prevented.

Finally we compare the classic surrogate algorithm for $p = 1$ with the approach proposed here. In particular we compare the results obtained for $p = 0.9$ and $q = 2$ with the case of $p = 1$, as they are presumably most similar. Figure 5 shows the data fit obtained by the ℓ_1-approach and compares this to the already presented result for $p = 0.9$ and $q = 2$. We observe that the quality of the obtained data fit is more or less identical and even slightly improved for $p = 0.9$. However, one may note that the reconstructed coefficient vector is more sparse for $p = 0.9$ than for $p = 1$. We observe twelve non-zero elements for the case of $p = 0.9$ and 13 non-zero elements for $p = 1$. Moreover, one may note that although all identified coefficients lie within the support of the true solution, in the case of $p = 1$ we observe that the reconstructed coefficient with index 2042 (most left coefficient in the lower left bar chart) has the wrong sign compared to the true solution. Eventually the shown comparison and the results presented for the cases of

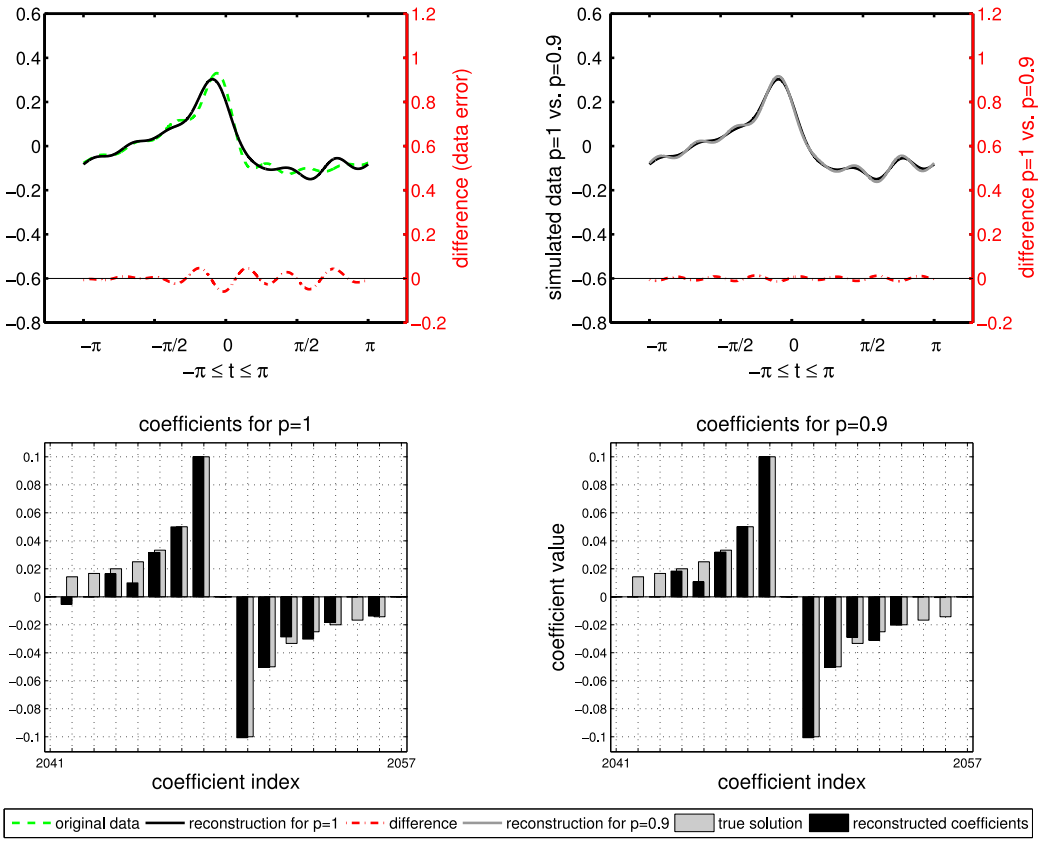

Figure 5: In the upper left corner the data fit obtained by the ℓ_1 surrogate approach is shown, similarly to Figure 2. On the upper right plot the ℓ_1 reconstruction is plotted against the obtained reconstruction for $p = 0.9$ and $q = 2$. Below the reconstructed coefficients are compared to the true solution for the two cases of $p = 1$ and $p = 0.9$.

