Trends in Mathematics

Bernd Hofmann
Antonio Leitão
Jorge P. Zubelli
Editors

New Trends in Parameter Identification for Mathematical Models

Birkhäuser

Birkhäuser

Trends in Mathematics

Trends in Mathematics is a series devoted to the publication of volumes arising from conferences and lecture series focusing on a particular topic from any area of mathematics. Its aim is to make current developments available to the community as rapidly as possible without compromise to quality and to archive these for reference.

Proposals for volumes can be submitted using the Online Book Project Submission Form at our website www.birkhauser-science.com.

Material submitted for publication must be screened and prepared as follows: All contributions should undergo a reviewing process similar to that carried out by journals and be checked for correct use of language which, as a rule, is English. Articles without proofs, or which do not contain any significantly new results, should be rejected. High quality survey papers, however, are welcome.

We expect the organizers to deliver manuscripts in a form that is essentially ready for direct reproduction. Any version of TEX is acceptable, but the entire collection of files must be in one particular dialect of TEX and unified according to simple instructions available from Birkhäuser.

Furthermore, in order to guarantee the timely appearance of the proceedings it is essential that the final version of the entire material be submitted no later than one year after the conference.

More information about this series at http://www.springer.com/series/4961

Bernd Hofmann • Antonio Leitão • Jorge P. Zubelli
Editors

New Trends in Parameter Identification for Mathematical Models

Birkhäuser

Editors

Bernd Hofmann
Fakultät für Mathematik
Technische Universität Chemnitz
Chemnitz, Germany

Antonio Leitão
Department of Mathematics
Federal University of Santa Catarin
Florianopolis, Brazil

Jorge P. Zubelli
Pura e Aplicada
Instituto Nacional de Matematica
Rio de Janeiro
Rio de Janeiro, Brazil

ISSN 2297-0215 ISSN 2297-024X (electronic)
Trends in Mathematics
ISBN 978-3-319-70823-2 ISBN 978-3-319-70824-9 (eBook)
https://doi.org/10.1007/978-3-319-70824-9

Library of Congress Control Number: 2018930337

Mathematics Subject Classification (2010): 35R30, 47A52, 47J06, 49N45, 65J20, 65F22, 65N21, 65R32

Printed on acid-free paper

This book is published under the trade name Birkhäuser, www.birkhauser-science.com
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The Proceedings volume contains a collection of 16 contributions, written by experts in the field of Inverse Problems in preparation and in the context of the IMPA conference "New Trends in Parameter Identification for Mathematical Models," Rio de Janeiro, Brazil, Oct 30–Nov 3, 2017, integrating the "Chemnitz Symposium on Inverse Problems on Tour." One aim of the conference was to foster the scientific collaboration between mathematicians and engineers from the Brazilian, American, European, and Asian communities. This conference has been part of the "Thematic Program on Parameter Identification in Mathematical Models" organized at IMPA in October and November 2017. The goal of this thematic program was to bring together leading scientists in Numerical Analysis and Mechanical Engineering, being all specialists in inverse problems, for a two-months period at IMPA in order to present a perspective concerning the current trends and to support the disciplinary and interdisciplinary collaboration between researchers of the diverse communities.

The contributions of this volume, which are original research papers with a high degree of novelty, have their focus on the following topics:

- Regularization methods for the stable approximate solution of ill-posed operator equations in Hilbert and Banach spaces, modeling linear and nonlinear inverse problems with applications in natural sciences and engineering
- Error analysis, regularization parameter choice, and convergence rates for a number of regularization approaches (variational regularization, iterated Tikhonov regularization, sparsity-promoting regularization, regularization by discretization, and ADMM)
- Problems of tomography (EIT, SPECT, terahertz tomography, and spherical surface wave tomography)
- Iterative regularization methods for inverse problems

- Novel methods for parameter identification in partial differential equations and integral equations
- Linear statistical inverse problems and Bayesian inverse problems

Chemnitz, Germany	Bernd Hofmann
Florianopolis, Brazil	Antonio Leitão
Rio de Janeiro, Brazil	Jorge P. Zubelli
November 2017	

Contents

Posterior Contraction in Bayesian Inverse Problems Under Gaussian Priors

Sergios Agapiou and Peter Mathé

Abstract We study Bayesian inference in statistical linear inverse problems with Gaussian noise and priors in a separable Hilbert space setting. We focus our interest on the posterior contraction rate in the small noise limit, under the frequentist assumption that there exists a fixed data-generating value of the unknown. In this Gaussian-conjugate setting, it is convenient to work with the concept of squared posterior contraction (SPC), which is known to upper bound the posterior contraction rate. We use abstract tools from regularization theory, which enable a unified approach to bounding SPC. We review and re-derive several existing results, and establish minimax contraction rates in cases which have not been considered until now. Existing results suffer from a certain saturation phenomenon, when the data-generating element is too smooth compared to the smoothness inherent in the prior. We show how to overcome this saturation in an empirical Bayesian framework by using a non-centered data-dependent prior.

1 Setup

We consider the following linear equation in real Hilbert space

$$y^\delta = Kx + \delta\eta,$$

where $K\colon X \to Y$ is a linear operator acting between the real separable Hilbert spaces X and Y, $\eta \sim \mathcal{N}(0, \Sigma)$ is an additive centered Gaussian noise, and $\delta > 0$ is a scaling constant modeling the size of the noise. Here, the covariance operator

S. Agapiou
Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus
e-mail: Agapiou.Sergios@ucy.ac.cy

P. Mathé (✉)
Weierstraß Institute for Applied Analysis and Stochastics, Berlin, Germany
e-mail: peter.mathe@wias-berlin.de

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_1

$\Sigma : Y \to Y$ is a self-adjoint and positive definite bounded linear operator. We formally pre-whiten this equation and get

$$z^\delta = \Sigma^{-1/2} y^\delta = \Sigma^{-1/2} K x + \delta \xi,$$

where now $\xi \sim \mathcal{N}(0, I)$ is Gaussian white noise. We assign $T := \Sigma^{-1/2} K$, and assume that this is bounded by imposing the condition $\mathscr{R}(K) \subset \mathscr{D}(\Sigma^{-1/2})$. We hence arrive to the data model

$$z^\delta = Tx + \delta \xi. \tag{1}$$

This model is to be understood in a weak sense. For each functional $b \in Y$ we have that the real valued random variable $\langle z^\delta, b\rangle$ is Gaussian $\mathcal{N}(\langle Tx, b\rangle, \delta^2 \|b\|^2)$. In this study we consider the Bayesian approach to the statistical inverse problem of finding x from the observation z^δ. We assume Gaussian priors on x, distributed according to $\mathcal{N}(0, \frac{\delta^2}{\alpha} C_0)$, where $C_0 : X \to X$ is a positive definite, self-adjoint and trace class linear operator, and $\alpha > 0$ is a scaling constant. Linearity suggests that the posterior is also Gaussian and in this paper we are interested in the asymptotic performance of the posterior in the small noise limit, $\delta \to 0$. Actually, it is well known that the posterior distribution is a tight Gaussian probability on X provided that the prior distribution was Gaussian and the noise ξ is a generalized Gaussian element, as this is assumed in (1), we refer to [14].

1.1 Squared Posterior Contraction

Consider a frequentist setting, in which we observe data z^δ generated from the model (1) for a fixed underlying true element $x^* \in X$ and corresponding to a noise level δ. It is then reasonable to expect that for small δ and for appropriate values of α, the posterior Gaussian distribution will concentrate around the true data-generating element x^*. As we discuss below, this concentration will be driven by the following function.

Squared posterior contraction (SPC):

$$\mathrm{SPC} := \mathbb{E}^{x^*} \mathbb{E}_\alpha^{z^\delta} \|x^* - x\|^2, \quad \alpha, \delta > 0. \tag{2}$$

Here, the outward expectation is taken with respect to the data generating distribution, that is, the distribution generating z^δ when x^* is given, and the inward expectation is taken with respect to the posterior distribution, given data z^δ and having chosen a parameter α. The Gaussian posterior distribution has a posterior mean, say $x_\alpha^\delta = x_\alpha^\delta(z^\delta; \alpha)$, and a posterior covariance, say $C^\delta(\alpha)$, which is independent from the data z^δ, and thus deterministic. Then the inner expectation obeys the usual bias-variance decomposition

$$\mathbb{E}_\alpha^{z^\delta} \|x^* - x\|^2 = \left\|x^* - x_\alpha^\delta\right\|^2 + \operatorname{tr}\left[C^\delta(\alpha)\right].$$

Applying the expectation with respect to the data-generating distribution, we obtain that

$$\mathbb{E}^{x^*}\mathbb{E}_\alpha^{z^\delta} \|x^* - x\|^2 = \mathbb{E}^{x^*} \left\|x^* - x_\alpha^\delta\right\|^2 + \operatorname{tr}\left[C^\delta(\alpha)\right].$$

The quantity $\mathbb{E}^{x^*} \left\|x^* - x_\alpha^\delta\right\|^2$ represents the mean integrated squared error (MISE) of the posterior mean viewed as an estimator of x^*, and it has again a bias-variance decomposition into squared bias $b_{x^*}^2(\alpha) := \left\|x^* - \mathbb{E}^{x^*} x_\alpha^\delta\right\|^2$ and estimation variance $V^\delta(\alpha) := \mathbb{E}^{x^*} \left\|x_\alpha^\delta - \mathbb{E}^{x^*} x_\alpha^\delta\right\|^2$. We have thus decomposed the squared posterior contraction into respectively the *squared estimation bias*, the *estimation variance*, and the *spread in the posterior distribution*

Decomposition of the SPC:

$$\mathrm{SPC}(\alpha, \delta) = b_{x^*}^2(\alpha) + V^\delta(\alpha) + \operatorname{tr}\left[C^\delta(\alpha)\right]. \tag{3}$$

We emphasize here, that the decomposition remains valid in the more general case of non-centered Gaussian priors.

First, how do the estimation variance $V^\delta(\alpha)$ and the posterior spread $\operatorname{tr}\left[C^\delta(\alpha)\right]$ relate? In previous studies, these quantities appear to be either of the same order, see proof of [11, Thm 4.1], or the posterior spread dominates the estimation variance, see proofs of [3, Thm 4.3] and [12, Thm 2.1].

The *posterior contraction rate* is concerned with the concentration rate of the posterior distribution around the truth, in the small noise limit $\delta \to 0$, and given a prior distribution. This rate is measured by the size of the smallest shrinking balls around the data generating true element, that contain most of the posterior probability, see [8]. In the assumed linear Gaussian-conjugate setting, it is well known that the square root of the convergence rate of SPC is a posterior contraction rate (see for example [1, Section 7]). Given the prior scaling assumed here, SPC decays to zero provided that the parameter α is chosen such that $\alpha = \alpha(\delta) \to 0$ in an appropriate manner. It is desirable that the rate of decay is optimal in the minimax sense for a data-generating element x^* of a certain smoothness class[1] (see [4] for a

[1] When considering the SPC uniformly over some class of inputs x^* then if follows from (3) that the best (uniform) contraction rate cannot be better than the corresponding minimax rate for statistical estimation.

review on the minimax theory for non-parametric statistical inverse problems, and the recent note [6]).

The study of this decay was the subject of the papers [1, 3, 11, 12] (see also [10, 18, 20] for results in more general, non Gaussian-conjugate settings). The obtained rates of convergence depend on the relationship between the regularity of the data-generating element x^* and the (maximal) regularity inherent in the prior (see [5, § 2.4] for details on the regularity of draws from Gaussian measures in Hilbert space). The general message is that if the prior regularity matches the regularity of x^*, then the convergence rate of SPC is the minimax-optimal rate even without rescaling the prior, that is for the scaling considered here, α should be chosen to be equal to δ^2. If there is a mismatch between the prior regularity and the regularity of the truth, then the minimax rate can be achieved by appropriately rescaling the prior. If the prior is smoother than the truth, then there exists an a priori parameter choice rule $\alpha = \alpha(\delta)$ such that $\frac{\delta^2}{\alpha} \to \infty$ as $\delta \to 0$, which gives the optimal rate. If however the prior is rougher than the truth, then the minimax rate can be achieved by appropriate choices $\alpha = \alpha(\delta)$ such that $\frac{\delta^2}{\alpha} \to 0$ as $\delta \to 0$, in general only up to a maximal smoothness of x^*. For true data-generating elements with smoothness higher than that maximal one, the achieved rate is suboptimal. As quoted in [11], rescaling can make the prior arbitrarily 'rougher' but not arbitrarily 'smoother'. A closer look at the situation reveals, and we shall highlight this in our subsequent analysis, that the estimation bias, which is part of the SPC in (3), is responsible for this phenomenon. Bounds for the bias depend on the inter-relation between the underlying solution smoothness and the capability of the chosen (Tikhonov-type since we have Gaussian priors) reconstruction by means of x_α^δ to take it into account. The capability of such a scheme to take smoothness into account is called *qualification* of the scheme, whereas the limited decay rate of the bias, as $\alpha \to 0$, due to the chosen reconstruction scheme, is called *saturation* of the scheme. Details will be given below. In Sect. 4 we shall review results known so far. But the present approach (using general link conditions, and also general smoothness) entails to derive these results in a unified manner. Moreover, we can establish results for settings which have not been known beforehand, some of them carry features, not expected beforehand.

From a statistical point of view, it is desirable to use priors which achieve the minimax-optimal rate for x^* in a range of smoothness classes, without the a priori knowledge of the exact smoothness of x^*. Such priors are called *rate adaptive*. The results referenced in the above paragraph show that Gaussian priors are not rate adaptive over e.g. the Sobolev smoothness class, but also suggest ways of overcoming this. In the literature there have been two strategies of building on Gaussian priors to obtain more elaborate priors that are rate adaptive. The first one, see e.g. in [19], is to attempt to learn the correct scaling from the data, either by using a maximum likelihood empirical Bayes approach, or by a fully hierarchical approach. The obtained results show that the minimax rate is achieved but unsurprisingly only up to a maximal regularity of the truth (what is actually surprising, is that this maximal regularity is smaller than the one for the oracle type choice of α). In statistical language, the corresponding priors are rate adaptive but only in a range of smoothness classes; they are not fully rate adaptive. The second strategy, see e.g. in [13], is to not rescale the prior but rather attempt to learn the correct regularity from the data, again either using a maximum likelihood empirical Bayes or a fully hierarchical approach. Indeed, the obtained results show that the minimax rate is achieved by both of the approaches, hence the corresponding priors are fully rate adaptive.

From a computational point of view, both of the above mentioned approaches to rate adaptivity can be difficult to implement. On the one hand as it is shown in [2], the implementation of the hierarchical approach in non-trivial problems is problematic in high dimensions and for small noise (it is much more difficult in the case of learning the regularity compared to the case of learning the scaling), while on the other hand the above empirical Bayes approaches involve solving an optimization problem which also becomes difficult for non-trivial problems.

Hence, another focus of this study is to discuss a simple way to overcome the saturation effect, which in turn will open up the possibility of designing other empirical Bayes approaches which are fully rate adaptive.

1.2 Paradigm

We consider the following alternative paradigm. Suppose we want to use a Gaussian prior with covariance C_0, and prior mean m_0 to perform posterior inference for the problem (1). The question we address is whether the prior center m_0 has a significant impact on the posterior contraction rate, and if so, how to choose it 'optimally' in the presence of data. The subsequent analysis will show that the convergence rate of SPC will improve by an appropriate adjustment of the prior if the underlying solution x^* has large smoothness. In terms of the previous discussion, for a prior of fixed smoothness this enables us to make a priori choices of $\alpha = \alpha(\delta)$ such that the posterior contraction rate is minimax-optimal even for higher smoothness of x^*, by choosing an appropriate center m_0 of the prior distribution. The proposed re-centering $m_0 = m_0(z^\delta; \alpha)$ of the prior depends on the data z^δ and the parameter α, it is not static. However, it can easily be managed by a regularization step preprocessing the Bayes step. We anticipate these results in the following Fig. 1. This figure highlights the results as described in Sect. 4.1.1, where the parameters $a, p > 0$ are explained.

We capture the advantages in a few lines:

- the user may choose a (centered) Gaussian prior of arbitrary smoothness;
- after observing data z^δ, a prior center, say $m_0 = m_0(z^\delta; \alpha)$ is determined by some deterministic regularization;
- if this preprocessing regularization has enough qualification, then the posterior distribution will contract order optimally regardless of the solution smoothness.

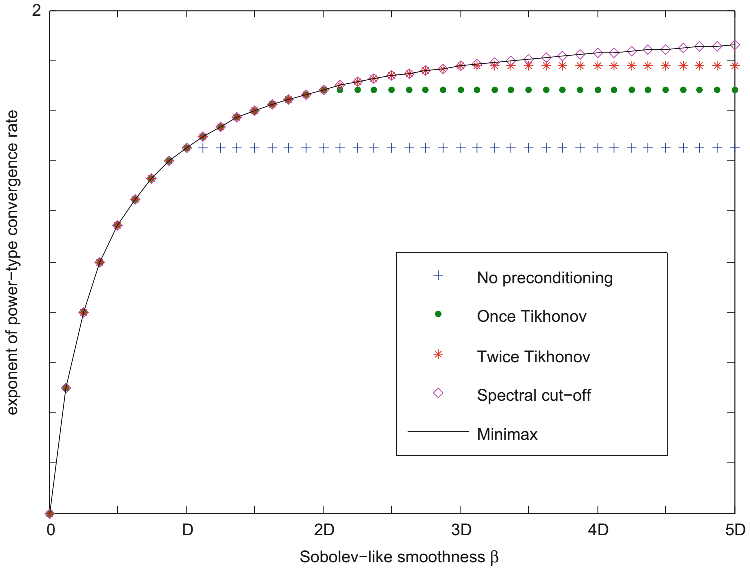


Fig. 1 Exponents of convergence rates of SPC plotted against Sobolev-like smoothness of the truth β, for different methods of choosing the prior mean m_α^δ, in the moderately ill-posed problem discussed in Sect. 4.1.1. We set $D := 1 + 2a + 2p$, the saturation point when no preconditioning of the prior mean is used. Rates calculated for $a = 0.5,\ p = 1$

If not, then the contraction rate is at least as good as the rate corresponding to a centered prior.

- this preprocessing step has no effect on the parameter choice; so *any* choice $\alpha = \alpha(\delta; z^\delta)$ which yields 'optimal' contraction without preprocessing will retain this property, and will eventually extend this optimality property for higher solution smoothness.

1.3 Outline

In order to explain the new paradigm we first study the impact of using a non-centered prior to the posterior mean and covariance. Then we specify the prior centering by means of using a linear regularization in Eq. (4), as such is known from regularization theory. Next, we provide explicit representations of the quantities involved in the subsequent analysis, the posterior mean, the posterior covariance, and formulas for the bias and estimation variance, see Eqs. (5)–(8).

The main results are given in Sect. 3, after confining ourselves to the case of commuting operators C_0 and T^*T, expressed in terms of a specific link condition. We first derive bounds for the estimation bias in Proposition 2, and these bounds are crucial for overcoming the saturation. Then we introduce the net posterior spread in

Sect. 3.3, which is the unscaled version of the posterior spread, and we highlight its properties. We then combine to obtain our main result on the convergence of SPC, which is Theorem 1.

To emphasize the significance of our results we discuss in Sect. 4 specific examples some of which were previously studied in [3, 11, 12]. In order to facilitate the reading of the study we postpone all proofs to the Appendix.

2 Setting the Pace

As mentioned above, we shall discuss a preprocessing of the prior by choosing it non-central, that is, we will introduce a shift m_0, such that the prior will be Gaussian with $\mathcal{N}(m_0, \frac{\delta^2}{\alpha}C_0)$. We are interested in understanding the impact of the shift m_0 on the convergence rate of SPC. We start with deriving (well-known) formulas for the posterior mean x_α^δ in this context.