$p = 0.4, 0.5$ and $p = 0.7$ (see Figure 2) indicate an approximation of the classic ℓ_1 surrogate approach for increasing p, which we would expect from the theory.

In summary the numerical deconvolution example confirms our analytical findings on the utilization of the transformation operator, the stable convergence and convergence rates. Additionally, the fact that the reconstructed solutions are always close to the true solution suggests that the algorithm reconstructs the global minimizer fitting the constructed data and thus providing good quality reconstructions. Moreover, the strong sparsity promoting feature of the considered regularization functionals and the principle idea of

exploiting the transformation operator in numerical algorithms is confirmed. Furthermore the comparison with the classic ℓ_1 surrogate approach suggests that the proposed approach can be seen as an extension of the ℓ_1 surrogate algorithm, with even increased sparsity promoting features. This allows a first rough assessment of the proposed algorithm within the framework of other sparsity promoting algorithms (cf. [41, 42, 43]). However, we would like to emphasize that these algorithms exclusively work for $p \geq 1$ and hence are not directly comparably.

7.2. Parameter Identification for a Chemical Reaction System - The Chlorite-Iodide Reaction

The second test example was taken from an application in physical chemistry. We demonstrate the advantages and capabilities of the suggested algorithm for a real world problem. In [11] Kügler et al. use sparsity promoting regularization for a parameter identification problem in a chemical reaction system, the so-called Chlorite-Iodide Reaction. This very well described chemical reaction network provides an attractive test example for the approach considered here. There are several motivations to enforce sparsity when identifying parameters in biological or chemical reaction systems (cf. [11, 44]). First, one may address the issue of manageable interpretations of the typically large networks. By identifying the most crucial parameters, still explaining the observed dynamics, one can obtain a reduced reaction network in the sense of model reduction. Secondly, the occurring species in these networks typically are not accessible to experimental measurements. Hence one inevitably lacks information and may encounter unidentifiable parameters. The sparsity promoting regularization is of particular interest, as it eliminates those unidentifiable parameters. They become zero, typically leading to zero rates or binding kinetics, and hence eliminate the respective reaction terms or species from the network. This is also in accordance with Ockham‘s razor, stating that the minimal solution is typically the most likely one. Especially when considering model based approaches this provides an attractive alternative to quantify these models by means of experimental data and reducing the probable model errors at the same time. Another application for sparsity promoting regularization arises for biological or chemical reaction systems, if we consider an already quantified model and want to incorporate additional data or different experimentally observed dynamics. By using the given parameter set as a prior in the regularization term we promote those solutions with a minimal number of changed parameters. In

this way previously identified parameters are not likely to change and moreover one might identify those mechanisms relevant for the newly observed data or dynamics.

The full chemical reaction system for the Chlorite-Iodide reaction consists of seven non-linear ODE equations, with 22 parameters, as shown below.

$$
\begin{aligned}
\frac{\mathrm{d}[ClO_2]}{\mathrm{d}t} &= -k_1\, ClO_2\, I^- + k_0\, INClO_2 - k_0\, ClO_2 \\
\frac{\mathrm{d}[HOCl]}{\mathrm{d}t} &= k_3\, HClO_2\, I^- + k_4\, HClO_2\, HOI + k_5\, HClO_2\, HIO_2 - k_6\, HOCl\, I^- \\
&\quad - k_7\, HOCl\, HIO_2 - k_0\, HOCl \\
\frac{\mathrm{d}[HIO_2]}{\mathrm{d}t} &= k_4\, HClO_2\, HOI - k_5\, HClO_2\, HIO_2 - k_7\, HOCl\, HIO_2 - k_{8f}\, HIO_2\, I^-\, H \\
&\quad + k_{8r}\, HOI^2 - 2\, k_9\, HIO_2^2 - k_{10}\, HIO_2\, H_2OI^+ - k_0\, HIO_2 \\
\frac{\mathrm{d}[TClO_2^-]}{\mathrm{d}t} &= k_1\, ClO_2\, I^- - k_3\, HClO_2\, I^- - k_4\, HClO_2\, HOI - k_5\, HClO_2\, HIO_2 \\
&\quad - k_0\, TClO_2^- \\
\frac{\mathrm{d}[THOI]}{\mathrm{d}t} &= k_{2af}\, I_2/H - k_{2ar}\, HOI\, I^- + k_{2bf}\, I_2 - k_{2br}\, H2OI^+\, I^- + k_3\, HClO_2\, I^- \\
&\quad - k_4\, HClO_2\, HOI + k_6\, HOCl\, I^- + 2\, \left(k_{8f}\, HIO_2\, I^-\, H - k_{8r}\, HOI^2\right) \\
&\quad + k_9\, HIO_2^2 - k_{10}\, HIO_2\, H_2OI^+ - k_0\, THOI \\
\frac{\mathrm{d}[TI^-]}{\mathrm{d}t} &= -k_1\, ClO_2\, I^- + k_{2af}\, I_2/H - k_{2ar}\, HOI\, I^- + k_{2bf}\, I_2 - k_{2br}\, H_2OI^+\, I^- \\
&\quad - k_3\, HClO_2\, I^- - k_6\, HOCl\, I^- - k_{8f}\, HIO_2\, I^-\, H + k_{8r}\, HOI^2 \\
&\quad + k_{10}\, HIO_2\, H_2OI^+ + k_0\, INI^- - k_0\, TI^- \\
\frac{\mathrm{d}[TI_2]}{\mathrm{d}t} &= 0.5\, k_1\, ClO_2\, I^- - k_{2af}\, I_2/H + k_{2ar}\, HOI\, I^- - k_{2bf}\, I_2 + k_{2br}\, H_2OI^+\, I^- \\
&\quad - k_0\, TI_2
\end{aligned}
$$