We first recall the representation of the posterior mean m and posterior covariance C when a centered prior $\mathcal{N}(0, \frac{\delta^2}{\alpha}C_0)$ is used. In this case we know, see for example [14, 16], that almost surely with respect to the joint distribution of (x, z^δ) the posterior is Gaussian, $\mathcal{N}(m, C)$, with mean $m = C_0^{1/2}(\alpha I + B^*B)^{-1} B^* z^\delta$, and covariance $C = \delta^2 C_0^{1/2}(\alpha I + B^*B)^{-1} C_0^{1/2}$, where we define the compact operator $B := TC_0^{\frac{1}{2}}$. Re-centering the prior towards m_0 does not affect the posterior covariance C. To obtain the shift in the posterior mean we rewrite (1) as

$$z^\delta - Tm_0 = T(x - m_0) + \delta\xi$$

Thus if $x \sim \mathcal{N}(m_0, C_0)$ then $x - m_0 \sim \mathcal{N}(0, C_0)$. We are in the usual context with centered prior but new data $z^\delta - Tm_0$.

Remark 1 We fix once and for all the function $s_\alpha(t) = \alpha/(\alpha + t)$, applied to the self-adjoint operator B^*B by using spectral calculus. This is the residual function for Tikhonov regularization. This is done in order to distinguish the (Tikhonov) regularization in the posterior mean due to the use of a Gaussian prior, from the chosen regularization for the prior preconditioning.

We obtain the representation for the posterior mean (shifting back towards m_0) as

$$\begin{aligned}
x_\alpha^\delta &= m_0 + C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^*(z^\delta - Tm_0) \\
&= C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^* z^\delta + m_0 - C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^* T m_0 \\
&= C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^* z^\delta + C_0^{1/2}\left(I - \left(\alpha I + B^*B\right)^{-1} B^*B\right) C_0^{-1/2} m_0 \\
&= C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^* z^\delta + C_0^{1/2} s_\alpha(B^*B) C_0^{-1/2} m_0.
\end{aligned}$$

It is well-understood from previous Bayesian analyses that a static choice of m_0 will not have impact on the posterior contraction. However, within our new paradigm we choose any regularization scheme g_α and assign the prior center as

$$m_0(z^\delta;\alpha) := m_\alpha^\delta = C_0^{1/2} g_\alpha(B^*B)B^*z^\delta. \tag{4}$$

We introduce linear regularization schemes as follows, cf. [7, 9].

Definition 1 (Linear Regularization) Let $b = \|B^*B\|$. A family of piecewise continuous functions $g_\alpha\colon (0,b] \to \mathbb{R}$, $\alpha > 0$, is called *regularization filter* with residual function $r_\alpha(t) = 1 - tg_\alpha(t),\ \alpha, 0 < t \le b$, if

1. $\sup_{0<t\le b} |r_\alpha(t)| \le \gamma_0$, for all $\alpha > 0$,
2. $\lim_{\alpha\to 0} r_\alpha(t) = 0$ for each $0 < t \le b$, and
3. $\sup_{0<t\le b} |g_\alpha(t)| \le \gamma_*/\alpha$, for all $\alpha > 0$.

Remark 2 The last assertion in Definition 1 is actually stronger than the one required in [9], but it is a convenient strengthening, and most known regularization schemes obey this stronger bound.

Remark 3 We use the following convention: if no preconditioning is used, that is, if $g_\alpha(t) \equiv 0$, then we assign the constant function $r_\alpha(t) \equiv 1$, in order to simplify the comparison of the different settings. Specifically, without preprocessing we would naturally (and statically) use $m_0 := 0$ as the prior mean.

Example 1 (Tikhonov Regularization) One of the commonly used regularization schemes is Tikhonov regularization, in which case the filter g_α is given as $g_\alpha(t) = 1/(\alpha + t),\ \alpha, t > 0$. Notice that in the case $m_0 = 0$, the posterior mean has the form of the right hand side in Eq. (4) with g_α being the Tikhonov filter.

Example 2 (k-fold Tikhonov Regularization) We may iterate Tikhonov regularization, starting from the trivial element $x_{0,\alpha} = 0$ as

$$x_{j,\alpha}^\delta := x_{j-1,\alpha}^\delta + \left(\alpha I + B^*B\right)^{-1} B^* \left(z^\delta - Bx_{j-1,\alpha}^\delta\right), \quad j = 1, \dots, k.$$

For $k = 1$ this gives Tikhonov regularization. The resulting linear regularization is given by the function $g_{k,\alpha} := \frac{1}{t}\left(1 - \left(\frac{\alpha}{\alpha+t}\right)^k\right),\ t > 0$, with corresponding residual function $r_{k,\alpha} = \left(\frac{\alpha}{\alpha+t}\right)^k,\ t > 0$. This regularization results in the prior center $m_\alpha^\delta = C_0^{1/2} x_{k,\alpha}^\delta = C_0^{\frac{1}{2}} g_{k,\alpha}(B^*B)B^*z^\delta$.

Example 3 (Spectral Cut-Off, Truncated SVD) This is a versatile scheme, which requires to know the singular value decomposition of the underlying operator. If this is available, then we let $g_\alpha(t) = 1/t$, for $t \ge \alpha$ and $g_\alpha(t) = 0$ else.

We summarize the previous considerations and fix the notation which will be used subsequently. Given prior mean m_α^δ from (4), we have that the posterior distribution is Gaussian with the following posterior mean and posterior covariance.

Posterior mean:

$$x_\alpha^\delta = C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^* z^\delta + C_0^{1/2} s_\alpha(B^*B) C_0^{-1/2} m_\alpha^\delta, \tag{5}$$

Posterior covariance $C := C^\delta(\alpha)$:

$$C^\delta(\alpha) = \delta^2 C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} C_0^{1/2}. \tag{6}$$

Recall the decomposition of the SPC from (3). We have that the spread is given as $\operatorname{tr}\left[C^\delta(\alpha)\right]$. We next give expressions for the corresponding estimation bias and estimation variance.

Lemma 1 *Let x_α^δ be as in (5). Then the estimation bias and estimation variances, with posterior mean as estimator, are*

$$b_{x^*}(\alpha) = \left\| C_0^{1/2} s_\alpha(B^*B) r_\alpha(B^*B) C_0^{-1/2} x^* \right\|, \quad \alpha > 0, \tag{7}$$

and

$$V^\delta(\alpha) = \delta^2 \operatorname{tr}\left[\left(I + \alpha g_\alpha(B^*B)\right)^2 \left(\alpha I + B^*B\right)^{-2} B^*BC_0\right],\ \alpha, \delta > 0, \tag{8}$$

respectively.

Proposition 1 *Let the prior center be obtained from any regularization (with corresponding constant γ_*). Then we have that*

$$V^\delta(\alpha) \leq (1+\gamma_*)^2 \operatorname{tr}\left[C^\delta(\alpha)\right]. \tag{9}$$

Consequently we have that

$$\mathbb{E}^{x^*}\left\|x^* - x_\alpha^\delta\right\|^2 \leq \mathrm{SPC}(\alpha,\delta) \leq b_{x^*}^2(\alpha) + \left(1 + (1+\gamma_*)^2\right) \operatorname{tr}\left[C^\delta(\alpha)\right].$$

Remark 4 The above analysis extends the previous bound from [15, Prop 2] to the present context (note that without preprocessing we have that $\gamma_* = 0$). We also note that the decay of the squared posterior contraction cannot be faster than the minimax error for statistical estimation.

We thus have that in order to (asymptotically) bound the squared posterior contraction, we only need to establish bounds for the bias and the posterior spread.

3 Assumptions and Main Results

We are now ready to present our main results. Before we do so, in Sect. 3.1 we introduce several concepts used in our formulation. First, we introduce *link conditions*, relating the two operators appearing in the setting at hand. Then we introduce *source sets*, which we use for expressing the regularity of the truth. Finally, we introduce the *qualification* of a regularization which quantifies its capability to take high smoothness into account. We then present our bounds for the bias and the posterior spread in Sects. 3.2 and 3.3 respectively. In Sect. 3.4 we present a priori bounds for the squared posterior contraction.

3.1 Link Conditions, Source Sets and Qualification

We call a function $\varphi\colon (0,\infty) \to \mathbb{R}^+$ an *index function* if it is a continuous non-decreasing function which can be extended to take the value zero at the origin.

Remark 5 The property of interest of an index function is its asymptotic behavior near the origin. In some cases the 'native' index function is not defined on $(0,\infty)$, but only on some sub-interval, say $(0,\bar{t})$. Consider for example the logarithmic function $\varphi(t) = \log^{-\mu}(1/t),\ 0 < t < \bar{t} = 1$ with $\phi(0) = 0$. Then one can extend the function ϕ at some interior point $0 < t_0 < \bar{t}$ in an increasing way, for instance as $\varphi(t) = \varphi(t_0) + (t - t_0),\ t \geq t_0$. By doing so we ensure that the extended function shares the same asymptotic properties near zero, that is, as $t \searrow 0$. In all subsequent (asymptotic) considerations it suffices to have such extensions, and this will not be mentioned explicitly.

To simplify the outline of the study we confine ourselves to commuting operators C_0 and T^*T. Specifically we do this as follows.

Assumption 1 (Link Condition) *There is an index function ψ such that*

$$\psi^2(C_0) = T^*T. \tag{10}$$

Along with the function ψ we introduce the function

$$\Theta_{\psi}(t) := \sqrt{t}\psi(t), \quad t > 0. \tag{11}$$

We draw the following consequence.

Lemma 2 *Let ψ be the index function for which Assumption 1 holds. Then the operators C_0 and T^*T commute. Moreover we have that*

$$\Theta_{\psi}^2(C_0) = B^*B.$$

Following the last lemma, we set

$$f(s) := \left(\left(\Theta_{\psi}^2\right)^{-1}(s)\right)^{1/2}, \quad s > 0. \tag{12}$$

We stress that the function f is an index function, since the function Θ_{ψ} was one. Moreover, the function Θ_{ψ}^2 is strictly increasing, such that its inverse is a well defined strictly increasing index function. Finally, as can be drawn from Lemma 2, we have that under Assumption 1 it holds

$$C_0^{1/2} = f(B^*B). \tag{13}$$

Remark 6 We remark the following about Assumption 1.

- The case that the operator T is the identity is not covered by this assumption. This would require the function $\psi \equiv 1$, which does not constitute an index function. However, for the subsequent analysis we shall only use Lemma 2. As seen from (13) we obtain that $\Theta_{\psi}(t) = \sqrt{t},\ t > 0$, in this case.
- If the prior C_0 has eigenvalues with multiplicities higher than one, then by Assumption 1 the operator T^*T also needs to have eigenvalues with higher multiplicities, since taking functions of operators preserves or increases the multiplicities of the eigenvalues. This is not realistic, hence one should choose a prior covariance with eigenvalues of multiplicity one. This can be achieved by a slight perturbation of the original choice.

In order to have a handy notation we agree to introduce the following partial ordering between index functions.

Notation 1 *Let f, g be index functions. We say that $f \prec g$ if the quotient g/f is non-decreasing. In other words $f \prec g$ if g decays to zero faster than f.*

For bounding the bias below we shall assume that the smoothness of the underlying true data-generating element x^*, is given as a source set with respect to C_0.

Definition 2 (Source Set) There is an index function φ such that

$$x^* \in A_\varphi := \{x, \quad x = \varphi(C_0)w, \ \|w\| \le 1\} .$$

The element $w \in X$ is called *source element*.

By Lemma 2 the source set A_φ can be rewritten as

$$A_\varphi = \left\{x, \quad x = \varphi(f^2(B^*B))w, \ \|w\| \le 1\right\} ,$$

with the function f from (12). Furthermore, under Assumption 1 the operators C_0 and B^*B commute, and hence the bias representation from (7) simplifies to

$$b_{x^*}(\alpha) = \|r_\alpha(B^*B)s_\alpha(B^*B)x^*\| . \tag{14}$$

Overall, if $x^* \in A_\varphi$ then

$$b_{x^*}(\alpha) \le \left\|r_\alpha(B^*B)s_\alpha(B^*B)\varphi(f^2(B^*B))\right\| = \sup_{0<t\le\|B^*B\|} |r_\alpha(t)|s_\alpha(t)\varphi(f^2(t)).$$

We shall bound this in terms of the parameter $\alpha > 0$, which directs us to the notion of a *qualification* of a regularization, see [9], again.

Definition 3 (Qualification) For an index function φ, a regularization g_α has qualification φ with constant γ, if

$$|r_\alpha(t)|\varphi(t) \le \gamma\varphi(\alpha), \quad \alpha > 0, \quad 0 < t \le \|B^*B\| .$$

The following result is a well-known consequence, see [9, Prop 2.7] again, albeit important for the subsequent analysis. We shall use the partial ordering from Definition 1.

Lemma 3 *Let g_α be a regularization with index function φ as a qualification (with constant γ). If ψ is an index function for which $\psi \prec \varphi$ then ψ is also a qualification (with constant γ).*

Example 4 (Tikhonov Regularization) Tikhonov regularization has (maximal) qualification $\varphi(t) = t,\ t > 0$. Thus, if for an index function ψ we have that $\psi(t) \prec t$ then ψ is a qualification. In particular, all concave index functions are qualifications of Tikhonov regularization with constant $\gamma = 1$.

Example 5 (Spectral Cut-Off) Spectral cut-off has arbitrary qualification, since $r_\alpha(t) = 0,\ t \geq \alpha$ and $r_\alpha(t) = 1$ elsewhere. Hence

$$r_\alpha(t)\varphi(t) = 0 \leq \varphi(\alpha),\ t \geq \alpha, \quad \text{and} \quad r_\alpha(t)\varphi(t) \leq \varphi(\alpha),\ t \leq \alpha.$$

Remark 7 We immediately see from (7) that the qualification of the regularization in the bias, can be raised from t (Tikhonov regularization) to t^{k+1}, if the residual function r_α of the regularization used for preconditioning the prior mean has qualification t^k, as is the case for k-fold Tikhonov regularization, see Example 2. If preconditioning is done by spectral cut-off, then the regularization in the bias has arbitrary qualification.

3.2 Bounding the Bias

We are now ready to present our bounds for the bias.

Proposition 2 *Suppose that $x^* \in A_\varphi$, and that m_α^δ uses a regularization g_α with constant γ_0 bounding the corresponding residual function.*

1. *If $\varphi \prec \Theta_\psi^2$, then $b_{x^*}(\alpha) \leq \gamma_0\varphi\left(f^2(\alpha)\right),\ \alpha > 0$.*
2. *If $\Theta_\psi^2 \prec \varphi$ and if there was no preconditioning, then there are constants $c_1, c_2 > 0$ (depending on x^*, φ, f^2, and on $\|B^*B\|$) such that $c_1\alpha \leq b_{x^*}(\alpha) \leq c_2\alpha,\ 0 < \alpha \leq 1$.*
3. *If $\Theta_\psi^2 \prec \varphi$ and if $t \mapsto \varphi\left(f^2(t)\right)/t$ is a qualification for the regularization g_α with constant γ, then $b_{x^*}(\alpha) \leq \gamma\varphi\left(f^2(\alpha)\right),\ \alpha > 0$.*

Remark 8 We mention that the above two cases $\varphi \prec \Theta_\psi^2$ or $\Theta_\psi^2 \prec \varphi$ are nearly disjoint, with $\varphi = \Theta_\psi^2$ being the only common member. Therefore the function Θ_ψ^2 may be viewed as the *'benchmark smoothness'*. However, note that the items (1) and (3) do not exhaust all possibilities since the function $\varphi\left(f^2(t)\right)/t$ may not be a qualification for g_α (in fact it may not even be an index function).

Remark 9 We stress that the bounds in item (2) show the saturation phenomenon in the bias if no preconditioning of the prior mean is used: for any sufficiently high smoothness the bias decays with the fixed rate α. In other words, if no preconditioning of the prior is used, the best achievable rate of decay for the bias is linear. Item (3) shows that appropriate preconditioning improves things, since for high smoothness the bias decays at the superlinear rate $\varphi(f^2(\alpha))$.

3.3 The Net Posterior Spread

Here we study the posterior spread, that is, the trace of the posterior covariance from (6), which will be needed for determining the contraction rate. In order to highlight the nature of the spread in the posterior within the assumed Bayesian framework, we make the following definition, for a given equation $z^\delta = Tx + \delta\xi$, with white noise ξ, as considered in (1).

Definition 4 (Net Posterior Spread) The function

$$S_{T,C_0}(\alpha) := \operatorname{tr}\left[C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} C_0^{1/2}\right], \quad \alpha > 0,$$

is called the *net posterior spread*.

Notice that with this function we have that $\operatorname{tr}\left[C^\delta(\alpha)\right] = \delta^2 S_{T,C_0}(\alpha)$. Moreover, using the cyclic commutativity of the trace, we get that

$$S_{T,C_0}(\alpha) = \operatorname{tr}\left[\left(\alpha I + B^*B\right)^{-1} C_0\right]. \tag{15}$$

With this more convenient representation at hand, we establish some fundamental properties of the net posterior spread, which are crucial for optimizing the convergence rate of SPC in the following subsection.

Lemma 4

1. *The function* $\alpha \mapsto S_{T,C_0}(\alpha)$ *is strictly decreasing and continuous for* $\alpha > 0$.
2. $\lim_{\alpha\to\infty} S_{T,C_0}(\alpha) = 0$, *and*
3. $\lim_{\alpha\to 0} S_{T,C_0}(\alpha) = \infty$.

3.4 Main Result: Bounding the Squared Posterior Contraction

It has already been highlighted that the squared posterior contraction as given in (2) is decomposed into the sum of the squared bias, estimation variance and posterior spread, see (3). By Proposition 1 we find that

$$b_{x^*}^2(\alpha) + \delta^2 S_{T,C_0}(\alpha) \le \mathrm{SPC}(\alpha) \le b_{x^*}^2(\alpha) + \left((1+\gamma_*)^2 + 1\right)\delta^2 S_{T,C_0}(\alpha).$$

In the asymptotic regime of $\delta \to 0$, the size of SPC is thus determined by the sum $b_{x^*}^2(\alpha) + \delta^2 S_{T,C_0}(\alpha)$. In Sect. 3.2 we have established bounds for the bias. Here we just constrain to the case where, given that $x^* \in A_\varphi$, the preconditioning is such that the size of the bias is bounded by (a multiple of) $\varphi(f^2(\alpha))$, see Proposition 2. Since $b_{x^*}^2(\alpha)$ is bounded by a non-decreasing function of α which decays to zero as $\alpha \searrow 0$, while by Lemma 4 the function $S_{T,C_0}(\alpha)$ is strictly decreasing, continuous and onto the positive half-line, the SPC is 'minimized' by the choice of α which balances the bound for the squared bias and the spread. This choice clearly exists and is unique and hence we immediately arrive to our main result which holds under Assumption 1.

Theorem 1 *Let φ be any index function, and assume that item (1) or item (3) in Proposition 2 hold. Consider the equation*

$$\varphi^2(f^2(\alpha)) = \delta^2 S_{T,C_0}(\alpha). \tag{16}$$

Equation (16) is uniquely solvable, and let $\alpha_* = \alpha_*(\varphi, \delta)$ be the solution. For $x^* \in A_\varphi$ we have that $\mathrm{SPC}(\alpha_*, \delta) = \mathcal{O}(\varphi^2(f^2(\alpha_*)))$ *as $\delta \to 0$.*

In Sect. 4 we show how to apply Theorem 1 to obtain rates of contraction in specific examples. In many cases, the obtained contraction rates of the SPC correspond to known minimax rates in statistical inverse problems. This can be seen in Propositions 4, 6, 8 and 10, below.

Remark 10 As emphasized in Remark 9, if no preconditioning is used, the best rate at which the bias can decay is linear. This effect, which is called saturation (of Tikhonov regularization), was discussed in a more general context in regularization theory, and we mention the study [17].

So, if no preconditioning is present, then the left hand side in (16) at best decays as α^2. We conclude that the best rate of decay of the SPC which can be established without preconditioning is α_*^2, where α_* is obtained from balancing $\alpha^2 = \delta^2 S_{T,C_0}(\alpha)$.

4 Examples and Discussion

We now study several examples, some which are standard in the literature, and some which exhibit new features. Our aim is to demonstrate both the simplicity of our method for deriving rates of posterior contraction as well as the benefits of preconditioning the prior. We shall provide a priori rates of posterior contraction, using Theorem 1.

Before we proceed we stress the following fact, which is not so accurately spelled out in other studies. It is important to distinguish the *degree of ill-posedness of the operator T* which governs Eq. (1), and which expresses the decay of its singular numbers, from the *degree of ill-posedness of the problem*, which corresponds to the operator T *and* the solution smoothness, and thus regards the achievable contraction rate.

Degree of ill-posedness of the operator and of the problem.

1. We call operator T *moderately ill-posed* if the singular numbers decay polynomially.
2. The operator T is *severely ill-posed* if the singular numbers decay exponentially.
3. The problem is *moderately ill-posed* if the contraction rate decays at some power.
4. The problem is *severely ill-posed* if the contraction rate is less than any power.
5. The problem is *mildly ill-posed* if the contraction rate is linear in the noise level up to some logarithmic factor.

As we will see in Sect. 4.2, below, the problem can have a significantly different degree of ill-posedness than the operator T.

To be specific we make the following assumptions.

Running assumption for the examples.

1. The prior covariance C_0 has spectrum that decays as $\{j^{-(1+2a)}\},\ a > 0$.
2. The operators C_0 and T^*T are simultaneously diagonalizable in an orthonormal basis $\{e_j\}$ which is complete in X.

In the first two examples, we present posterior contraction rates under the assumption that we have the a priori knowledge that the truth belongs to the Sobolev ellipsoid

$$S^\beta = \{x \in X : \sum_{j=1}^{\infty} j^{2\beta} x_j^2 \le 1\}, \tag{17}$$

for some $\beta > 0$ and where $x_j := \langle x, e_j \rangle$. We emphasize that such ellipsoids are examples of source sets as in Definition 2. Indeed, (17) is defining the corresponding source element w, having (square summable) coefficients $w_j := j^\beta x_j,\ j = 1, 2, \ldots$

Sobolev smoothness:
Relative to C_0, the index function defining the source set A_φ in Definition 2, is in this case $\varphi(t) = t^{\frac{\beta}{1+2a}}$.

We will recover the moderately and severely ill-posed problems, as for example studied in [11], and [3, 12], respectively.

In the other two examples, we present posterior contraction rates under analytic smoothness of the truth, that is, we assume that we have the a priori knowledge that the truth belongs to the ellipsoid

$$\mathscr{A}^{\beta} = \{x \in X : \sum_{j=1}^{\infty} e^{2\beta j} x_j^2 \leq 1\}, \tag{18}$$

for some $\beta > 0$. Again, this corresponds to a source set.

Analytic-type smoothness:
Relative to C_0, the index function defining the source set A_φ in Definition 2, is in this case $\varphi(t) = \exp(-\beta t^{-\frac{1}{1+2a}})$.

To our knowledge these cases have not been studied before in a Bayesian context. First, we once more study the moderately ill-posed operator problem, which we will see that under analytic-type smoothness of the truth leads to what we call a *mildly ill-posed* problem. Then, we study a problem with severely ill-posed operator, which as we will see, under analytic-type smoothness of the truth leads to a *moderately ill-posed* problem.

We shall use the following handy symbols for describing rates.

Notation 2 *Given two positive functions $k, h : \mathbb{R}^+ \to \mathbb{R}^+$, we use $k \asymp h$ to denote that $k = \mathscr{O}(h)$ and $h = \mathscr{O}(k)$ as $s \to 0$. Furthermore, the notation $h(s) \gg k(s)$, means that $k(s) = \mathscr{O}(h(s)s^{\mu})$ as $s \to 0$ for some positive power $\mu > 0$.*

4.1 Sobolev Smoothness

4.1.1 Moderately Ill-Posed Operator

We consider the moderately ill-posed setup studied in [11], in which the operator T^*T has spectrum which decays as $\{j^{-2p}\}$ for some $p \geq 0$, and thus the singular numbers of B^*B decay as $s_j(B^*B) \asymp j^{-(1+2a+2p)}$. In the present case Assumption 1, which expresses the operator T^*T as a function of the prior covariance operator C_0, is satisfied for $\psi^2(t) = t^{\frac{2p}{1+2a}}$. Next, we find that the function Θ_ψ in (11), which expresses the operator B^*B as a function of C_0, is given as $\Theta_\psi(t) = t^{\frac{1+2a+2p}{2(1+2a)}}$, hence the benchmark smoothness is $\Theta_\psi^2(t) = t^{\frac{1+2a+2p}{1+2a}}$. Finally, we have that the function f in (12), which expresses C_0 as a function of B^*B is given by $f(s) = s^{\frac{1+2a}{2(1+2a+2p)}}$.

Bounding the Bias

We now have all the ingredients required to bound the bias. The following result is an immediate consequence of Proposition 2 and the considerations of the previous paragraph.

Proposition 3 *Suppose that $x^* \in S^\beta$, for some $\beta > 0$. Then as $\alpha \to 0$:*

1. *If $\beta \le 1 + 2a + 2p$, and independently of whether preconditioning of the prior is used or not, we have that $b_{x^*}(\alpha) = \mathscr{O}(\alpha^{\frac{\beta}{1+2a+2p}})$;*
2. *if $\beta > 1+2a+2p$ and no preconditioning of the prior is used, then $b_{x^*}(\alpha) \asymp \alpha$;*
3. *if $\beta > 1+2a+2p$ and m_α^δ uses a regularization g_α with qualification $t^{\frac{\beta-1-2a-2p}{1+2a+2p}}$, then $b_{x^*}(\alpha) = \mathscr{O}(\alpha^{\frac{\beta}{1+2a+2p}})$.*

We stress here that our contribution is item (3). In particular, item (3) implies that if we choose the prior mean m_α^δ using the k-fold Tikhonov regularization filter (cf. Example 2), which has maximal qualification t^k, then for $\beta \le (k+1)(1+2a+2p)$ we have that $b_{x^*}(\alpha) = \mathscr{O}(\alpha^{\frac{\beta}{1+2a+2p}})$, that is, the saturation in the bias is delayed. If we choose m_α^δ using the spectral cut-off regularization filter, which as we saw in Example 5 has arbitrary qualification, then for any $\beta > 0$, we have that $b_{x^*}(\alpha) = \mathscr{O}(\alpha^{\frac{\beta}{1+2a+2p}})$, that is, there is no saturation in the bias.

A Priori Bounds of the SPC

To see the impact of this result to the SPC rate, we apply Theorem 1. In order to do so, we first need to calculate the net posterior spread which in this case is such that $S_{T,C_0}(\alpha) \asymp \alpha^{-\frac{1+2p}{1+2a+2p}}$, see [11, Thm 4.1]. Concatenating we get the following result.

Proposition 4 *Suppose that $x^* \in S^\beta$, $\beta > 0$. Then as $\delta \to 0$:*

1. *if $\beta \le 1 + 2a + 2p$ and independently of whether preconditioning of the prior is used or not, for $\alpha = \delta^{\frac{2(1+2a+2p)}{1+2p+2\beta}}$ we have that $SPC = \mathscr{O}(\delta^{\frac{4\beta}{1+2\beta+2p}})$;*
2. *if $\beta > 1 + 2a + 2p$ and no preconditioning of the prior is used, then for any choice $\alpha = \alpha(\delta, \beta)$ we have that $SPC \gg \delta^{\frac{4\beta}{1+2\beta+2p}}$;*
3. *if $\beta > 1+2a+2p$ and m_α^δ uses a regularization g_α with qualification $t^{\frac{\beta-1-2a-2p}{1+2a+2p}}$, for $\alpha = \delta^{\frac{2(1+2a+2p)}{1+2p+2\beta}}$ we have that $SPC = \mathscr{O}(\delta^{\frac{4\beta}{1+2\beta+2p}})$.*

As before, our contribution is item (3), which in particular implies that if we choose the prior mean m_α^δ using the k-fold Tikhonov regularization filter, then for $\beta \le (k+1)(1+2a+2p)$ we achieve the optimal (minimax) rate $\delta^{\frac{4\beta}{1+2\beta+2p}}$, that is the saturation in the SPC is also delayed. If we choose m_α^δ using the spectral cut-off regularization filter, then for any $\beta \ge 0$ we achieve the optimal rate $\delta^{\frac{4\beta}{1+2\beta+2p}}$, that is, there is no saturation in the SPC! Note that the optimal scaling of the prior, as a function of the noise level δ, is the same whether we use preconditioning or not. We depict the findings in Fig. 1.

4.1.2 Severely Ill-Posed Operator

We now consider the severely ill-posed setup studied in [3, 12], in which the operator T^*T has spectrum which decays as $\{e^{-2qj^b}\}$ for some $q, b > 0$, and thus the singular numbers of B^*B decay as $s_j(B^*B) \asymp j^{-(1+2a)}e^{-2qj^b}$.

In this case Assumption 1, which expresses the operator T^*T as a function of the prior covariance operator C_0, is satisfied for $\psi^2(t) = \exp(-2qt^{-\frac{b}{1+2a}})$. Next, we find that the function Θ_ψ in (11), which expresses the operator B^*B as a function of C_0, is given as $\Theta_\psi(t) = t^{\frac{1}{2}}\exp(-qt^{-\frac{b}{1+2a}})$, and hence the benchmark smoothness is $\Theta_\psi^2(t) = t\exp(-2qt^{-\frac{b}{1+2a}})$. Finally, we have that as $s \to 0$, the function f in (12) which expresses C_0 as a function of B^*B behaves as $f(s) \sim (\log(s^{-\frac{1}{2q}}))^{-\frac{1+2a}{2b}}$, see Lemma 5 in Sect. 4.3.

Bounding the Bias

In this example we have that $\Theta_\psi^2(t)$ decays exponentially, while $\varphi(t)$ polynomially, hence for any Sobolev-like smoothness of the truth β, it holds $\varphi \prec \Theta_\psi^2$. In other words, even without preconditioning there is no saturation in the bias and we are always in case (1) in Proposition 2. However, our theory still works and we can easily derive the rate for the bias and SPC. The next result follows immediately from the considerations in the previous paragraph and Proposition 2.

Proposition 5 *Suppose that $x^* \in S^\beta$, $\beta > 0$. Then independently of whether preconditioning of the prior is used or not, we have that $b_{x^*}(\alpha) = \mathscr{O}\big((\log(\alpha^{-1}))^{-\frac{\beta}{b}}\big)$, as $\alpha \to 0$.*

A Priori Bounds of the SPC

We now apply Theorem 1 in order to calculate the SPC rate. Again, we first need to calculate the net posterior spread, which in this case is such that $S_{T,C_0}(\alpha) \asymp \frac{1}{\alpha}(\log(\alpha^{-1}))^{-\frac{2a}{b}}$, see [3, Thm 4.2]. We prove the following result, which agrees with [12, Thm 2.1] and [3, Thm 4.3].

Table 1 Outline of SPC rates for Sobolev-type smoothness of the truth, $\varphi(t) = t^{\beta/(1+2a)},\ t > 0$

		$s_j(T^*T) \asymp j^{-2p}$	$s_j(T^*T) \asymp e^{-2qj^b}$
Link	ψ	$t^{p/(1+2a)}$	$\exp\left(-2qt^{-b/(1+2a)}\right)$
Benchmark	Θ_ψ^2	$t^{(1+2a+2p)/(1+2a)}$	$t\exp\left(-2qt^{-b/(1+2a)}\right)$
Saturation	$\varphi = \Theta_\psi^2$	$\beta = 1 + 2a + 2p$	always $\varphi \prec \Theta_\psi^2$
Contraction	SPC	$\delta^{4\beta/(1+2a+2p)}$	$\log^{-2\beta/b}(1/\delta)$

Proposition 6 *Suppose that $x^* \in S^\beta$, $\beta > 0$. Then independently of whether preconditioning of the prior is used or not, for any $\sigma > 0$, any parameter choice rule $\alpha = \alpha(\delta)$ such that $\delta^2(\log(\delta^{-2}))^{\frac{2\beta-2a}{b}} \leq \alpha \leq \delta^{2\sigma}$, gives the (minimax) rate $SPC = \mathscr{O}((\log(\delta^{-2}))^{-\frac{2\beta}{b}}))$, as $\delta \to 0$.*

We outline the previous results in Table 1.

4.2 Analytic-Type Smoothness

4.2.1 Moderately Ill-Posed Operator

We now consider the moderately ill-posed operator setup studied in Sect. 4.1.1 with the difference that here we assume that we have the a priori knowledge that the truth has a certain analytic smoothness. The functions ψ, Θ_ψ and f which have to do with the relationship between the forward operator and the prior covariance are as in Sect. 4.1.1, but the function φ which describes analytic smoothness of the truth as in (18), is now $\varphi(t) = \exp(-\beta t^{-\frac{1}{1+2a}})$. In particular, since φ is exponential while the benchmark smoothness Θ_ψ^2 is of power type, we are always in the high smoothness case $\Theta_\psi^2 \prec \varphi$.

Bounding the Bias

The following is an immediate consequence of Proposition 2 and the considerations in the previous paragraph.

Proposition 7 *Suppose that $x^* \in \mathscr{A}^\beta$, for some $\beta > 0$. Then as $\alpha \to 0$:*

1. *if no preconditioning is used, $b_{x^*}(\alpha) \asymp \alpha$;*
2. *if m_α^δ uses a regularization g_α with qualification* $\exp(-\beta t^{-1})$, *then we have that $b_{x^*}(\alpha) = \mathscr{O}(\exp(-\beta\alpha^{-\frac{1}{1+2a+2p}}))$.*

Remark 11 If no preconditioning is used, the bias convergence rate is always saturated. The qualification as formulated in item (2) is a *sufficient condition*, while the actual form can be calculated easily. The given form highlights that exponential type qualification is required to overcome the limitation of the power type prior covariance in order to treat analytic smoothness. We stress here that such qualification is hard to achieve. For example, iterated Tikhonov can never achieve such exponential qualification, while even Landweber iteration which has qualification t^ν, for any $\nu > 0$, only achieves this qualification for values β which are not too big. On the other hand, $\exp(-\beta t^{-1})$ is a qualification for spectral cut-off for any positive value of β.

A Priori Bounds of the SPC

We again apply Theorem 1 in order to calculate the SPC rate. The net posterior spread is as in Sect. 4.1.1, $S_{T,C_0}(\alpha) \asymp \alpha^{-\frac{1+2p}{1+2a+2p}}$. We prove the following result, using the convention from Definition 2.

Proposition 8 *Suppose that $x^* \in \mathscr{A}^\beta$, $\beta > 0$. Then as $\delta \to 0$:*

1. *if no preconditioning of the prior is used, then for any choice $\alpha = \alpha(\delta, \beta)$ we have that SPC $\gg \delta^2$;*
2. *if m_α^δ uses a regularization g_α with qualification* $\exp(-\beta t^{-1})$*, for* $\alpha = (\log(\delta^{-1/\beta}))^{-(1+2a+2p)}$ *we have that SPC* $= \mathscr{O}(\delta^2(\log(\delta^{-1}))^{1+2p})$.

Remark 12 We stress that according to item (1), without preconditioning we have that δ^2/SPC decays at an algebraic rate, while the optimal achievable (also minimax) rate is of power two up to some logarithmic factor. Since the optimal achievable rate in this case is of power two up to logarithmic factors, it is reasonable to call such problems *mildly ill-posed*, as they are almost well-posed.

4.2.2 Severely Ill-Posed Operator

We now consider the severely ill-posed operator setup studied in Sect. 4.1.2 with the difference that here we assume that we have the a priori knowledge that the truth has a certain analytic smoothness. For simplicity, we concentrate on the case $b = 1$, which corresponds for example to the Cauchy problem for the Helmholtz equation, see [3, Section 5] for details.

The functions ψ, Θ_ψ and f which have to do with the relationship between the forward operator and the prior covariance are as in Sect. 4.1.2 for the value $b = 1$, but the function φ which describes analytic smoothness of the truth as in (18), is now $\varphi(t) = \exp(-\beta t^{-\frac{1}{1+2a}})$. In particular, since both φ and the benchmark smoothness Θ_ψ^2 are exponential, unlike Sect. 4.1.2 we now have a saturation phenomenon.

Bounding the Bias

The following is an immediate consequence of Proposition 2 and the considerations in the previous paragraph.

Proposition 9 *Suppose that $x^* \in \mathscr{A}^\beta$, for some $\beta > 0$. Then as $\alpha \to 0$:*

1. *if $\beta \le 2q$ and independently of whether preconditioning of the prior is used or not, we have that $b_{x^*}(\alpha) = \mathscr{O}(\alpha^{\frac{\beta}{2q}})$;*
2. *if $\beta > 2q$ and no preconditioning is used $b_{x^*}(\alpha) \asymp \alpha$;*
3. *if $\beta > 2q$ and m_α^δ uses a regularization g_α with qualification $t^{\frac{\beta-2q}{2q}}$, then we have that $b_{x^*}(\alpha) = \mathscr{O}(\alpha^{\frac{\beta}{2q}})$.*

The benefits of preconditioning are once more clear and can be seen in item (3). If for example we choose the prior mean m_α^δ using the k-fold Tikhonov regularization filter, then for $\beta \le (k+1)2q$ we have that $b_{x^*}(\alpha) = \mathscr{O}(\alpha^{\frac{\beta}{2q}})$, that is the saturation in the bias is delayed. If we use spectral cut-off, then there is no saturation at all.

A Priori Bounds of the SPC

We again apply Theorem 1 in order to calculate the SPC rate. The net posterior spread is as in Sect. 4.1.2, $S_{T,C_0}(\alpha) \asymp \frac{1}{\alpha}(\log(\alpha^{-1}))^{-2a}$. We prove the following result.

Proposition 10 *Suppose that $x^* \in \mathscr{A}^\beta$, $\beta > 0$. Then as $\delta \to 0$:*

1. *If $\beta \le 2q$ and independently of whether preconditioning of the prior is used or not, for $\alpha = \delta^{\frac{2q}{\beta+q}}$ we have that SPC $= \mathscr{O}(\delta^{\frac{2\beta}{\beta+q}})$;*
2. *if $\beta > 2q$ and no preconditioning of the prior is used, then for any choice $\alpha = \alpha(\delta, \beta)$ we have that SPC $\gg \delta^{\frac{2\beta}{\beta+q}}$;*
3. *if $\beta > 2q$ and m_α^δ uses a regularization g_α with qualification $t^{\frac{\beta-2q}{2q}}$, for $\alpha = \delta^{\frac{2q}{\beta+q}}$ we have that SPC $= \mathscr{O}(\delta^{\frac{2\beta}{\beta+q}})$.*

The benefits of preconditioning can again be seen in item (3). If for example we choose the prior mean m_α^δ using the k-fold Tikhonov regularization filter, then for $\beta \le (k+1)2q$ we achieve the optimal (minimax) rate $\delta^{\frac{2\beta}{\beta+q}}$, that is the saturation in the SPC is delayed. If we use spectral cut-off, then there is no saturation at all. Note again that the optimal scaling of the prior, as a function of the noise level δ, is the same whether we use preconditioning or not.

We outline the results in Table 2.

4.3 Summary and Discussion

We succinctly summarize the above examples, in which we confined to power-type decay of the spectrum of the prior C_0, that is, $s_j(C_0) \asymp j^{-(1+2a)}$, $j = 1, 2, \ldots$, for some $a > 0$.