where

$$
\begin{aligned}
I^- &= TI^- - (K_{16} + TI^- + TI_2)/2 - \sqrt{(K_{16} + TI^- + TI_2)^2/4 - TI^-\, TI_2} \\
I_2 &= TI_2 - (K_{16} + TI^- + TI_2)/2 - \sqrt{(K_{16} + TI^- + TI_2)^2/4 - TI^-\, TI_2} \\
HClO_2 &= TClO_2^-\, \frac{H}{(K_{14} + H)} \\
H_2OI^+ &= THOI\, \frac{H}{(K_{15} + H)} \\
HOI &= THOI\, \frac{K_{15}}{(K_{15} + H)}
\end{aligned}
$$

molecular formula	name
$\mathbf{ClO_2}$	Chlorine Dioxide
$\mathbf{HOCl}$	Hypochlorous Acid
$\mathbf{HIO_2}$	Iodous Acid
$\mathbf{ClO_2^-}$	Chlorite Ion
$\mathbf{HOI}$	Hypoiodous Acid
$\mathbf{I^-}$	Iodide Ion
$\mathbf{I_2}$	Iodine
$\mathbf{HClO_2}$	Chlorous Acid
$\mathbf{H_2OI^+}$	Protonated Hypoiodous Acid

Table 2: The table shows the molecular formulas and names of the species occurring in the considered model of the Chlorite-Iodide reaction system.

Table 2 gives a list of the occurring species. Additionally the prefixes T and IN denote the total concentration of the respective species and the influx into the batch system. The parameters mostly denote reaction rates or binding constants, which are assumed to be constant for the experiment. For an exact derivation and explanation of the species and parameters we refer to [11]. Eventually the experimental setup can be formulated by means of the shown ODE system and the algebraic equations below.

The Chlorite-Iodide Reaction is a so-called "chemical clock" and therefore exhibits sudden rapid changes in compound concentration. This causes the mathematical ODE model to be highly non-linear, stiff and consequently increases the computational load. We use the adjoint state technique for an efficient calculation of the gradient of the objective. Further we consider only a single data set, i.e. we assume the pH-value to be constant (cf. [11]). Naturally this is likely to reduce the number of identifiable parameters. In accordance with the findings in [11], we subsequently present even sparser solutions for the single data set.

Figure 6 shows the result of the identification routine for $p = 0.7$ and $q = 1.2$. The data was generated based on the results presented in [11]. This way we obtain a reasonable size of the problem (i.e. identifying all parameters by means of the time course of the ODE species from a single experiment), with realistic parameters and known true solution and a added relative noise level of about five percent. In order to reduce the computational load, we