Table 2 Outline of SPC rates for analytic-type smoothness of the truth, $\varphi(t) = \exp(-\beta t^{-\frac{1}{1+2a}}),\ t > 0$

		$s_j(T^*T) \asymp j^{-2p}$	$s_j(T^*T) \asymp e^{-2qj^b}$
Link	ψ	$t^{p/(1+2a)}$	$\exp\left(-2qt^{-b/(1+2a)}\right)$
Benchmark	Θ^2_ψ	$t^{(1+2a+2p)/(1+2a)}$	$t \exp\left(-2qt^{-b/(1+2a)}\right)$
Saturation	$\varphi = \Theta^2_\psi$	always $\Theta^2_\psi \prec \varphi$	$\beta = 2q$
Contraction	SPC	$\delta^2 \log^{1+2p}(1/\delta)$	$\delta^{2\beta/(\beta+q)}$

First, in Sect. 4.1 we specified the solution element to belong to some Sobolev-type ball as in (17), characterized by $\beta > 0$. The distinction between moderately and severely ill-posed problems then comes from the decay of the singular numbers of the operator T governing Eq. (1).

Then, in Sect. 4.2 we considered analytic type smoothness of the truth as in (18), again characterized by $\beta > 0$. As commented earlier on, to our knowledge we are the first to study these examples. Our findings show that the overall problem degree of ill-posedness can be significantly different than the degree of ill-posedness of the operator.

Finally we stress that the rates exhibited in Tables 1 and 2, correspond to the minimax rates as given in [4, Tbl. 1].

Appendix

Proof (of Lemma 1) We first express the element x_α^δ in terms of z^δ.

$$\begin{aligned} x_\alpha^\delta &= C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^*z^\delta + C_0^{1/2} s_\alpha(B^*B) C_0^{-1/2} m_\alpha^\delta \\ &= C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} B^*z^\delta + C_0^{1/2} s_\alpha(B^*B) g_\alpha(B^*B) B^* z^\delta \\ &= C_0^{1/2}\left[\left(\alpha I + B^*B\right)^{-1} + s_\alpha(B^*B) g_\alpha(B^*B)\right] B^* z^\delta . \end{aligned}$$

We notice that

$$\left(\alpha I + B^*B\right)^{-1} + s_\alpha(B^*B) g_\alpha(B^*B) = \left(\alpha I + B^*B\right)^{-1}\left(I + \alpha g_\alpha(B^*B)\right).$$

The expectation of the posterior mean with respect to the distribution generating z^δ when x^* is given, is thus

$$\mathbb{E}^{x^*} x_\alpha^\delta = C_0^{1/2}\left[\left(\alpha I + B^*B\right)^{-1}\left(I + \alpha g_\alpha(B^*B)\right)\right] B^*BC_0^{-1/2} x^* .$$

For the next calculations we shall use that

$$\begin{aligned} I - \left(\alpha I + B^*B\right)^{-1} \left(I + \alpha g_\alpha(B^*B)\right) B^*B \\ &= \left(\alpha I + B^*B\right)^{-1} \alpha \left(I - g_\alpha(B^*B)B^*B\right) \\ &= s_\alpha(B^*B) r_\alpha(B^*B). \end{aligned}$$

Therefore we rewrite

$$\begin{aligned} x^* - \mathbb{E}^{x^*} x_\alpha^\delta &= C_0^{1/2} \left[I - \left(\alpha I + B^*B\right)^{-1} \left(I + \alpha g_\alpha(B^*B)\right) B^*B\right] C_0^{-1/2} x^* \\ &= C_0^{1/2} s_\alpha(B^*B) r_\alpha(B^*B) C_0^{-1/2} x^*, \end{aligned}$$

which proves the first assertion. The variance is $\mathbb{E}^{x^*} \left\| x_\alpha^\delta - \mathbb{E}^{x^*} x_\alpha^\delta \right\|^2$, and this can be written as in (8), by using similar reasoning as for the bias term.

Proof (of Proposition 1) We notice that $\| I + \alpha g_\alpha(B^*B) \| \leq 1 + \gamma_*$, which gives

$$\begin{aligned} V^\delta(\alpha) &= \delta^2 \mathrm{tr}\left[\left(I + \alpha g_\alpha(B^*B)\right)^2 \left(\alpha I + B^*B\right)^{-2} B^*BC_0\right] \\ &\leq \delta^2 \left(1 + \gamma_*\right)^2 \mathrm{tr}\left[\left(\alpha I + B^*B\right)^{-2} B^*BC_0\right] \end{aligned}$$

Since $\left\| (\alpha + B^*B)^{-1} B^*B \right\| \leq 1$ we see that

$$V^\delta(\alpha) \leq (1 + \gamma_*)^2 \delta^2 \mathrm{tr}\left[\left(\alpha I + B^*B\right)^{-1} C_0\right] = (1 + \gamma_*)^2 \mathrm{tr}\left[C^\delta(\alpha)\right],$$

and the proof is complete.

Proof (of Lemma 2) Since C_0 has finite trace, it is compact, and we use the eigenbasis (arranged by decreasing eigenvalues) $u_j,\ j = 1, 2, \ldots$ Under Assumption 1 this is also the eigenbasis for T^*T. If $t_j,\ j = 1, 2, \ldots$ denote the eigenvalues then we see that

$$T^*T = \sum_{j=1}^{\infty} \tau_j u_j \otimes u_j.$$

Correspondingly, $C_0 = \sum_{j=1}^{\infty} \left(\psi^2\right)^{-1} (\tau_j) u_j \otimes u_j$, which gives the first assertion. Moreover, the latter representation yields that

$$C_0^{1/2} = \sum_{j=1}^{\infty} \left(\left(\psi^2\right)^{-1} (\tau_j)\right)^{1/2} u_j \otimes u_j,$$

such that

$$\begin{aligned}
B^*B &= C_0^{1/2}T^*TC_0^{1/2}\\
&= \sum_{j=1}^{\infty}\left(\left(\psi^2\right)^{-1}(\tau_j)\right)^{1/2}\tau_j\left(\left(\psi^2\right)^{-1}(\tau_j)\right)^{1/2}u_j\otimes u_j\\
&= \sum_{j=1}^{\infty}\left(\left(\psi^2\right)^{-1}(\tau_j)\right)\tau_j u_j\otimes u_j\\
&= \sum_{j=1}^{\infty}\psi^2\left(\left(\left(\psi^2\right)^{-1}(\tau_j)\right)\right)\left(\left(\psi^2\right)^{-1}(\tau_j)\right)u_j\otimes u_j\\
&= \sum_{j=1}^{\infty}\Theta_\psi^2\left(\left(\psi^2\right)^{-1}(\tau_j)\right)u_j\otimes u_j\\
&= \Theta_\psi^2\left(C_0\right),
\end{aligned}$$

and the proof is complete.

Proof (of Proposition 2) For the first item (1), we notice that $\varphi \prec \Theta_\psi^2$ if and only if $\varphi(f^2(t)) \prec t$. The linear function $t \mapsto t$ is a qualification of Tikhonov regularization with constant $\gamma = 1$. Thus, by Lemma 3 we have

$$b_{x^*}(\alpha) \leq \|r_\alpha(B^*B)\| \left\|s_\alpha(B^*B)\varphi(f^2(B^*B))\right\| \leq \gamma_0\varphi(f^2(\alpha)),$$

which completes the proof for this case. For item (2), we have that

$$b_{x^*}(\alpha) = \|s_\alpha(B^*B)x^*\|\,.$$

For any $0 < \alpha \leq 1$, we have $\alpha + t \leq 1 + t$, hence

$$b_{x^*}(\alpha) = \alpha\left\|(\alpha I + B^*B)^{-1}x^*\right\| \geq \alpha\left\|(I + B^*B)^{-1}x^*\right\|\,.$$

We conclude that there exists a constant $c_1 = c_1(x^*, \|B^*B\|)$, such that for small α it holds

$$b_{x^*}(\alpha) \geq c_1\alpha.$$

On the other hand, since $t \prec \varphi(f^2(t))$, there exists a constant $c_2 > 0$ which depends only on the index functions φ, f and on $\|B^*B\|$, such that

$$b_{x^*}(\alpha) = \alpha\left\|(\alpha I + B^*B)^{-1}x^*\right\| \leq \alpha\left\|(B^*B)^{-1}\varphi(f^2(B^*B))w\right\| \leq c_2\alpha.$$

For item (3), we have that

$$\begin{aligned} b_{x^*}(\alpha) &\le \left\| r_\alpha(B^*B)s_\alpha(B^*B)\varphi(f^2(B^*B))\right\| \\ &\le \|s_\alpha(B^*B)B^*B\| \left\| r_\alpha(B^*B)\varphi(f^2(B^*B))\left(B^*B\right)^{-1}\right\| \\ &\le \alpha\gamma\frac{\varphi(f^2(\alpha))}{\alpha} = \gamma\varphi(f^2(\alpha)), \end{aligned}$$

and the proof is complete.

Proof (of Lemma 4) The continuity is clear. For the monotonicity we use the representation (15) to get

$$\begin{aligned} S_{T,C_0}(\alpha) - S_{T,C_0}(\alpha') &= \operatorname{tr}\left[\left(\alpha I + B^*B\right)^{-1} C_0\right] - \operatorname{tr}\left[\left(\alpha' + B^*B\right)^{-1} C_0\right] \\ &= \operatorname{tr}\left[\left(\alpha I + B^*B\right)^{-1}(\alpha' - \alpha)\left(\alpha' + B^*B\right)^{-1} C_0\right] \\ &= (\alpha' - \alpha)\operatorname{tr}\left[\left(\alpha I + B^*B\right)^{-1}\left(\alpha' + B^*B\right)^{-1} C_0\right]. \end{aligned}$$

The trace on the right hand side is positive. Indeed, if (s_j, u_j, u_j) denotes the singular value decomposition of B^*B then this trace can be written as

$$\operatorname{tr}\left[\left(\alpha I + B^*B\right)^{-1}\left(\alpha' + B^*B\right)^{-1} C_0\right] = \sum_{j=1}^{\infty} \frac{1}{\alpha + s_j}\frac{1}{\alpha' + s_j}\langle C_0 u_j, u_j\rangle,$$

where the right hand side is positive since the operator C_0 is positive definite. Thus, if $\alpha < \alpha'$ then $S_{T,C_0}(\alpha) - S_{T,C_0}(\alpha')$ is positive, which proves the first assertion.

The proof of the second assertion is simple, and hence omitted. To prove the last assertion we use the partial ordering of self-adjoint operators in Hilbert space, that is, we write $A \le B$ if $\langle Ax, x\rangle \le \langle Bx, x\rangle,\ x \in X$, for two self-adjoint operators A and B. Plainly, with $a := \|T^*T\|$, we have that $T^*T \le aI$. Multiplying from the left and right by $C_0^{1/2}$ this yields $B^*B \le aC_0$, and thus for any $\alpha > 0$ that $\alpha I + B^*B \le \alpha I + aC_0$. The function $t \mapsto -1/t,\ t > 0$ is operator monotone, which gives $(\alpha I + aC_0)^{-1} \le (\alpha I + B^*B)^{-1}$. Multiplying from the left and right by $C_0^{1/2}$ again, we arrive at

$$C_0^{1/2}(\alpha I + aC_0)^{-1} C_0^{1/2} \le C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} C_0^{1/2}.$$

This in turn extends to the traces and gives that

$$\operatorname{tr}\left[C_0^{1/2}(\alpha I + aC_0)^{-1} C_0^{1/2}\right] \le \operatorname{tr}\left[C_0^{1/2}\left(\alpha I + B^*B\right)^{-1} C_0^{1/2}\right] = S_{T,C_0}(\alpha).$$

Now, let us denote by $t_j,\ j \in \mathbb{N}$, the singular numbers of C_0, then we can bound

$$S_{T,C_0}(\alpha) \geq \operatorname{tr}\left[(\alpha I + aC_0)^{-1} C_0\right] \geq \sum_{t_j \geq \alpha/a} \frac{t_j}{\alpha + at_j} \geq \frac{1}{2a}\#\left\{j,\ t_j \geq \frac{\alpha}{a}\right\}.$$

If $S_{T,C_0}(\alpha)$ were uniformly bounded from above, then there would exist a finite natural number, say N, such that $t_N \geq \frac{\alpha}{a} > t_{N+1}$, for $\alpha > 0$ small enough. But this would imply that $t_{N+1} = 0$, which contradicts the assumption that C_0 is positive definite.

Lemma 5 *For $t > 0$ let $\Theta_\Psi^2(t) = t\exp(-2qt^{-\frac{b}{1+2a}})$, for some $q, b, a > 0$. Then for small s we have $(\Theta_\Psi^2)^{-1}(s) \sim (\log s^{-\frac{1}{2q}})^{-\frac{1+2a}{b}}$.*

Proof Let

$$s = \Theta_\Psi^2(t) > 0 \tag{19}$$

and observe that t is small if and only if s is small. Applying [3, Lem 4.5] for $x = t^{-1}$ we get the result.

Proof (of Proposition 6) In this example the explicit solution of Eq. (16) in Theorem 1 is more difficult. However, as discussed in Sect. 3.4, it suffices to asymptotically balance the squared bias and the posterior spread using an appropriate parameter choice $\alpha = \alpha(\delta)$. Indeed, under the stated choice of α the squared bias is of order

$$(\log(\alpha^{-1}))^{-\frac{2\beta}{b}} \leq \sigma^{-\frac{2\beta}{b}} \log(\delta^{-2})^{-\frac{2\beta}{b}}$$

while the posterior spread term is of order

$$\frac{\delta^2}{\alpha}(\log(\alpha^{-1}))^{-\frac{2a}{b}} \leq \log(\delta^{-2}))^{-\frac{2\beta}{b}}.$$

Proof (of Proposition 8) According to the considerations in Remark 10, it is straightforward to check that without preconditioning the best SPC rate that can be established is $\delta^{\frac{4+8a+8p}{3+4a+6p}}$ which proves item (1). In the preconditioned case, the explicit solution of Eq. (16) in Theorem 1, which in this case has the form

$$\exp(-2\beta\alpha^{-\frac{1}{1+2a+2p}}) = \delta^2\alpha^{-\frac{1+2p}{1+2a+2p}},$$

is again difficult. However, as discussed in Sect. 3.4, it suffices to asymptotically balance the squared bias and the posterior spread using an appropriate parameter choice $\alpha = \alpha(\delta)$. Indeed, using [3, Lem 4.5] we have that the solution to the above equation behaves asymptotically as the stated choice of α, and substitution gives the claimed rate.

Proof (of Proposition 10) We begin with items (1) and (3). The explicit solution of Eq. (16) in Theorem 1, which in this case has the form

$$\alpha^{\frac{\beta}{q}} = \frac{\delta^2}{\alpha}(\log(\alpha^{-1})^{-2a},$$

is difficult. As discussed in Sect. 3.4, it suffices to asymptotically balance the squared bias and the posterior spread using an appropriate parameter choice $\alpha = \alpha(\delta)$. Indeed, under the stated choice of α both quantities are bounded from above by $\delta^{\frac{2\beta}{\beta+q}}$. For item (2), according to the considerations in Remark 10, it is straightforward to check that without preconditioning the best SPC rate that can be established is $\delta^{\frac{4q}{\beta+q}}$.

References

1. S. Agapiou, S. Larsson, A.M. Stuart, Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. Stoch. Process. Appl. **123**(10), 3828–3860 (2013). http://doi.org/10.1016/j.spa.2013.05.001
2. S. Agapiou, J.M. Bardsley, O. Papaspiliopoulos, A.M. Stuart, Analysis of the Gibbs sampler for hierarchical inverse problems. SIAM/ASA J. Uncertain. Quantif. **2**(1), 511–544 (2014)
3. S. Agapiou, A.M. Stuart, Y.X. Zhang, Bayesian posterior contraction rates for linear severely ill-posed inverse problems. J. Inverse Ill-Posed Prob. **22**(3), 297–321 (2014). http://doi.org/10.1515/jip-2012-0071
4. L. Cavalier, Nonparametric statistical inverse problems. Inverse Prob. **24**(3), 034004, 19 (2008). http://doi.org/10.1088/0266-5611/24/3/034004
5. M. Dashti, A.M. Stuart, The Bayesian approach to inverse problems (2013). ArXiv e-prints
6. L.T. Ding, P. Mathé, Minimax rates for statistical inverse problems under general source conditions (2017). ArXiv e-prints. https://arxiv.org/abs/1707.01706. https://doi.org/10.1515/cmam-2017-0055
7. H.W. Engl, M. Hanke, A. Neubauer, Regularization of inverse problems, in *Mathematics and its Applications*, vol. 375 (Kluwer Academic, Dordrecht, 1996). http://doi.org/10.1007/978-94-009-1740-8
8. S. Ghosal, H.K. Ghosh, A.W. Van Der Vaaart, Convergence rates of posterior distributions. Ann. Stat. **28**(2), 500–531 (2000). http://doi.org/10.1214/aos/1016218228
9. B. Hofmann, P. Mathé, Analysis of profile functions for general linear regularization methods. SIAM J. Numer. Anal. **45**(3), 1122–1141(electronic) (2007). http://doi.org/10.1137/060654530
10. B. Knapik, J.B. Salomond, A general approach to posterior contraction in nonparametric inverse problems. Bernoulli (to appear). arXiv preprint arXiv:1407.0335
11. B.T. Knapik, A.W. van der Vaart, J.H. van Zanten, Bayesian inverse problems with Gaussian priors. Ann. Stat. **39**(5), 2626–2657 (2011). http://doi.org/10.1214/11-AOS920
12. B.T. Knapik, A.W. van der Vaart, J.H. van Zanten, Bayesian recovery of the initial condition for the heat equation. Comm. Stat. Theory Methods **42**(7), 1294–1313 (2013). http://doi.org/10.1080/03610926.2012.681417
13. B.T. Knapik, B.T. Szabó, A.W. van der Vaart, J.H. van Zanten, Bayes procedures for adaptive inference in inverse problems for the white noise model. Probab. Theory Relat. Fields **164**, 1–43 (2015)
14. M.S. Lehtinen, L. Päivärinta, E. Somersalo, Linear inverse problems for generalised random variables. Inverse Prob. **5**(4), 599–612 (1989). http://stacks.iop.org/0266-5611/5/599

15. K. Lin, S. Lu, P. Mathé, Oracle-type posterior contraction rates in Bayesian inverse problems. Inverse Prob. Imaging **9**(3), 895–915 (2015). http://doi.org/10.3934/ipi.2015.9.895
16. A. Mandelbaum, Linear estimators and measurable linear transformations on a Hilbert space. Z. Wahrsch. Verw. Gebiete **65**(3), 385–397 (1984). http://doi.org/10.1007/BF00533743
17. P. Mathé, Saturation of regularization methods for linear ill-posed problems in Hilbert spaces. SIAM J. Numer. Anal. **42**(3), 968–973 (electronic) (2004). http://doi.org.pugwash.lib.warwick.ac.uk/10.1137/S0036142903420947
18. K. Ray, Bayesian inverse problems with non-conjugate priors. Electron. J. Stat. **7**, 2516–2549 (2013). http://doi.org/10.1214/13-EJS851
19. B.T. Szabó, A.W. van der Vaart, J.H. van Zanten, Empirical Bayes scaling of Gaussian priors in the white noise model. Electron. J. Stat. **7**, 991–1018 (2013). http://doi.org/10.1214/13-EJS798
20. S.J. Vollmer, Posterior consistency for Bayesian inverse problems through stability and regression results. Inverse Prob. **29**(12), 125011 (2013). https://doi.org/10.1088/0266-5611/29/12/125011

Convex Regularization of Discrete-Valued Inverse Problems

Christian Clason and Thi Bich Tram Do

Abstract This work is concerned with linear inverse problems where a distributed parameter is known a priori to only take on values from a given discrete set. This property can be promoted in Tikhonov regularization with the aid of a suitable convex but nondifferentiable regularization term. This allows applying standard approaches to show well-posedness and convergence rates in Bregman distance. Using the specific properties of the regularization term, it can be shown that convergence (albeit without rates) actually holds pointwise. Furthermore, the resulting Tikhonov functional can be minimized efficiently using a semi-smooth Newton method. Numerical examples illustrate the properties of the regularization term and the numerical solution.

1 Introduction

We consider Tikhonov regularization of inverse problems, where the unknown parameter to be reconstructed is a distributed function that only takes on values from a given discrete set (i.e., the values are known, but not in which points they are attained). Such problems can occur, e.g., in nondestructive testing or medical imaging; a similar task also arises as a sub-step in segmentation or labelling problems in image processing. The question we wish to address here is the following: If such strong a priori knowledge is available, how can it be incorporated in an efficient manner? Specifically, if X and Y are function spaces, $F : X \to Y$ denotes the parameter-to-observation mapping, and $y^\delta \in Y$ is the given noisy data, we would wish to solve the constrained Tikhonov functional

$$\min_{u \in U} \frac{1}{2} \| F(u) - y^\delta \|_Y \tag{1}$$

C. Clason (✉) · T. B. Tram Do
Faculty of Mathematics, University Duisburg-Essen, 45117 Essen, Germany
e-mail: christian.clason@uni-due.de; tram.do@uni-due.de

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_2

for

$$U := \{u \in X : u \in \{u_1, \dots, u_d\} \text{ pointwise}\}, \tag{2}$$

where $u_1, \dots, u_d \in \mathbb{R}$ are the known parameter values. However, this set is nonconvex, and hence the functional in (1) is not weakly lower-semicontinuous and can therefore not be treated by standard techniques. (In particular, it will in general not admit a minimizer.) A common strategy to deal with such problems is by convex relaxation, i.e., replacing U by its convex hull

$$\operatorname{co} U = \{u \in X : u \in [u_1, u_d] \text{ pointwise}\}.$$

This turns (1) into a classical *bang-bang* problem, whose solution is known to generically take on only the values u_1 or u_d; see, e.g., [4, 24]. If $d > 2$, intermediate parameter values are therefore lost in the reconstruction. (Here we would like to remark that a practical regularization should not only converge as the noise level tends to zero but also yield informative reconstructions for fixed—and ideally, a large range of—noise levels.) As a remedy, we propose to add a convex regularization term that promotes reconstructions in U (rather than merely in $\operatorname{co} U$) for the convex relaxation. Specifically, we choose the convex integral functional

$$\mathcal{G} : X \to \mathbb{R}, \qquad \mathcal{G}(u) := \int g(u(x))\, dx,$$

for a convex integrand $g : \mathbb{R} \to \mathbb{R}$ with a polyhedral epigraph whose vertices correspond to the known parameter values $u_1, \dots, u_d$. Just as in L^1 regularization for sparsity (and in linear optimization), it can be expected that minimizers are found at the vertices, thus yielding the desired structure.

This approach was first introduced in [8] in the context of linear optimal control problems for partial differential equations, where the so-called *multi-bang* (as a generalization of bang-bang) penalty $\mathcal{G}$ was obtained as the convex envelope of a (nonconvex) L^0 penalization of the constraint $u \in U$. The application to nonlinear control problems and the limit as the L^0 penalty parameter tends to infinity were considered in [9], and our particular choice of $\mathcal{G}$ is based on this work. The extension of this approach to vector-valued control problems was carried out in [10].

Our goal here is therefore to investigate the use of the multi-bang penalty from [9] as a regularization term in inverse problems, in particular addressing convergence and convergence rates as the noise level and the regularization parameter tend to zero. Due to the convexity of the penalty, these follow from standard results on convex regularization if convergence is considered with respect to the Bregman distance. The main contribution of this work is to show that due to the structure of the pointwise penalty, this convergence can be shown to actually hold pointwise. Since the focus of our work is the novel convex regularization term, we restrict ourselves to linear problems for the sake of presentation. However, all results carry over in a straightforward fashion to nonlinear problems. Finally, we describe following [8, 9] the computation of Tikhonov minimizers using a path-following semismooth Newton method.

Let us briefly mention other related literature. Regularization with convex nonsmooth functionals is now a widely studied problem, and we only refer to the monographs [17, 21, 23] as well as the seminal works [6, 13, 15, 20]. To the best of our knowledge, this is the first work treating regularization of general inverse problems with discrete-valued distributed parameters. As mentioned above, similar problems occur frequently in image segmentation or, more generally, image labelling problems. The former are usually treated by (multi-phase) level set methods [27] or by a combination of total variation minimization and thresholding [7]. More general approaches to image labelling problems are based on graph-cut algorithms [1, 16] or, more recently, vector-valued convex relaxation [14, 19]. Both multi-phase level sets and vector-valued relaxations, however, have the disadvantage that the dimension of the parameter space grows quickly with the number of admissible values, which is not the case in our approach. On the other hand, our approach assumes, similar to [16], a linear ordering of the desired values which is not necessary in the vector-valued case; see also [10].

This work is organized as follows. In Sect. 2, we give the concrete form of the pointwise multi-bang penalty g and summarize its relevant properties. Section 3 is concerned with well-posedness, convergence, and convergence rates of the corresponding Tikhonov regularization. Our main result, the pointwise convergence of the regularized solutions to the true parameter, is the subject of Sect. 4. We also briefly discuss the structure of minimizers for given y^δ and fixed $\alpha > 0$ in Sect. 5. Finally, we address the numerical solution of the Tikhonov minimization problem using a semismooth Newton method in Sect. 6 and apply this approach to an inverse source problem for a Poisson equation in Sect. 7.

2 Multi-Bang Penalty

Let $u_1 < \cdots < u_d \in \mathbb{R}$, $d \geq 2$, be the given admissible parameter values and $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, be a bounded domain. Following [9, § 3], we define the corresponding multi-bang penalty

$$\mathcal{G} : L^2(\Omega) \to \overline{\mathbb{R}}, \qquad \mathcal{G}(u) = \int_\Omega g(u(x))\, dx,$$

for $g : \mathbb{R} \to \overline{\mathbb{R}}$ defined by

$$g(v) = \begin{cases} \frac{1}{2}\left((u_i + u_{i+1})v - u_i u_{i+1}\right) & \text{if } v \in [u_i, u_{i+1}], \quad 1 \leq i < d, \\ \infty & \text{else.} \end{cases}$$

(Note that we have now included the convex constraint $u \in \operatorname{co} U$ in the definition of $\mathcal{G}$.) This choice can be motivated as the convex hull of $\frac{1}{2}\|\cdot\|^2_{L^2(\Omega)} + \delta_U$, where

δ_U denotes the indicator function of the set U defined in (2) in the sense of convex analysis, i.e., $\delta_U(u) = 0$ if $u \in U$ and ∞ else; see [9, § 3]. Setting

$$g_i(v) := \frac{1}{2}\left((u_i + u_{i+1})v - u_i u_{i+1}\right), \qquad 1 \le i < d,$$

it is straightforward to verify that

$$g(v) = \max_{1 \le i < d} g_i(v), \qquad v \in [u_1, u_d],$$

and hence g is the pointwise supremum of affine functions and therefore convex and continuous on the interior of its effective domain $\operatorname{dom} g = [u_1, u_d]$.

We can thus apply the sum rule and maximum rule of convex analysis (see, e.g., [22, Props. 4.5.1 and 4.5.2, respectively]), and obtain for the convex subdifferential at $v \in \operatorname{dom} g$ that

$$\begin{aligned}\partial g(v) &= \partial\left(\max_{1\le i<d} g_i + \delta_{[u_1,u_d]}\right)(v)\\ &= \partial\left(\max_{1\le i<d} g_i\right)(v) + \partial\delta_{[u_1,u_d]}(v)\\ &= \operatorname{co}\left(\bigcup_{i:g(v)=g_i(v)} g_i'(v)\right) + \partial\delta_{[u_1,u_d]}(v).\end{aligned}$$

Using the definition of g_i together with the classical characterization of the subdifferential of an indicator function via its normal cone yields the explicit characterization

$$\partial g(v) = \begin{cases} \left(-\infty, \frac{1}{2}(u_1 + u_2)\right] & \text{if } v = u_1,\\ \left\{\frac{1}{2}(u_i + u_{i+1})\right\} & \text{if } v \in (u_i, u_{i+1}), \quad 1 \le i < d,\\ \left[\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right] & \text{if } v = u_i, \quad 1 < i < d,\\ \left[\frac{1}{2}(u_{d-1} + u_d), \infty\right) & \text{if } v = u_d,\\ \emptyset & \text{else.}\end{cases} \tag{3}$$

In Sects. 5 and 6, we will also make use of the subdifferential of the Fenchel conjugate g^* of g. Here we can use the fact that g is convex and hence $q \in \partial g(v)$ if and only if $v \in \partial g^*(q)$ (see, e.g., [22, Prop. 4.4.4]) to obtain

$$\partial g^*(q) \in \begin{cases} \{u_1\} & \text{if } q \in \left(-\infty, \frac{1}{2}(u_1 + u_2)\right),\\ [u_i, u_{i+1}] & \text{if } q = \frac{1}{2}(u_i + u_{i+1}), \quad 1 \le i < d,\\ \{u_i\} & \text{if } q \in \left(\frac{1}{2}(u_{i-1} + u_i), \frac{1}{2}(u_i + u_{i+1})\right), \quad 1 < i < d,\\ \{u_d\} & \text{if } q \in \left(\frac{1}{2}(u_{d-1} + u_d), \infty\right),\\ \emptyset & \text{else.}\end{cases} \tag{4}$$

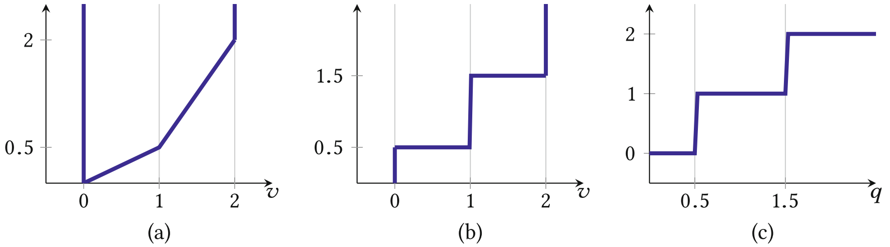


Fig. 1 Structure of pointwise multibang penalty for the choice $(u_1, u_2, u_3) = (0, 1, 2)$. (**a**) g, (**b**) ∂g, (**c**) ∂g^*

(Note that subdifferentials are always closed.) We illustrate these characterizations for a simple example in Fig. 1.

Finally, since g is proper, convex, and lower semi-continuous by construction, the corresponding integral functional $\mathcal{G} : L^2(\Omega) \to \overline{\mathbb{R}}$ is proper, convex and weakly lower semicontinuous as well; see, e.g., [2, Proposition 2.53]. Furthermore, the subdifferential can be computed pointwise as

$$\partial\mathcal{G}(u) = \left\{v \in L^2(\Omega) : v(x) \in \partial g(u(x)) \quad \text{for almost every } x \in \Omega\right\}, \tag{5}$$

see, e.g., [2, Prop. 2.53]. The same is true for the Fenchel conjugate $\mathcal{G}^* : L^2(\Omega) \to \overline{\mathbb{R}}$ and hence for $\partial\mathcal{G}^*$ (which is thus an element of $L^\infty(\Omega)$ instead of $L^2(\Omega)$); see, e.g., [12, Props. IV.1.2, IX.2.1].

3 Multi-Bang Regularization

We consider for a linear operator $K : X \to Y$ between the Hilbert spaces $X = L^2(\Omega)$ and Y and exact data $y^\dagger \in Y$ the inverse problem of finding $u \in X$ such that

$$Ku = y^\dagger. \tag{6}$$

We assume that K is weakly closed, i.e., $u_n \rightharpoonup u$ and $Ku_n \rightharpoonup y$ imply $y = Ku$. For the sake of presentation, we also assume that (6) admits a solution $u^\dagger \in X$. Let now $y^\delta \in Y$ be given noisy data with $\| y^\delta - y^\dagger\|_Y \leq \delta$ for some noise level $\delta > 0$. The *multi-bang regularization* of (6) for $\alpha > 0$ then consists in solving

$$\min_{u\in X} \frac{1}{2}\|Ku - y^\delta\|_Y^2 + \alpha\mathcal{G}(u). \tag{7}$$

Since $\mathcal{G}$ is proper, convex and semi-continuous with bounded effective domain co U, and K is weakly closed, the following results can be proved by standard semi-continuity methods; see also [9, 10].

Proposition 1 (Existence and Uniqueness) *For every $\alpha > 0$, there exists a minimizer u_α^δ to (7). If K is injective, this minimizer is unique.*

Proposition 2 (Stability) *Let $\{y_n\}_{n\in\mathbb{N}} \subset Y$ be a sequence converging strongly to $y^\delta \in Y$ and $\alpha > 0$ be fixed. Then the corresponding sequence of minimizers $\{u_n\}_{n\in\mathbb{N}}$ to (7) contains a subsequence converging weakly to a minimizer u_α^δ.*

We now address convergence for $\delta \to 0$. Recall that an element $u^\dagger \in X$ is called a $\mathcal{G}$-minimizing solution to (6) if it is a solution to (6) and $\mathcal{G}(u^\dagger) \leq \mathcal{G}(u)$ for all solutions u to (6). The following result is standard as well; see, e.g., [17, 21, 23].

Proposition 3 (Convergence) *Let $\{y^{\delta_n}\}_{n\in\mathbb{N}} \subset Y$ be a sequence of noisy data with $\|y^{\delta_n} - y^\dagger\|_Y \leq \delta_n \to 0$, and choose $\alpha_n := \alpha_n(\delta_n)$ satisfying*

$$\lim_{n\to\infty} \frac{\delta_n^2}{\alpha_n} = 0 \qquad \textit{and} \qquad \lim_{n\to\infty} \alpha_n = 0.$$

Then the corresponding sequence of minimizers $\{u_{\alpha_n}^{\delta_n}\}_{n\in\mathbb{N}}$ to (7) contains a subsequence converging weakly to a $\mathcal{G}$-minimizing solution $u^\dagger$.

For convex nonsmooth regularization terms, convergence rates are usually derived in terms of the Bregman distance [5], which is defined for $u_1, u_2 \in X$ and $p_1 \in \partial\mathcal{G}(u_1)$ as

$$d_{\mathcal{G}}^{p_1}(u_2, u_1) = \mathcal{G}(u_2) - \mathcal{G}(u_1) - \langle p_1, u_2 - u_1\rangle_X.$$

From the convexity of $\mathcal{G}$, it follows that $d_{\mathcal{G}}^{p_1}(u_2, u_1) \geq 0$ for all $u_2 \in X$. Furthermore, we have from, e.g., [17, Lem. 3.8] the so-called *three-point identity*

$$d_{\mathcal{G}}^{p_1}(u_3, u_1) = d_{\mathcal{G}}^{p_2}(u_3, u_2) + d_{\mathcal{G}}^{p_1}(u_2, u_1) + (p_2 - p_1)(u_3 - u_2) \tag{8}$$

for any $u_1, u_2, u_3 \in X$ and $p_1 \in \mathcal{G}(u_1)$ and $p_2 \in \partial\mathcal{G}(u_2)$. Finally, we point out that due to the pointwise characterization (5) of the subdifferential of the integral functional $\mathcal{G}$, we have that

$$d_{\mathcal{G}}^{p}(u_2, u_1) = \int_\Omega d_g^{p(x)}(u_2(x), u_1(x))dx \tag{9}$$

for

$$d_g^q(v_2, v_1) = g(v_2) - g(v_1) - q(v_2 - v_1).$$

Standard arguments can then be used to show convergence rates for a priori and a posteriori parameter choice rules under the usual source conditions; see, e.g., [6, 17, 20, 21, 23]. Here we follow the latter and assume that there exists a $w \in Y$ such that

$$p^\dagger := K^*w \in \partial\mathcal{G}(u^\dagger). \tag{10}$$

Under the a priori choice rule

$$\alpha = c\delta \qquad \text{for some } c > 0, \tag{11}$$

we obtain the following convergence rate from, e.g., [17, Cor. 3.4].

Proposition 4 (Convergence Rate, A Priori) *Assume that the source condition* (10) *holds and that* $\alpha = \alpha(\delta)$ *is chosen according to* (11)*. Then there exists a* $C > 0$ *such that*

$$d_{\mathcal{G}}^{p^\dagger}(u_\alpha^\delta, u^\dagger) \le C\delta.$$

We obtain the same rate under the classical Morozov discrepancy principle

$$\delta < \|Ku_\alpha^\delta - y^\delta\|_Y \le \tau\delta, \tag{12}$$

for some $\tau > 1$ from, e.g., [17, Thm. 3.15].

Proposition 5 (Convergence Rate, A Posteriori) *Assume that the source condition* (10) *holds and that* $\alpha = \alpha(\delta)$ *is chosen according to* (12)*. Then there exists a* $C > 0$ *such that*

$$d_{\mathcal{G}}^{p^\dagger}(u_\alpha^\delta, u^\dagger) \le C\delta.$$

4 Pointwise Convergence

The pointwise definition (9) of the Bregman distance together with the explicit pointwise characterization (3) of subgradients allows us to show that the convergence in Proposition 3 is actually pointwise if $u^\dagger(x) \in \{u_1, \dots, u_d\}$ almost everywhere. The following lemma provides the central argument for pointwise convergence.

Lemma 1 *Let* $v^\dagger \in \{u_1, \dots, u_d\}$ *and* $q^\dagger \in \partial g(v^\dagger)$ *satisfying*

$$q^\dagger \in \begin{cases} \left\{\frac{1}{2}(u_i + u_{i+1})\right\} & \text{if } v^\dagger \in (u_i, u_{i+1}), \quad 1 \le i < d, \\ \left(\frac{1}{2}(u_i + u_{i-1}), \frac{1}{2}(u_i + u_{i+1})\right), & \text{if } v^\dagger = u_i, \qquad\quad 1 < i < d \\ \left(-\infty, \frac{1}{2}(u_1 + u_2)\right), & \text{if } v^\dagger = u_1 \\ \left(\frac{1}{2}(u_d + u_{d-1}), \infty\right), & \text{if } v^\dagger = u_d \end{cases} \tag{13}$$

Furthermore, let $\{v_n\}_{n\in\mathbb{N}} \subset [u_1, u_d]$ *be a sequence with*

$$d_g^{q^\dagger}(v_n, v^\dagger) \to 0.$$

Then, $v_n \to v^\dagger$.

Proof We argue by contraposition: Assume that v_n does not converge to $v^\dagger = u_i$ for some $1 \le i \le d$. Then there exists an $\varepsilon > 0$ such that for every $n_0 \in \mathbb{N}$, there is an $n \ge n_0$ with $|v_n - v^\dagger| > \varepsilon$, i.e., either $v_n > u_i + \varepsilon$ or $v_n < u_i - \varepsilon$. We now further discriminate these two cases. (Note that some cases cannot occur if $i = 1$ or $i = d$.)