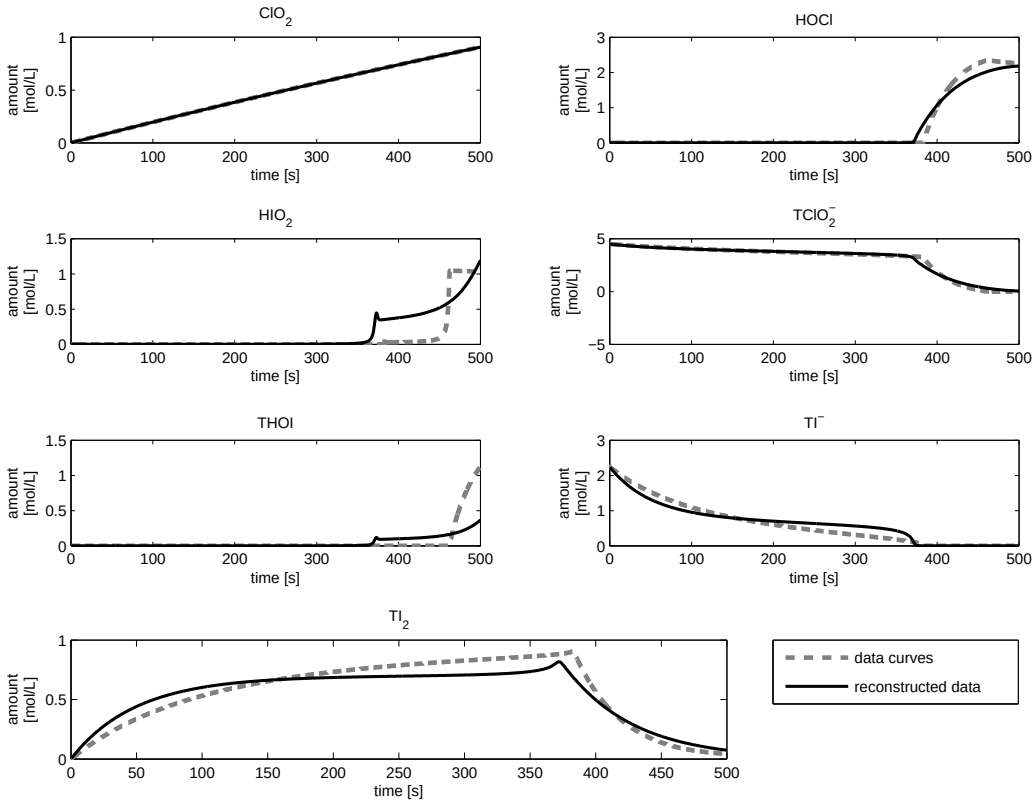

Figure 6: Concentration of the ODE species obtained by sparsity promoting regularization with $p = 0.7$ and $q = 1.2$.

first identify a parameter set by means of a standard l_2 regularization. The obtained solution is then used as an initial guess. Further the solution can be used for an efficient weighting of the regularization term for all non-zero coefficients of the l_2 solution. This is essential, as the parameters of the chemical model vary by more than 10^{20}. With exception of the second coefficient all parameters obtained by the ℓ_2 fit are non-zero and have been used for weighting.

Figure 7 shows the identified parameters, where eight out of 17 parameters are different from zero. Again we observed the enhanced sparsity promoting features by means of the regularization and of the transformation operator respectively. Figure 8 shows the progress of the iteration procedure. The size of the individual parameters decreases monotonously. One may note that the only zero component in the initial solution remains zero throughout

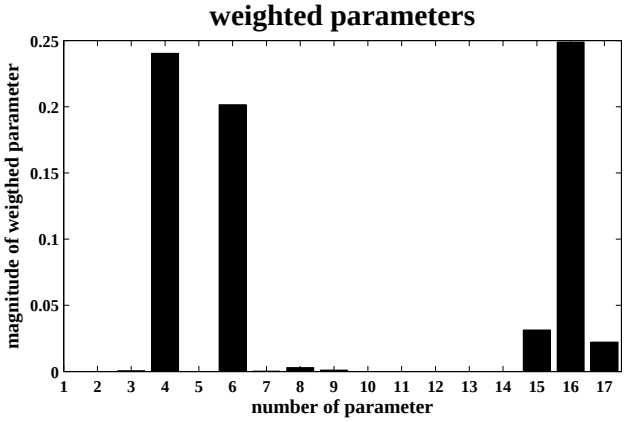

Figure 7: Parameters identified for the Chlorite-Iodide Reaction system with $p = 0.7$ and $q = 1.2$.

the iteration.