(i) $v_n > u_{i+1}$: Then, $v_n \in (u_k, u_{k+1}]$ for some $k \ge i+1$. The three point identity (8) yields that

$$d_g^{q^\dagger}(v_n, v^\dagger) = d_g^{q_{i+1}}(v_n, u_{i+1}) + d_g^{q^\dagger}(u_{i+1}, v^\dagger) + (q_{i+1} - q^\dagger)(v_n - u_{i+1})$$

for $q_{i+1} \in \partial g(u_{i+1})$. We now estimate each term separately. The first term is nonnegative by the properties of Bregman distances. For the last term, we can use the assumption (13) and the pointwise characterization (3) to obtain

$$q^\dagger \in \left(\tfrac{1}{2}(u_i + u_{i-1}), \tfrac{1}{2}(u_i + u_{i+1})\right) \quad \text{and} \quad q_{i+1} \in \left[\tfrac{1}{2}(u_{i+1} + u_i), \tfrac{1}{2}(u_{i+1} + u_{i+2})\right],$$

which implies that $q_{i+1} - q^\dagger > 0$. By assumption we have $v_n - u_{i+1} > 0$, which together implies that the last term is strictly positive. For the second term, we can use that $v^\dagger, u_{i+1} \in [u_i, u_{i+1}]$ to simplify the Bregman distance to

$$d_g^{q^\dagger}(u_{i+1}, v^\dagger) = \frac{1}{2}(u_{i+1} - u_i)(u_{i+1} + u_i - 2q^\dagger) > 0,$$

again by assumption (13). Since this term is independent of n, we obtain the estimate

$$d_g^{q^\dagger}(v_n, v^\dagger) > d_g^{q^\dagger}(u_{i+1}, v^\dagger) =: \varepsilon_1 > 0.$$

(ii) $u_i < v_n \le u_{i+1}$: In this case, we can again simplify

$$d_g^{q^\dagger}(v_n, v^\dagger) = \frac{1}{2}(u_{i+1} + u_i - 2q^\dagger)(v_n - v^\dagger) > C_1\varepsilon,$$

since $C_1 := \frac{1}{2}(u_{i+1} + u_i - 2q^\dagger) > 0$ by assumption (13) and $v_n - v^\dagger > \varepsilon$ by hypothesis.

(iii) $v_n < u_i$: We argue similarly to either obtain

$$d_g^{q^\dagger}(v_n, v^\dagger) > d_g^{q^\dagger}(u_{i-1}, v^\dagger) =: \varepsilon_2 > 0$$

or

$$d_g^{q^\dagger}(v_n, v^\dagger) > C_2\varepsilon$$

for $C_2 := -\frac{1}{2}(u_{i-1} + u_i - 2q^\dagger) > 0$.

Thus if we set $\tilde{\varepsilon} := \min\{\varepsilon_1, \varepsilon_2, C_1\varepsilon, C_2\varepsilon\}$, for every $n_0 \in \mathbb{N}$ we can find $n \geq n_0$ such that $d_g^{q^\dagger}(v_n, v^\dagger) > \tilde{\varepsilon} > 0$. Hence, $d_g^{q^\dagger}(v_n, v^\dagger)$ cannot converge to 0. □

Assumption (13) can be interpreted as a strict complementarity condition for $q^\dagger$ and $v^\dagger$. Comparing (13) to (3), we point out that such a choice of $q^\dagger$ is always possible. If $v^\dagger \notin \{u_1, \dots, u_d\}$, on the other hand, convergence in Bregman distance is uninformative.

Lemma 2 *Let $v^\dagger \in (u_i, u_{i+1})$ for some $1 \leq i < d$ and $q^\dagger \in \partial g(v^\dagger)$. Then we have*

$$d_{\mathcal{G}}^{q^\dagger}(v, v^\dagger) = 0 \qquad \textit{for any} \quad v \in [u_i, u_{i+1}].$$

Proof By the definition of the Bregman distance and the characterization (3) of $\partial g(v^\dagger)$ (which is single-valued under the assumption on $v^\dagger$), we directly obtain

$$\begin{aligned} d_g^{q^\dagger}(v, v^\dagger) = \frac{1}{2}\left[(u_i + u_{i+1})v - u_i u_{i+1}\right] - \frac{1}{2}[(u_i + u_{i+1})v^\dagger - u_i u_{i+1}] \\ - \frac{1}{2}(u_i + u_{i+1})(v - v^\dagger) = 0 \end{aligned}$$

for any $v \in [u_i, u_{i+1}]$. □

Lemma 1 allows us to translate the weak convergence from Proposition 3 to pointwise convergence, which is the main result of our work.

Theorem 1 *Assume the conditions of Proposition 3 hold. If $u^\dagger(x) \in \{u_1, \dots, u_d\}$ almost everywhere, the subsequence $u_{\alpha_n}^{\delta_n} \to u^\dagger$ pointwise almost everywhere.*

Proof From Proposition 3, we obtain a subsequence $\{u_n\}_{n\in\mathbb{N}}$ of $\{u_{\alpha_n}^{\delta_n}\}_{n\in\mathbb{N}}$ converging weakly to $u^\dagger$. Since $\mathcal{G}$ is convex and lower semicontinuous, we have that

$$\mathcal{G}(u^\dagger) \leq \liminf_{n\to\infty} \mathcal{G}(u_n) \leq \lim_{n\to\infty} \mathcal{G}(u_n). \tag{14}$$

By the minimizing properties of $\{u_n\}_{n\in\mathbb{N}}$ and the nonnegativity of the discrepancy term, we further obtain that

$$\alpha_n \mathcal{G}(u_n) \leq \frac{1}{2}\|Ku_n - y^{\delta_n}\|_Y^2 + \alpha_n \mathcal{G}(u_n) \leq \frac{\delta_n^2}{2} + \alpha_n \mathcal{G}(u^\dagger).$$

Dividing this inequality by α_n and passing to the limit $n \to \infty$, the assumption on α_n from Proposition 3 yields that

$$\lim_{n\to\infty} \mathcal{G}(u_n) \leq \mathcal{G}(u^\dagger),$$

which combined with (14) gives $\lim_{n\to\infty} \mathcal{G}(u_n) = \mathcal{G}(u^\dagger)$. Hence, $u_n \rightharpoonup u^\dagger$ implies that $d_{\mathcal{G}}^{p^\dagger}(u_n, u^\dagger) \to 0$ for any $p^\dagger \in \partial\mathcal{G}(u^\dagger)$. By the pointwise characterization (9) and the nonnegativity of Bregman distances, this implies that $d_g^{p^\dagger(x)}(u_n(x), u^\dagger(x)) \to 0$ for almost every $x \in \Omega$. Choosing now $p^\dagger \in \partial\mathcal{G}(u^\dagger)$ such that (13) holds for $q^\dagger = p^\dagger(x)$ and $v^\dagger = u^\dagger(x)$ almost everywhere, the claim follows from Lemma 1. □

Since $u_n(x) \in [u_1, u_d]$ by construction, the subsequence $\{u_n\}_{n\in\mathbb{N}}$ is bounded in $L^\infty(\Omega)$ and hence also converges strongly in $L^p(\Omega)$ for any $1 \leq p < \infty$ by Lebesgue's dominated convergence theorem. We remark that since Lemma 1 applied to $u_n(x)$ and $u^\dagger(x)$ does not hold uniformly in Ω, we cannot expect that the convergence rates from Propositions 4 and 5 hold pointwise or strongly as well.

5 Structure of Minimizers

We now briefly discuss the structure of reconstructions obtained by minimizing the Tikhonov functional in (7) for given $y^\delta \in Y$ and fixed $\alpha > 0$, based on the necessary optimality conditions for (7). Since the discrepancy term is convex and differentiable, we can apply the sum rule for convex subdifferentials. Furthermore, the standard calculus for Fenchel conjugates and subdifferentials (see, e.g., [22]) yields for $\mathcal{G}_\alpha := \alpha\mathcal{G}$ that $\mathcal{G}_\alpha^*(p) = \alpha\mathcal{G}^*(\alpha^{-1}p)$ and hence that $p \in \partial\mathcal{G}_\alpha(u)$ if and only if $u \in \partial\mathcal{G}_\alpha^*(p) = \partial\mathcal{G}^*(\frac{1}{\alpha}p)$. We thus obtain as in [8] that $\bar{u} := u_\alpha^\delta \in L^2(\Omega)$ is a solution to (7) if and only if there exists a $\bar{p} \in L^2(\Omega)$ satisfying

$$\begin{cases} \bar{p} = K^*(y^\delta - K\bar{u}) \\ \bar{u} \in \partial\mathcal{G}_\alpha^*(\bar{p}) := \begin{cases} \{u_i\} & \bar{p}(x) \in Q_i, \qquad 1 \leq i \leq d, \\ [u_i, u_{i+1}] & \bar{p}(x) \in Q_{i,i+1} \quad 1 \leq i < d. \end{cases} \end{cases} \tag{15}$$

for

$$\begin{aligned} Q_1 &= \left\{q : q < \tfrac{\alpha}{2}(u_1 + u_2)\right\}, \\ Q_i &= \left\{q : \tfrac{\alpha}{2}(u_{i-1} + u_i) < q < \tfrac{\alpha}{2}(u_i + u_{i+1})\right\}, \quad 1 < i < d, \\ Q_d &= \left\{q : q > \tfrac{\alpha}{2}(u_{d-1} + u_d)\right\}, \\ Q_{i,i+i} &= \left\{q : q = \tfrac{\alpha}{2}(u_i + u_{i+1})\right\}, \qquad\qquad 1 \leq i < d. \end{aligned}$$

Here we have made use of the pointwise characterization in (4) and reformulated the case distinction in terms of $\bar{p}(x)$ instead of $\frac{1}{\alpha}\bar{p}(x)$.

First, we obtain directly from (15) the desired structure of the reconstruction $\bar{u}$: Apart from a singular set

$$\mathcal{S} := \left\{x \in \Omega : \bar{p}(x) = \tfrac{\alpha}{2}(u_i + u_{i+1}) \text{ for some } 1 \leq i < d\right\},$$

we always have $\bar{u}(x) \in \{u_1, \dots, u_d\}$. For operators K where K^*w cannot be constant on a set of positive measure unless $w = 0$ locally (as is the case for many operators involving solutions to partial differential equations; see [8, Prop. 2.3]) and $y^\delta \notin \operatorname{ran} K$, the singular set $\mathcal{S}$ has zero measure and hence the "multi-bang" structure $\bar{u} \in \{u_1, \dots, u_d\}$ almost everywhere can be guaranteed a priori for any $\alpha > 0$.

Furthermore, we point out that the regularization parameter α only enters via the case distinction. In particular, increasing α shifts the conditions on $\bar{u}(x)$ such that the smaller values among the u_i become more preferred. In fact, if $\bar{p}$ is bounded, we can expect that there exists an $\alpha_0 > 0$ such that $\bar{u} \equiv u_1$ for all $\alpha > \alpha_0$. Conversely, for $\alpha \to 0$, the second line of (15) reduces to

$$\bar{u}(x) \in \begin{cases} \{u_1\} & \text{if } \bar{p}(x) < 0, \\ \{u_d\} & \text{if } \bar{p}(x) > 0, \\ [u_1, u_d] & \text{if } \bar{p}(x) = 0, \end{cases}$$

i.e., (15) coincides with the well-known optimality conditions for bang-bang control problems; see, e.g., [25, Lem. 2.26]. Since in the context of inverse problems, we only have $\alpha = \alpha(\delta) \to 0$ if $\delta \to 0$, the limit system (15) will contain consistent data and hence $\bar{p} \equiv 0$. This allows recovery of $u^\dagger(x) \in \{u_2, \dots, u_{d-1}\}$ on a set of positive measure, consistent with Theorem 3. However, if $u^\dagger(x) \in \{u_1, \dots, u_d\}$ does not hold almost everywhere, we can only expect weak and not strong convergence, cf. [10, Prop. 5.10 (ii)].

6 Numerical Solution

In this section we address the numerical solution of the Tikhonov minimization problem (7) for given $y^\delta \in Y$ and $\alpha > 0$, following [9]. For the sake of presentation, we omit the dependence on α and δ from here on. We start from the necessary (and, due to convexity, sufficient) optimality conditions (15). To apply a semismooth Newton method, we replace the subdifferential inclusion $\bar{u} \in \partial \mathcal{G}_\alpha^*(\bar{p})$ by its single-valued Moreau–Yosida regularization, i.e., we consider for $\gamma > 0$ the regularized optimality conditions

$$\begin{cases} p_\gamma = K^*(y^\delta - Ku_\gamma) \\ u_\gamma = (\partial \mathcal{G}_\alpha^*)_\gamma(p_\gamma). \end{cases} \tag{16}$$

The Moreau–Yosida regularization can also be expressed as

$$H_\gamma := (\partial \mathcal{G}_\alpha^*)_\gamma = \partial(\mathcal{G}_{\alpha,\gamma})^*$$

for

$$\mathcal{G}_{\alpha,\gamma}(u) := \alpha\mathcal{G}(u) + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega)},$$

see, e.g., [3, Props. 13.21, 12.29]. This implies that for (u_γ, p_γ) satisfying (16), u_γ is a solution to the strictly convex problem

$$\min_{u\in L^2(\Omega)} \frac{1}{2}\|Ku - y^\delta\|^2_Y + \alpha\mathcal{G}(u) + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega)},$$

so that existence of a solution can be shown by the same arguments as for (7). Note that by regularizing the conjugate subdifferential, we have not smoothed the nondifferentiability but merely made the functional (more) strongly convex. The regularization of $\mathcal{G}^*_\alpha$ instead of $\mathcal{G}^*$ also ensures that the regularization is robust for $\alpha \to 0$. From [9, Prop. 4.1], we obtain the following convergence result.

Proposition 6 *The family $\{u_\gamma\}_{\gamma>0}$ satisfying (16) contains at least one subsequence $\{u_{\gamma_n}\}_{n\in\mathbb{N}}$ converging to a global minimizer of (7) as $n \to \infty$. Furthermore, for any such subsequence, the convergence is strong.*

From [11, Appendix A.2] we further obtain the pointwise characterization

$$[H_\gamma(p)](x) = \begin{cases} u_i & \text{if } p(x) \in Q^\gamma_i, & 1 \le i \le d,\\ \frac{1}{\gamma}\left(p(x) - \frac{\alpha}{2}(u_i + u_{i+1})\right) & \text{if } p(x) \in Q^\gamma_{i,i+1}, & 1 \le i < d,\end{cases}$$

where

$$\begin{aligned}
Q^\gamma_1 &= \left\{q : q < \tfrac{\alpha}{2}\left((1+2\gamma)u_1 + u_2\right)\right\},\\
Q^\gamma_i &= \left\{q : \tfrac{\alpha}{2}\left(u_{i-1} + (1+2\gamma)u_i\right) < q < \tfrac{\alpha}{2}\left((1+2\gamma)u_i + u_{i+1}\right)\right\} \quad \text{for } 1 < i < d,\\
Q^\gamma_d &= \left\{q : \tfrac{\alpha}{2}\left(u_{d-1} + (1+2\gamma)u_d\right) < q\right\},\\
Q^\gamma_{i,i+1} &= \left\{q : \tfrac{\alpha}{2}\left((1+2\gamma)u_i + u_{i+1}\right) \le q \le \tfrac{\alpha}{2}\left(u_i + (1+2\gamma)u_{i+1}\right)\right\} \quad \text{for } 1 \le i < d.
\end{aligned}$$

Since H_γ is a superposition operator defined by a Lipschitz continuous and piecewise differentiable scalar function, H_γ is Newton-differentiable from $L^r(\Omega) \to L^2(\Omega)$ for any $r > 2$; see, e.g., [18, Example 8.12] or [26, Theorem 3.49]. A Newton derivative at p in direction h is given pointwise almost everywhere by

$$[D_N H_\gamma(p)h](x) = \begin{cases} \frac{1}{\gamma}h(x) & \text{if } p(x) \in Q^\gamma_{i,i+1}, \quad 1 \le i < d,\\ 0 & \text{else.}\end{cases}$$

Hence if the range of K^* embeds into $L^r(\Omega)$ for some $r > 2$ (which is the case, e.g., for many convolution operators and solution operators for partial differential equations) and the semismooth Newton step is uniformly invertible, the corresponding Newton iteration converges locally superlinearly. We address this for the concrete example considered in the next section. In practice, the local convergence can be addressed by embedding the Newton method into a continuation strategy, i.e., starting for γ large and then iteratively reducing γ, using the previous solution as a starting point.

7 Numerical Examples

We illustrate the proposed approach for an inverse source problem for the Poisson equation, i.e., we choose $K = A^{-1} : L^2(\Omega) \to L^2(\Omega)$ for $\Omega = [0,1]^2$ and $A = -\Delta$ together with homogeneous boundary conditions. We note that since Ω is a Lipschitz domain, we have that $\operatorname{ran} A^{-*} = \operatorname{ran} A^{-1} = H^2(\Omega) \cap H^1_0(\Omega)$, and hence this operator satisfies the conditions discussed in Sect. 5 that guarantee that $u^\delta_\alpha(x) \in \{u_1, \dots, u_d\}$ almost everywhere if $y^\delta \notin \operatorname{ran} K$; see [8, Prop. 2.3]. For the computational results below, we use a finite element discretization on a uniform triangular grid with 256 × 256 vertices.

The specific form of K can be used to reformulate the optimality condition (and hence the Newton system) into a more convenient form. Introducing $y_\gamma = A^{-1}u_\gamma$ and eliminating u_γ using the second relation of (16), we obtain as in [8] the equivalent system

$$\begin{cases} A^* p_\gamma + y_\gamma - y^\delta = 0, \\ \quad A y_\gamma - H_\gamma(p_\gamma) = 0. \end{cases} \tag{17}$$

Setting $V := H^1_0(\Omega)$, we can consider this as an equation from $V \times V$ to $V^* \times V^*$, which due to the embedding $V \hookrightarrow L^p(\Omega)$ for $p > 2$ provides the necessary norm gap for Newton differentiability of H_γ. By the chain rule for Newton derivatives from, e.g., [18, Lem. 8.4], the corresponding Newton step therefore consists of solving for $(\delta y, \delta p) \in V \times V$ given $(y^k, p^k) \in V \times V$ in

$$\begin{pmatrix} \mathrm{Id} & A^* \\ A & -D_N H_\gamma(p^k) \end{pmatrix} \begin{pmatrix} \delta y \\ \delta p \end{pmatrix} = -\begin{pmatrix} A^* p^k + y - y^\delta \\ A y^k - H_\gamma(p^k) \end{pmatrix} \tag{18}$$

and setting

$$y^{k+1} = y^k + \delta y, \qquad p^{k+1} = p^k + \delta p.$$

Note that the reformulated Newton matrix is symmetric, which in general is not the case for nonsmooth equations. Following [8, Prop. 4.3], the Newton step (18) is uniformly boundedly invertible, from which local superlinear convergence to a solution of (17) follows.

In practice, we include the continuation strategy described above as well as a simple backtracking line search based on the residual norm in (17) to improve robustness. Since the forward operator is linear and H_γ is piecewise linear, the semi-smooth Newton method has the following finite termination property: If $H_\gamma(p^{k+1}) = H_\gamma(p^k)$, then (y^{k+1}, p^{k+1}) satisfy (17); cf. [18, Rem. 7.1.1]. We then recover $u^{k+1} = H_\gamma(p^{k+1})$. In the implementation, we also terminate if more than 100 Newton iterations are performed, in which case the continuation is also terminated and the last successful iterate is returned. Otherwise we terminate if $\gamma < 10^{-12}$. In all results reported below, the continuation is terminated successfully. The implementation of this approach used to obtain the following results can be downloaded from https://github.com/clason/discreteregularization.

The first example illustrates the convergence behavior of the Tikhonov regularization. Here, the true parameter is chosen as

$$\begin{aligned} u^\dagger(x) &= u_1 + u_2\, \chi_{\{x:(x_1-0.45)^2+(x_2-0.55)^2<0.1\}}(x) \\ &\qquad + (u_3-u_2)\, \chi_{\{x:(x_1-0.4)^2+(x_2-0.6)^2<0.02\}}(x) \end{aligned} \tag{19}$$

for $(u_1, u_2, u_3) = (0, 0.1, 0.15)$; see Fig. 2a. (This might correspond to, e.g., material properties of background, healthy tissue, and tumor, respectively.) The noisy data is constructed pointwise via

$$y^\delta = y^\dagger + (\tilde{\delta}\|y^\dagger\|_\infty)\xi,$$

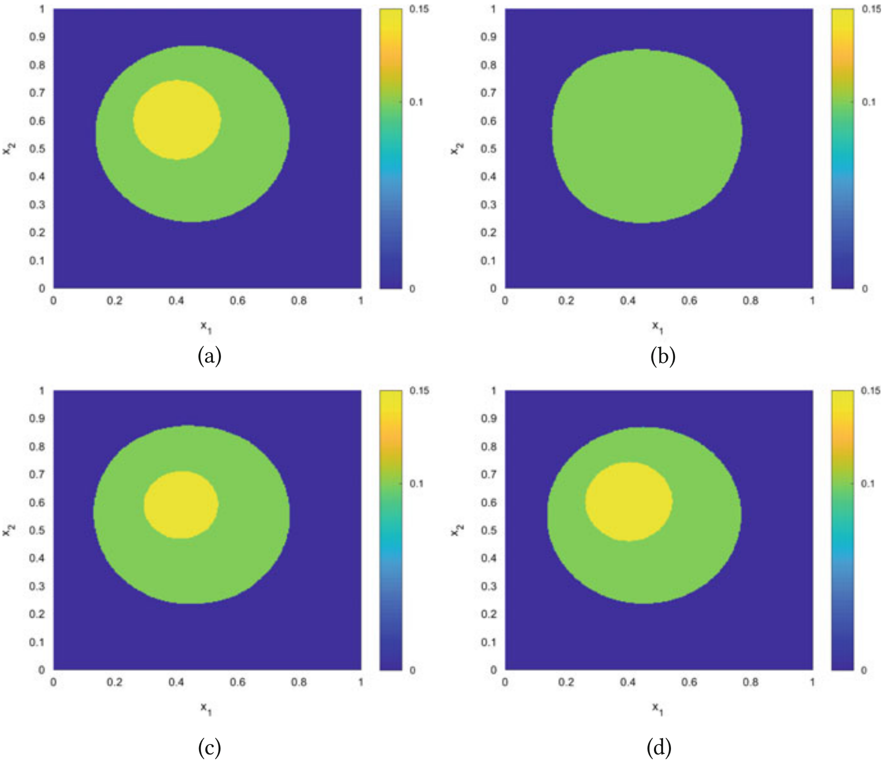


Fig. 2 True parameter $u^\dagger$ for $u_3 = 0.15$ and reconstructions u_α^δ for different values of δ. (**a**) $u^\dagger$. (**b**) u_α^δ for $\delta \approx 1.89 \times 10^{-1}$. (**c**) u_α^δ for $\delta \approx 2.37 \times 10^{-2}$. (**d**) u_α^δ for $\delta \approx 3.69 \times 10^{-4}$

Table 1 Convergence behavior as $\delta \to 0$ for $u_3 = 0.15$: noise level δ, regularization parameter α, L^2-error e_2, L^∞-error e_∞

δ	α	e_2	e_∞	δ	α	e_2	e_∞
1.52e+0	1.00e−2	1.60e+1	1.50e−1	7.44e−4	6.10e−7	6.86e−1	1.00e−1
7.59e−1	1.25e−3	8.64e+0	1.00e−1	3.69e−4	3.05e−7	4.74e−1	1.00e−1
3.78e−1	6.25e−4	6.18e+0	1.00e−1	1.85e−4	1.53e−7	2.91e−1	7.82e−2
1.89e−1	3.13e−4	4.26e+0	1.00e−1	9.28e−5	7.63e−8	2.27e−1	7.67e−2
9.48e−2	7.81e−5	4.32e+0	1.00e−1	4.64e−5	3.81e−8	1.29e−1	5.73e−2
4.73e−2	3.91e−5	3.67e+0	1.00e−1	2.32e−5	1.91e−8	9.19e−2	4.91e−2
2.37e−2	1.95e−5	2.97e+0	1.00e−1	1.16e−5	9.54e−9	9.32e−2	4.03e−2
1.19e−2	9.77e−6	2.33e+0	1.00e−1	5.79e−6	4.77e−9	4.61e−2	2.30e−2
5.90e−3	4.88e−6	1.76e+0	1.00e−1	2.89e−6	2.38e−9	1.13e−1	5.00e−2
2.95e−3	2.44e−6	1.33e+0	1.00e−1	1.44e−6	5.96e−10	1.70e−2	4.39e−3
1.49e−3	1.22e−6	9.47e−1	1.00e−1				

where ξ is a vector of identically and independently normally distributed random variables with mean 0 and variance 1, and $\tilde{\delta} \in \{2^0, \dots, 2^{-20}\}$. For each value of $\tilde{\delta}$, the corresponding regularization parameter α is chosen according to the discrepancy principle (12) with $\tau = 1.1$. Details on the convergence history are reported in Table 1, which shows the effective noise level $\delta := \| y^\delta - y^\dagger \|_2$, the parameter α selected as satisfying the Morozov discrepancy principle, the L^2-error $e_2 := \|u_\alpha^\delta - u^\dagger\|_2$ and the L^∞-error $e_\infty := \|u_\alpha^\delta - u^\dagger\|_\infty$. First, we note that the a posteriori choice approximately follows the a priori choice $\alpha \sim \delta$. Similarly, for larger values of δ, the L^2-error behaves as $e_2 \sim \delta$, which is no longer true for $\delta \to 0$ (and cannot be expected due to the nonsmooth regularization). The L^∞-error e_∞ is initially dominated by the jump in admissible parameter values: As long as there is a single point $x \in \Omega$ with $u_\alpha^\delta(x) = u_i \neq u_j = u^\dagger(x)$, we necessarily have $e_\infty \geq \min_{1 \leq i < d} u_{i+1} - u_i$. (Recall that we do not have a convergence rate and thus an error bound for pointwise convergence.) Later, e_∞ becomes smaller than this threshold value, which indicates that apart from points in the regularized singular set (i.e., where $p_\gamma(x) \in Q_{i,i+1}^\gamma$, which in these cases happens for 20 out of 256×256 vertices), the reconstruction is exact. Here we point out that since γ is independent of α, the Moreau–Yosida regularization for fixed γ becomes more and more active as $\alpha \to 0$. Nevertheless, in all cases $\gamma \ll \alpha$, and hence the multi-bang regularization dominates.

The pointwise convergence can also be seen clearly from Fig. 2, which shows the true parameter $u^\dagger$ together with three representative reconstructions for different noise levels. It can be seen that for large noise, the corresponding large regularization suppresses the smaller inclusion; see Fig. 2b. This is consistent with the discussion at the end of Sect. 5. For smaller noise, the inclusion is recovered well (Fig. 2c), and for $\delta \approx 3.69 \times 10^{-4}$, the reconstruction is visually indistinguishable from the true parameter (Fig. 2d).

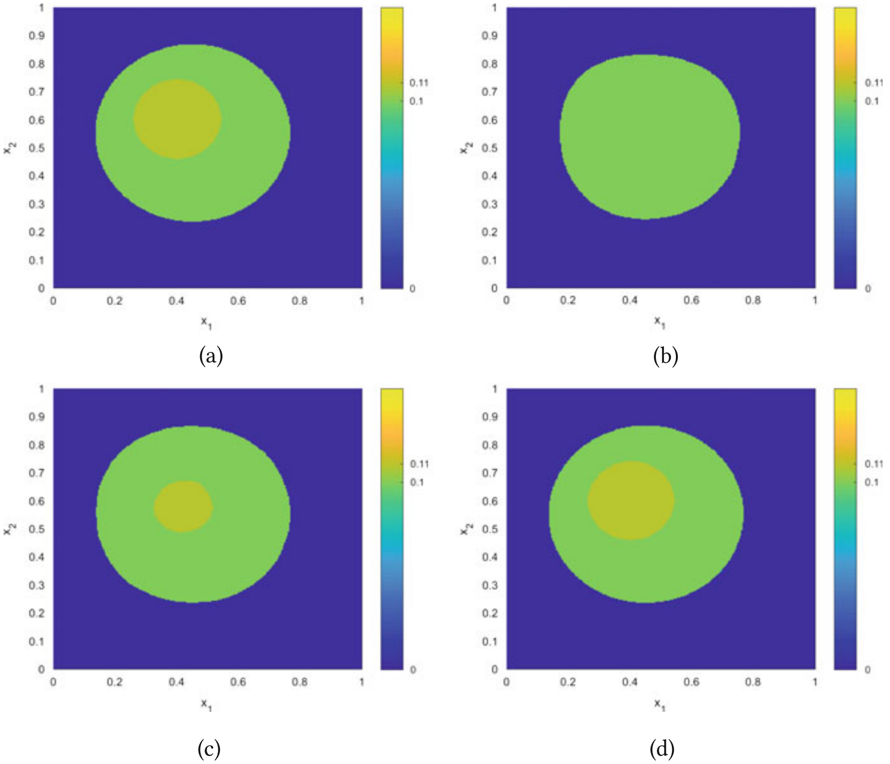


Fig. 3 True parameter $u^\dagger$ for $u_3 = 0.11$ and reconstructions u_α^δ for different values of δ. (**a**) $u^\dagger$. (**b**) u_α^δ for $\delta \approx 1.68 \times 10^{-1}$. (**c**) u_α^δ for $\delta \approx 2.17 \times 10^{-2}$. (**d**) u_α^δ for $\delta \approx 3.29 \times 10^{-4}$

The behavior is essentially the same if we set $(u_1, u_2, u_3) = (0, 0.1, 0.11)$ in (19) (i.e., a contrast of 10% instead of 50% for the inner inclusion), demonstrating the robustness of the multi-bang regularization; see Fig. 3 and Table 2.

To illustrate the behavior if the true parameter does not satisfy the assumption $u^\dagger \in \{u_1, \dots, u_d\}$ almost everywhere, we repeat the above for

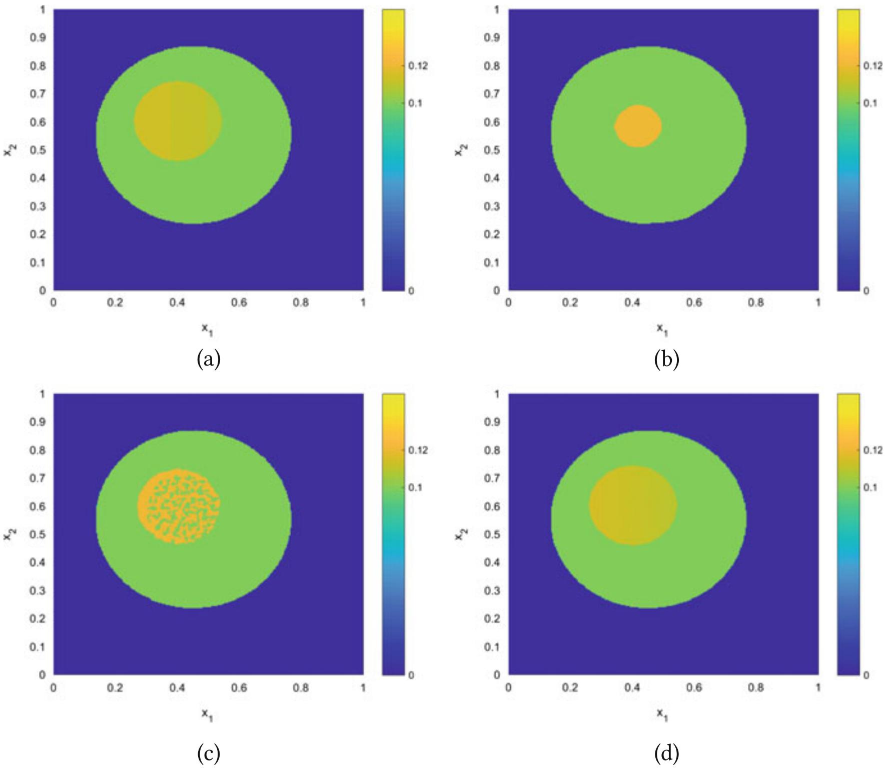
$$
\begin{aligned}
u^\dagger(x) = u_1 &+ u_2\, \chi_{\{x:(x_1-0.45)^2+(x_2-0.55)^2<0.1\}}(x) \\
&+ (u_3 - u_2)(1 - x_1)\, \chi_{\{x:(x_1-0.4)^2+(x_2-0.6)^2<0.02\}}(x)
\end{aligned}
$$

with $(u_1, u_2, u_3) = (0, 0.1, 0.12)$; see Fig. 4a. While for large noise level and regularization parameter value, the multi-bang regularization behaves as before (see Fig. 4b), the reconstruction for smaller noise and regularization (Fig. 4c) shows the typical checkerboard pattern expected from weak but not strong convergence; cf. [8, Rem. 4.2]. Nevertheless, as $\delta \to 0$, we still observe convergence to the true parameter; see Fig. 4d and Table 3.

Finally, we address the qualitative dependence of the reconstruction on the regularization parameter α. Figure 5 shows reconstructions for the true parameter

Table 2 Convergence behavior as $\delta \to 0$ for $u_3 = 0.11$: noise level δ, regularization parameter α, L^2-error e_2, L^∞-error e_∞

δ	α	e_2	e_∞	δ	α	e_2	e_∞
1.34e+0	2.50e−3	1.16e+0	1.10e−1	6.56e−4	6.10e−7	4.55e−1	1.00e−1
6.73e−1	1.25e−3	9.13e+0	1.00e−1	3.29e−4	3.05e−7	2.94e−1	1.00e−1
3.36e−1	6.25e−4	6.89e+0	1.00e−1	1.64e−4	1.53e−7	2.20e−1	6.15e−2
1.68e−1	3.13e−4	4.91e+0	1.00e−1	8.27e−5	7.63e−8	1.87e−1	8.55e−2
8.41e−2	1.56e−4	3.27e+0	1.00e−1	4.11e−5	3.81e−8	6.75e−2	3.35e−2
4.20e−2	3.91e−5	1.90e+0	1.00e−1	2.07e−5	1.91e−8	4.34e−2	1.44e−2
2.17e−2	1.95e−5	1.57e+0	1.00e−1	1.03e−5	9.54e−9	3.72e−2	1.46e−2
1.05e−3	9.77e−6	1.19e+0	1.00e−1	5.12e−6	4.77e−9	3.29e−2	1.31e−2
5.25e−3	4.88e−6	9.81e−1	1.00e−1	2.56e−6	2.38e−9	3.85e−2	1.00e−2
2.64e−3	2.44e−6	8.14e−1	1.00e−1	1.29e−6	2.98e−10	1.65e−1	1.79e−2
1.32e−4	1.22e−6	6.70e−1	1.00e−1				



(a)

(b)

(c)

(d)

Fig. 4 True parameter $u^\dagger$ and reconstructions u_α^δ for different values of δ. (**a**) $u^\dagger$. (**b**) u_α^δ for $\delta \approx 2.11 \times 10^{-2}$. (**c**) u_α^δ for $\delta \approx 3.29 \times 10^{-4}$. (**d**) u_α^δ for $\delta \approx 1.29 \times 10^{-6}$

Table 3 Convergence behavior as $\delta \to 0$ for $u^\dagger$: noise level δ, regularization parameter α, L^2-error e_2, L^∞-error e_∞

δ	α	e_2	e_∞	δ	α	e_2	e_∞
1.36e+0	2.50e−3	1.17e+1	1.15e−1	6.60e−4	6.10e−7	8.46e−1	1.00e−1
6.77e−1	1.25e−3	9.08e+0	1.00e−1	3.29e−4	1.53e−7	7.23e−1	1.00e−1
3.39e−1	6.25e−4	6.84e+0	1.00e−1	1.66e−4	7.63e−8	6.20e−1	5.63e−2
1.69e−1	3.12e−4	4.81e+0	1.00e−1	8.25e−5	3.81e−8	6.04e−1	5.60e−2
8.48e−2	1.56e−4	3.12e+0	1.00e−1	4.12e−5	1.91e−8	5.69e−1	1.83e−2
4.22e−2	3.91e−5	2.03e+0	1.00e−1	2.06e−5	9.54e−9	5.82e−1	5.60e−2
2.11e−2	1.95e−5	1.67e+0	1.00e−1	1.03e−5	4.77e−9	4.95e−1	5.66e−2
1.05e−2	9.77e−6	1.45e+0	1.00e−1	5.15e−6	2.38e−9	3.39e−1	1.47e−2
5.29e−3	4.88e−6	1.29e+0	1.00e−1	2.58e−6	5.96e−10	2.70e−1	2.61e−2
2.66e−3	2.44e−6	1.18e+0	1.00e−1	1.29e−6	3.73e−11	1.65e−1	1.48e−2
1.32e−3	1.22e−6	9.82e−1	1.00e−1				

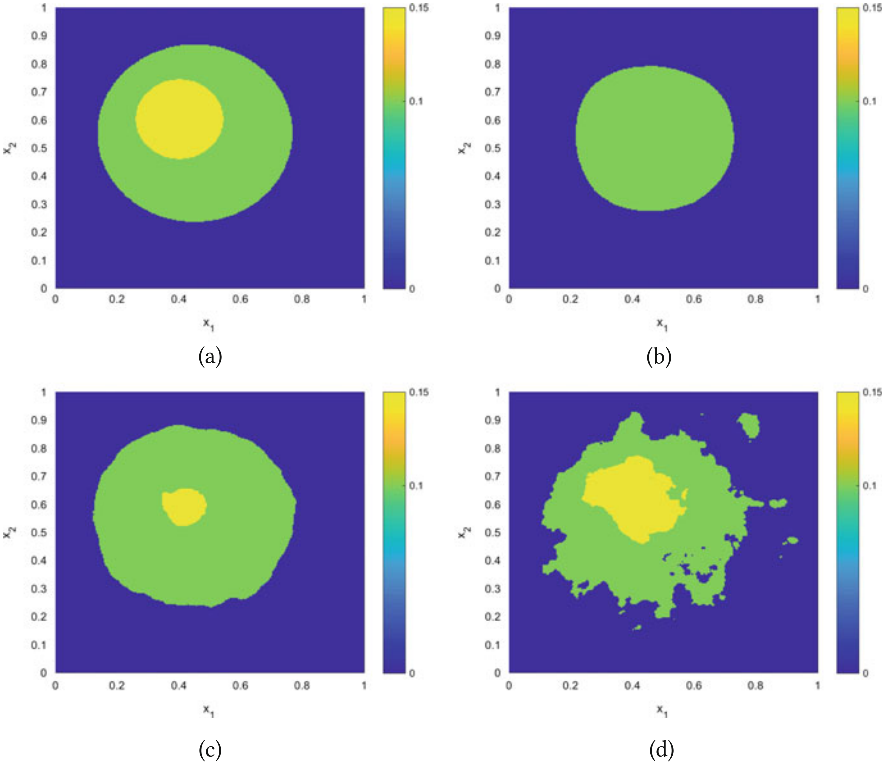


Fig. 5 True parameter $u^\dagger$ and reconstructions u_α^δ for $u_3 = 0.15$, $\delta \approx 7.59 \times 10^{-1}$, and different α. (**a**) $u^\dagger$. (**b**) u_α^δ for $\alpha = 1.25 \times 10^{-3}$. (**c**) u_α^δ for $\alpha = 10^{-4}$. (**d**) u_α^δ for $\alpha = 10^{-5}$

$u^\dagger$ from (19) again with $(u_1, u_2, u_3) = (0, 0.1, 0.15)$ for an effective noise level $\delta \approx 0.759$ and different values of α. First, Fig. 5b presents the reconstruction for the value $\alpha = 1.25 \times 10^{-3}$, where as before the volume corresponding to u_2 is reduced and the inner inclusion corresponding to u_3 is suppressed completely. If the parameter is chosen smaller as $\alpha = 10^{-4}$, however, the reconstruction of the outer volume is essentially correct, while the inner inclusion—although reduced—is also localized well; see Fig. 5c. Visually, this value yields a better reconstruction than the one obtained by the discrepancy principle. The trade-off is a loss of spatial regularity, manifested in more irregular level lines, which becomes even more pronounced for smaller $\alpha = 10^{-5}$; see Fig. 5d. This behavior is surprising insofar that the pointwise definition of the multi-bang penalty itself imposes no spatial regularity on the reconstruction at all; as is evident from (15), any regularity of the solution $\bar{u}$ is solely due to that of the level sets of $\bar{p}$ (which in this case has the regularity of a solution to a Poisson equation).