As observed for the deconvolution problem, our algorithm showed stable convergence for the non-linear problem. However, we observed an increased computational effort, due to the highly non-linear operator. In particular local minima increased the overall computation time and we have found an even increased sensitivity w.r.t. to the choice of the regularization parameter. Therefore we use a very fine grid for the different values of α ($0.9^{\{0,1,2,\ldots,\}}$). The fine grid for α and the addressed stiff and non-linear character of the considered ODE system required an efficient ODE solver. The evaluation time of the forward operator and the computation time of the gradient strongly depended on the respective parameter set and the current iterate. To reduce the computation time we used the CVODES ODE solver library from SUNDIALS. Usually the evaluation of the forward operator takes a few milliseconds up to some seconds, whereas the computation time for the gradient is typically slightly increased and lies between some milliseconds and up to several seconds. The use of the SUNDIALS package decreased the computation time about thirty percent. Moreover one may note that only non-negative values for the concentration of the species and the parameter values are realistic and acceptable. However, due to numerical artifacts negative values might occur during the ODE integration. We control this by stringent absolute error tolerances for the solver, as suggest by the developers of the solver library.

In summary we can conclude that the proposed algorithm provides a reasonable extension of the surrogate approach for non-convex sparsity pro-

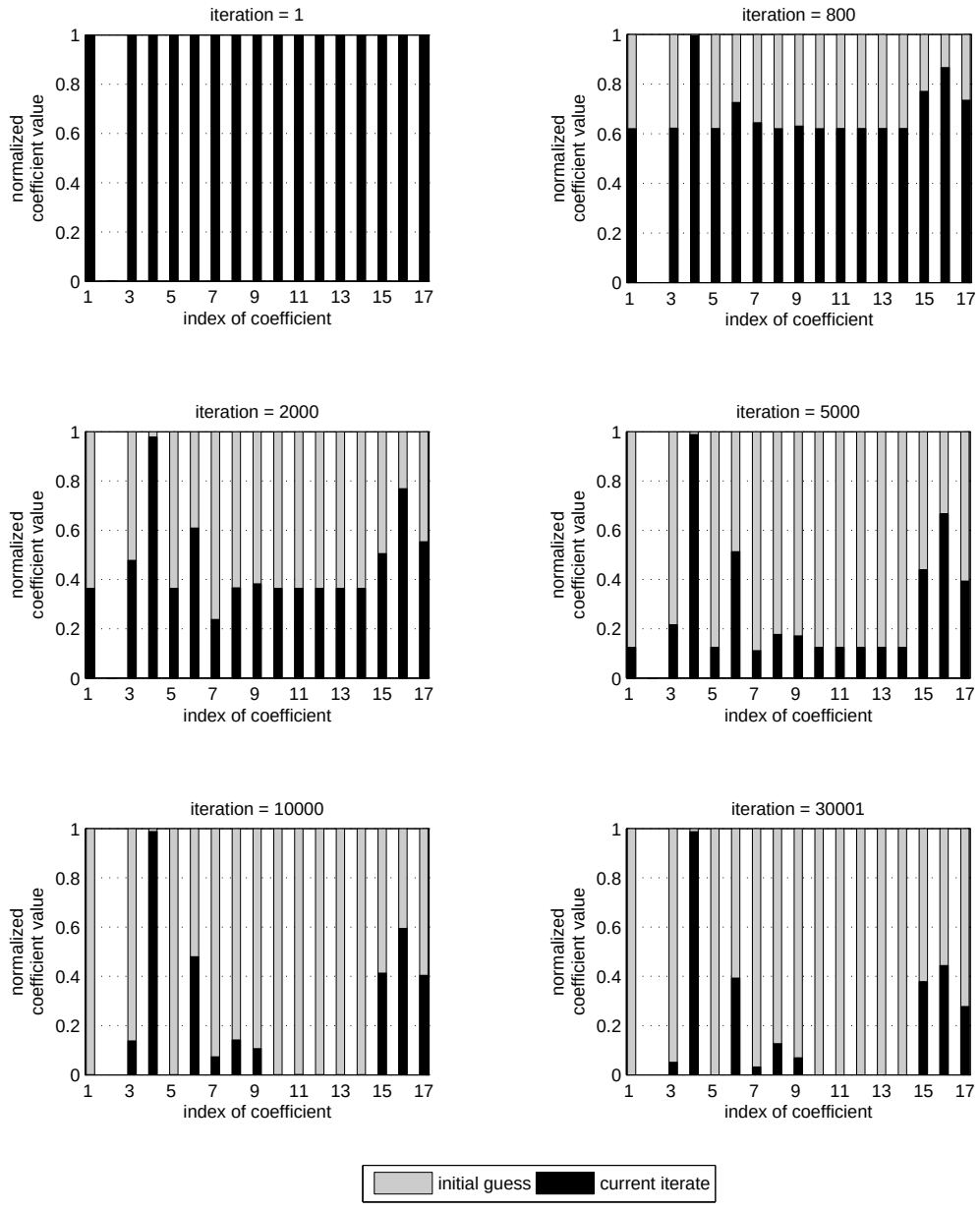

Figure 8: The six bar charts show the progress of the solution for the iterations 1, 800, 2000, 5000, 10000, 30001 for the considered example with $p = 0.7$ and $q = 1.2$. Only normalized values of coefficients greater than 10^{-6} are plotted. Bars of the current iterate are additionally scaled w.r.t. the initial solution.

moting regularization terms in sequence spaces. It was successfully applied to the deconvolution problem, leading to a linear operator equation, as well as to the parameter identification problem with a highly non-linear operator. In both cases the strong sparsity promoting feature was observed. Moreover we showed that the technique of the transformation operator potentially allows to transform the ℓ_p regularization problem for $p < 1$ to a general ℓ_q problem with $q \geq 1$. This is especially of interest as numerous techniques for ℓ_q regularization with $q \geq 1$ exist, which can then be utilized. Particularly, methods which have already been shown to have sparsity promoting features (e.g. ℓ_1 regularization) provide attractive alternatives. The transformation operator technique then would act as a sparsity enhancing map. For our future work we plan to investigate those possibilities and analyze the impact of the transformation operator.

Acknowledgments

R. Ramlau acknowledges support by the FWF Grants P19496-N18 and P20237-N14. C. A. Zarzer thanks H. W. Engl for his support and supervision and P. Kügler for his support and many fruitful discussions. C. A. Zarzer acknowledges support by the WWTF Grant MA07-030.

References

[1] D. Donoho, P. Stark, Uncertainty principles and signal recovery, SIAM J. Appl. Math. 49 (1989) 906–931.

[2] D. Donoho, J. Tanner, Sparse nonnegative solutions of underdetermined linear equations by linear programming, Proc. Nat. Acad. Sci. 102 (2005) 9446–9451.

[3] E. Candes, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math 59 (2006) 1207–1223.

[4] E. Candes, T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (2005) 4203–4215.

[5] D. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 52 (2006) 1289–1306.

[6] D. Donoho, High-dimensional centrally symmetric polytypes with neighborliness proportional to dimension, Discrete Comput. Geom. 35 (2006) 617–652.

[7] R. Gribonval, M. Nielsen, Highly sparse representations from dictionaries are unique and independent of the sparseness measure, Appl. Comput. Harmon. Anal. 22 (2007) 335–355.

[8] E. Candes, T. Tao, Near optimal signal recovery from random projections: universal encoding strategies?, IEEE Trans. Inform. Theory 52 (2006) 5406–5425.

[9] A. Cohen, W. Dahmen, R. DeVore, Compressed sensing and best k-term approximation, J. Amer.Math. Soc. 22 (2009) 211–231.

[10] I. Daubechies, R. Devore, M. Fornasier, S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, Report (2009).

[11] P. Kügler, E. Gaubitzer, S. Müller, Parameter identification for chemical reaction systems using sparsity enforcing regularization: A case study for the chlorite–iodide reaction, The Journal of Physical Chemistry A 113 (2009) 2775–2785.

[12] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math. 51 (2004) 1413–1541.

[13] R. Ramlau, G. Teschke, Tikhonov replacement functionals for iteratively solving nonlinear operator equations, Inverse Problems 21 (2005) 1571–1592.

[14] R. Ramlau, G. Teschke, A thresholding iteration for nonlinear operator equations with sparsity constraints, Numer. Math. 104 (2006) 177–203.

[15] R. Ramlau, G. Teschke, An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector-valued regimes and an application to color image inpainting., Inverse Problems 23 (2007) 1851–1870.

[16] M. Fornasier, R. Ramlau, G. Teschke, The application of joint sparsity and total variation minimization algorithms to a real-life art restoration problem, Adv. Comput. Math. 31 (2009) 157–184.

[17] K. Bredies, D. Lorenz, P. Maass, A generalized conditional gradient method and its connection to an iterative shrinkage method, Comput. Optim. Appl. 42 (2008) 173–193.

[18] T. Bonesky, K. Bredies, D. A. Lorenz, P. Maass, A generalized conditional gradient method for nonlinear operator equations with sparsity constraints, Inverse Problems 23 (2007) 2041–2058.

[19] R. Griesse, D. Lorenz, A semismooth newton method for tikhonov functionals with sparsity constraints, Inverse Problems 24 (2008) 035007.

[20] J. Bect, L. Blanc Feraud, G. Aubert, A. Chambolle, A/1-unified variational framework for image restoration, pp. Vol IV: 1–13.

[21] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Model. Simul. 4 (2005) 1168–1200 (electronic).

[22] K. Bredies, D. A. Lorenz, Iterated hard shrinkage for minimization problems with sparsity constraints, SIAM J. Sci. Comput. 30 (2008) 657–683.

[23] R. Ramlau, Regularization properties of tikhonov regularization with sparsity constraints, ETNA 30 (2008) 54–74.

[24] R. Ramlau, E. Resmerita, Convergence rates for regularization with sparsity, ETNA 37 (2010) 87–104.

[25] M. Grasmair, M. Haltmeier, O. Scherzer, Sparse regularization with lq penalty term, Inverse Problems 24 (2008) 055020 (13pp).

[26] D. A. Lorenz, Convergence rates and source conditions for Tikhonov regularization with sparsity constraints, J. Inverse Ill-Posed Probl. 16 (2008) 463–478.

[27] M. Grasmair, Well-posedness and convergence rates for sparse regularization with sublinear l^q penalty term, Inverse Probl. Imaging 3 (2009) 383–387.

[28] C. A. Zarzer, On Tikhonov regularization with non-convex sparsity constraints, Inverse Problems 25 (2009) 025006.

[29] M. Nikolova, Markovian reconstruction in computed imaging and fourier synthesis, in: ICIP (2), pp. 690–694.

[30] R. Pytlak, Conjugate gradient algorithms in nonconvex optimization, volume 89 of *Nonconvex Optimization and its Applications*, Springer-Verlag, Berlin, 2009.

[31] R. Ramlau, G. Teschke, Sparse recovery in inverse problems, in: M. fornasier, editor, *theoretical foundations and numerical methods for sparse recovery*, Radon Series on Computational and Applied Mathematics 9, De Gruyter, Berlin, 2010 (2010).

[32] R. Ramlau, A steepest descent algorithm for the global minimization of the Tikhonov– functional, Inverse Problems 18 (2002) 381–405.

[33] R. Ramlau, TIGRA–an iterative algorithm for regularizing nonlinear ill–posed problems, Inverse Problems 19 (2003) 433–467.

[34] H. W. Engl, G. Landl, Convergence rates for maximum entropy regularization, SIAM J. Numer. Anal. 30 (1993) 1509–1536.

[35] I. Cioranescu, Geometry of Banach spaces, duality mappings and nonlinear problems, Kluwer, Dordrecht, 1990.

[36] O. Scherzer, A modified landweber iteration for solving parameter estimation problems, Appl. Math. Optim. 38 (1998) 45–68.

[37] K. Bredies, D. Lorenz, Linear convergence of iterated soft-tresholding, J. Fourier Anal. Appl. 14 (2008) 813–837.

[38] F. Bauer, S. Kindermann, Recent results on the quasi-optimality principle, J. Inverse Ill-Posed Probl. 17 (2009) 5–18.

[39] F. Bauer, S. Kindermann, The quasi-optimality criterion for classical inverse problems, Inverse Problems 24 (2008) 035002, 20.

[40] S. W. Anzengruber, R. Ramlau, Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators, Inverse Problems 26 (2010) 025001, 17.

[41] J. M. Bioucas-Dias, M. A. T. Figueiredo, A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration, IEEE Trans. Image Process. 16 (2007) 2992–3004.

[42] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009) 183–202.

[43] I. Loris, M. Bertero, C. De Mol, R. Zanella, L. Zanni, Accelerating gradient projection methods for l_1-constrained signal recovery by steplength selection rules, Appl. Comput. Harmon. Anal. 27 (2009) 247–254.

[44] H. W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, P. Schuster, Inverse problems in systems biology, Inverse Problems 25 (2009) 123014.