8 Conclusion

Reconstructions in inverse problems that take on values from a given discrete admissible set can be promoted via a convex penalty that leads to a convergent regularization method. While convergence rates can be shown with respect to the usual Bregman distance, if the true parameter to be reconstructed takes on values only from the admissible set, the convergence (albeit without rates) is actually pointwise. A semismooth Newton method allows the efficient and robust computation of Tikhonov minimizers.

This work can be extended in several directions. First, Fig. 5 demonstrates that regularization parameters chosen according to the discrepancy principle are not optimal with respect to the visual reconstruction quality. This motivates the development of new, heuristic, parameter choice rules that are adapted to the discrete-valued, pointwise, nature of the multi-bang penalty. It would also be interesting to investigate whether an active set condition in the spirit of [28, 29] based on (13) can be used to obtain strong or pointwise convergence rates. A natural further step is the extension to nonlinear parameter identification problems, making use of the results of [9]. Finally, Fig. 5c, d suggest combining the multi-bang penalty with a total variation penalty to also promote regularity of the level lines of the reconstruction. The resulting problem is challenging both analytically and numerically, but would open up the possibility of application to electrical impedance tomography, which can be formulated as parameter identification problem for the diffusion coefficient in an elliptic equation.

Acknowledgements This work was supported by the German Science Fund (DFG) under grant CL 487/1-1. The authors also wish to thank Daniel Wachsmuth for several helpful remarks.

References

1. E. Bae, X.C. Tai, Graph cut optimization for the piecewise constant level set method applied to multiphase image segmentation (Springer, Berlin/Heidelberg, 2009), pp. 1–13, http://doi.org/10.1007/978-3-642-02256-2_1
2. V. Barbu, T. Precupanu, *Convexity and Optimization in Banach Spaces*. Springer Monographs in Mathematics, 4th edn. (Springer, Dordrecht, 2012). http://doi.org/10.1007/978-94-007-2247-7
3. H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC (Springer, New York, 2011). http://doi.org/10.1007/978-1-4419-9467-7
4. M. Bergounioux, F. Tröltzsch, Optimality conditions and generalized bang-bang principle for a state-constrained semilinear parabolic problem. Numer. Funct. Anal. Optim. **17**(5–6), 517–536 (1996). htttp://doi.org/10.1080/01630569608816708
5. L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
6. M. Burger, S. Osher, Convergence rates of convex variational regularization. Inverse Prob. **20**(5), 1411 (2004). http://doi.org/10.1088/0266-5611/20/5/005
7. X. Cai, R. Chan, T. Zeng, A two-stage image segmentation method using a convex variant of the Mumford–Shah model and thresholding. SIAM J. Imag. Sci. **6**(1), 368–390 (2013). http://doi.org/10.1137/120867068
8. C. Clason, K. Kunisch, Multi-bang control of elliptic systems. Ann. Inst. H. Poincaré Anal. Non Linéaire **31**(6), 1109–1130 (2014). http://doi.org/10.1016/j.anihpc.2013.08.005
9. C. Clason, K. Kunisch, A convex analysis approach to multi-material topology optimization. ESAIM: Math. Model. Numer. Anal. **50**(6), 1917–1936 (2016). http://doi.org/10.1051/m2an/2016012
10. C. Clason, C. Tameling, B. Wirth, Vector-valued multibang control of differential equations (2016). arXiv 1611(07853). http://www.arxiv.org/abs/1611.07853
11. C. Clason, K. Ito, K. Kunisch, A convex analysis approach to optimal controls with switching structure for partial differential equations. ESAIM Control, Optimisation and Calculus of Variations **22**(2), 581–609 (2016). http://doi.org/10.1051/cocv/2015017
12. I. Ekeland, R. Témam, *Convex Analysis and Variational Problems*. Classics in Applied Mathematics, vol. 28 (SIAM, Philadelphia, 1999). http://doi.org/10.1137/1.9781611971088
13. J. Flemming, B. Hofmann, Convergence rates in constrained Tikhonov regularization: equivalence of projected source conditions and variational inequalities. Inverse Prob. **27**(8), 085001 (2011). http://doi.org/10.1088/0266-5611/27/8/085001
14. B. Goldluecke, D. Cremers, Convex relaxation for multilabel problems with product label spaces (Springer, Berlin/Heidelberg, 2010), pp. 225–238. http://doi.org/10.1007/978-3-642-15555-0_17
15. B. Hofmann, B. Kaltenbacher, C. Pöschl, O. Scherzer, A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Prob. **23**(3), 987–1010 (2007). http://doi.org/10.1088/0266-5611/23/3/009
16. H. Ishikawa, Exact optimization for Markov random fields with convex priors. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1333–1336 (2003). http://doi.org/10.1109/TPAMI.2003.1233908
17. K. Ito, B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*. Series on Applied Mathematics, vol. 22 (World Scientific, Singapore, 2014). http://doi.org/10.1142/9789814596206_0001
18. K. Ito, K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, Advances in Design and Control, vol. 15 (SIAM, Philadelphia, PA, 2008). http://doi.org/10.1137/1.9780898718614

19. J. Lellmann, C. Schnörr, Continuous multiclass labeling approaches and algorithms. SIAM J. Imag. Sci. **4**(4), 1049–1096 (2011). http://doi.org/10.1137/100805844
20. E. Resmerita, Regularization of ill-posed problems in Banach spaces: convergence rates. Inverse Prob. **21**(4), 1303 (2005). http://doi.org/10.1088/0266-5611/21/4/007
21. O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, F. Lenzen, *Variational Methods in Imaging*. Applied Mathematical Sciences, vol. 167 (Springer, Cham, 2009). http://doi.org/10.1007/978-0-387-69277-7
22. W. Schirotzek, *Nonsmooth Analysis*. Universitext (Springer, Berlin, 2007). http://doi.org/10.1007/978-3-540-71333-3
23. T. Schuster, B. Kaltenbacher, B. Hofmann, K.S. Kazimierski, *Regularization Methods in Banach Spaces*. Radon Series on Computational and Applied Mathematics, vol. 10 (De Gruyter, Berlin, 2012). http://doi.org/10.1515/9783110255720
24. F. Tröltzsch, A minimum principle and a generalized bang-bang principle for a distributed optimal control problem with constraints on control and state. Z. Angew. Math. Mech. **59**(12), 737–739 (1979). http://doi.org/10.1002/zamm.19790591208
25. F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, Translated from the German by Jürgen Sprekels (American Mathematical Society, Providence, 2010). http://doi.org/10.1090/gsm/112
26. M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM Series on Optimization, vol. 11 (SIAM, Philadelphia, PA, 2011). http://doi.org/10.1137/1.9781611970692
27. L.A. Vese, T.F. Chan, A multiphase level set framework for image segmentation using the Mumford and Shah model. Int. J. Comput. Vis. **50**(3), 271–293 (2002). http://doi.org/10.1023/A:1020874308076
28. D. Wachsmuth, G. Wachsmuth, Regularization error estimates and discrepancy principle for optimal control problems with inequality constraints. Control. Cybern. **40**(4), 1125–1158 (2011)
29. G. Wachsmuth, D. Wachsmuth, Convergence and regularization results for optimal control problems with sparsity functional. ESAIM Control Optim. Calc. Var. **17**(3), 858–886 (2011). http://doi.org/10.1051/cocv/2010027

Algebraic Reconstruction of Source and Attenuation in SPECT Using First Scattering Measurements

Evelyn Cueva, Axel Osses, Juan Carlos Quintana, Cristián Tejos, Matías Courdurier, and Pablo Irarrazaval

Abstract Here we present an Algebraic Reconstruction Technique (ART) for solving the identification problem in Single Photon Emission Computed Tomography (SPECT). Traditional reconstruction for SPECT is done by finding the radiation source, nevertheless the attenuation of the surrounding tissue affects the data. In this context, ballistic and first scattering information are used to recover source and attenuation simultaneously. Both measurements are related with the Attenuated Radon Transform and a Klein-Nishina angular type dependency is considered for the scattering. The proposed ART algorithm allow us to obtain good reconstructions of both objects in a few number of iterations.

E. Cueva · A. Osses (✉)
Departamento de Ingeniería Matemática and Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile
e-mail: ecueva@dim.uchile.cl; axosses@dim.uchile.cl

J. C. Quintana
Departamento de Radiología, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: jcquinta@med.puc.cl

C. Tejos
Departamento de Ingeniería Eléctrica, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: ctejos@ing.puc.cl

M. Courdurier
Departamento de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: mcourdurier@mat.uc.cl

P. Irarrazaval
Departamento de Ingeniería Eléctrica and Instituto de Ingeniería Biológica y Médica, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: pim@ing.puc.cl

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_3

1 Introduction

In this work we propose a new reconstruction technique for medical single-photon emission computed tomography (SPECT) imaging. We seek to simultaneously obtain the internal radioactive sources and the attenuation map using not only ballistic measurements but also first-order scattering measurements under very specific scattering regime. The problem is modeled using the radiative transfer equation by means of an explicit non-linear operator that gives the ballistic and scattering measurements as a function of the radioactive source f and attenuation distribution a. In scattering measurements we face one more difficulty, the source has an angular dependency, which in general can not be solved.

The identification problem has motivated several numerical studies. In many of them [6, 9, 10], the focus is to first obtain a good approximation of the attenuation map instead of treating (a,f) as a pair, called attenuation algorithms. Other numerical aspects and reconstructions are presented in [1–3, 5, 7, 8, 11].

This work is based on the results presented in [4]. We made a numerical approach related to the simultaneously source and attenuation reconstruction using an ART algorithm.

In the second section we present the model considered to explain the ballistic and first scattering measurements. In the third section, we describe the Albedo operator related with this inverse problem. This operator explicits the structure of measurements in terms of attenuated Radon transform (AtRT). In the fourth section, we present a discretization of these measurements, in order to represent the ArRT as a linear system, and the ART method for this case is presented. In the last section, numerical results are shown.

2 Model Description

Before describing the model, we introduce some integral operators that appear throughout our study.

2.1 Integral Operators

Let $S^1 = \{\theta \in \mathbb{R}^2 \colon |\theta| = 1\}$ be the set of directions in $\mathbb{R}^2$, and for $\theta = (\theta_1, \theta_2) \in S^1$, let $\theta^\perp = (-\theta_2, \theta_1)$ be its $\pi/2$ counterclockwise rotation.

Definition 1 (Weighted Radon Transform) Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be a function and $w\colon \mathbb{R}^2 \times S^1 \to \mathbb{R}$ be a weight function, the weighted Radon transform of f, with the weight w, is defined as

$$I_w f(s, \theta) = \int_{\mathbb{R}} w(s\theta^\perp + t\theta, \theta) f(s\theta^\perp + t\theta) dt, \quad s \in \mathbb{R}, \quad \theta \in S^1.$$

Definition 2 (Beam Transform) The beam transform of the function $a:\mathbb{R}^2 \to \mathbb{R}$ at the point $x \in \mathbb{R}^2$, in the direction $\theta \in S^1$ is defined as

$$(Ba)(x,\theta) = \int_0^\infty a(x+t\theta)dt, \quad x \in \mathbb{R}^2, \quad \theta \in S^1.$$

The weighted Radon transform with the exponential of the Beam transform as a weight is called the attenuated Radon transform.

Definition 3 (Attenuated Radon Transform) Let $a,f:\mathbb{R}^2 \to \mathbb{R}$, then the AtRT of f, with attenuation a, is defined as

$$R_a f(s,\theta) = \int_{\mathbb{R}} f(s\theta^\perp + t\theta)e^{(Ba)(s\theta^\perp + t\theta,\theta)}dt, \quad s \in \mathbb{R}, \quad \theta \in S^1.$$

When $a \equiv 0$, this is called the Radon transform of f and it is denoted as $Rf(s,\theta)$.

2.2 Ballistic and First Scattering Measurements

In order to describe the inverse problem related to the simultaneous reconstruction of source and attenuation, we make use of the radiative transfer equation (RTE) which is extensively used in medical imaging techniques related with photon transport.

Let $\sigma(x,\theta,\theta')$ be a scattering kernel that describes which photons at the spatial point $x \in \mathbb{R}^2$, coming from direction $\theta \in S^1$ are scattered in the direction $\theta' \in S^1$. So the RTE for an attenuation a, source f and scattering σ is, for all $x \in \mathbb{R}^2$ and $\theta \in S^1$:

$$\begin{aligned}\theta \cdot \nabla_x u(x,\theta) + a(x)u(x,\theta) + \int_{S^1} u(x,\theta)\sigma(x,\theta,\theta')d\theta' \\ = f(x) + \int_{S^1} u(x,\theta')\sigma(x,\theta',\theta)d\theta', \quad \forall x \in \mathbb{R}^2,\ \theta \in S^1. \qquad (1)\end{aligned}$$

The first integral term corresponds to the effect of photons that are scattered away from the path defined by (x,θ). The second integral term is the opposite, and represents the gamma rays traveling in the spatial point $x \in \mathbb{R}^2$ coming from any direction that by a scattering process take the path defined by (x,θ). By introducing the total attenuation:

$$a_T(x,\theta) = a(x) + \int_{S^1} \sigma(x,\theta,\theta')d\theta',$$

Eq. (1) can be rewritten as

$$\theta \cdot \nabla_x u(x,\theta) + a_T(x)u(x,\theta) = f(x) + \int_{S^1} u(x,\theta')\sigma(x,\theta',\theta)d\theta', \ \forall x \in \mathbb{R}^2, \theta \in S^1. \qquad (2)$$

Defining $u_i(x, \theta)$ as the intensity of photons that have been scattered i time, we can decompose the total intensity u as

$$u(x, \theta) = \sum_{i=0}^{\infty} u_i(x, \theta),$$

hence Eq. (2) becomes the system

$$\begin{aligned}
\theta \cdot \nabla_x u_0(x, \theta) + a_T(x, \theta) u_0(x, \theta) &= f(x), \quad \forall x \in \mathbb{R}^2,\ \theta \in S^1 \\
\theta \cdot \nabla_x u_i(x, \theta) + a_T(x, \theta) u_i(x, \theta) &= \int_{S^1} \sigma(x, \theta, \theta') u_{i-1}(x, \theta') d\theta', \quad \forall i \geq 1, \\
\lim_{t \to +\infty} u_i(x - t\theta, \theta) &= 0, \quad \forall i \geq 0,\ \forall x \in \mathbb{R}^2,\ \theta \in S^1.
\end{aligned} \tag{3}$$

We first assume isotropy of the scattering kernel $\sigma(x, \theta, \theta') = \sigma(x, \theta \cdot \theta')$, i.e. the scattering process only depends on the angle at which photons are scattered. Moreover, we assume we can separate variables for the scattering kernel

$$\sigma(x, \theta \cdot \theta') = k(x) \varphi(\theta \cdot \theta').$$

where φ is well known by Klein–Nishina formula. Compton scattering is not equally probable at all energies or scattering angles. The probability of scattering is given by the Klein Nishina equation:

$$\begin{aligned}
\frac{d\sigma}{d\Omega} = Zr_0^2 \left(\frac{1}{1 + \alpha(1 + \cos\theta_C)} \right)^2 \left(\frac{1 + \cos^2\theta_C}{2} \right) \\
\left(1 + \frac{\alpha^2 (1 - \cos\theta_C)^2}{(1 + \cos^2\theta_C)(1 + \alpha\{1 - \cos\theta_C\})} \right)
\end{aligned} \tag{4}$$

where $d\sigma/d\Omega$ is the differential cross-section, Z is the atomic number of the scattering material, r_0 is the classical electron radius, and $\alpha = E_\gamma / m_0 c^2$. E_γ is the photon energy and α is the fine structure constant ($\sim$ 1/137.04). Here $\theta_C = \cos^{-1}(\theta \cdot \theta')$, so φ will be completely determined by θ_C.

Secondly, we assume that the function $k(x)$ is proportional to the attenuation map, i.e. $k(x) = Ca(x)$. Then the system becomes

$$\begin{aligned}
\theta \cdot \nabla_x u_0(x, \theta) + a(x, \theta) u_0(x, \theta) &= f(x), \\
\theta \cdot \nabla_x u_i(x, \theta) + a(x, \theta) u_i(x, \theta) &= Ca(x) \int_{S^1} \varphi(\theta \cdot \theta') u_{i-1}(x, \theta') d\theta', \quad \forall i \geq 1, \\
\lim_{t \to +\infty} u_i(x - t\theta, \theta) &= 0, \quad \forall i \geq 0.
\end{aligned} \tag{5}$$

Proposition 1 *If a and f are uniformly line-integrable then the system* (5) *has a unique solution:*

$$u_0(x,\theta) = \int_{-\infty}^{0} f(x+t\theta)e^{\int_{-\infty}^{0} a_T(x+s\theta)ds}dt, \forall x \in \mathbb{R}^2,\ \theta \in S^1 \tag{6}$$

$$u_i(x,\theta) = C\int_{-\infty}^{0} a(x+t\theta)\int_{S^1} \varphi(\theta\cdot\theta')u_{i-1}(x+t\theta,\theta')d\theta'\, e^{\int_{-\infty}^{0} a(x+s\theta)ds}\, dt,\ \forall i\geq 0. \tag{7}$$

Proof The proof is a generalization of [4, Proposition 1].

As measurements we assume that we are able to record $u_0(x,\theta)$, the ballistic photons, and $u_1(x,\theta)$, the first-order scattering photons, as they exit the patient, i.e. we assume the knowledge of u_0 and u_1 at all points outside the support of a and f.

> In summary, the inverse problem is the reconstruction of the source and attenuation maps f and a from the measurements of the ballistic and first-order scattering photons.

3 Inverse Problem

In this section we present the Albedo operator and the principal results related with its inversion. For this, we assume that $\varphi = 1+\delta\varphi : [-1,1] \to \mathbb{R}$ (quite far from true Klein–Nishina formula at 140 KeV). Proofs are omitted since they are not difficult to obtain as generalizations of propositions and theorems presented in [4].

3.1 Albedo Operator

Defining

$$M_\varphi[a,f](x,\theta) = \int_{S^1} \varphi(\theta\cdot\theta')u_0(x,\theta')d\theta',$$

then (6) and (7) with $i = 1$ becomes:

$$u_0(x,\theta) = \int_{-\infty}^{0} f(x+t\theta)e^{-\int_t^0 a(x+s\theta)ds}dt, \qquad \forall x \in \mathbb{R}^2,\ \theta \in S^1$$

$$u_1(x,\theta) = C\int_{-\infty}^{0} a(x+t\theta)M_\varphi[a,f](x+t\theta,\theta)e^{-\int_t^0 a(x+s\theta)ds}dt, \quad \forall x \in \mathbb{R}^2,\ \theta \in S^1.$$

Therefore, the ballistic and first-order scattering photons exiting the domain correspond to $\mathscr{A}_0$, $\mathscr{A}_1$, respectively, are

$$\mathscr{A}_i(x,\theta) = \lim_{\tau\to+\infty} u_i(x+\tau\theta,\theta), \qquad (x,\theta)\in\mathbb{R}^2\times S^1, \quad i=0,1.$$

The inverse problem can be rephrased as the construction of f and a from knowledge of the Albedo operator

$$\mathscr{A}[a,f] = (\mathscr{A}_0,\mathscr{A}_1) = \left\{(\mathscr{A}_0(x,\theta),\mathscr{A}_1(x,\theta)),\ (x,\theta)\in\mathbb{R}^2\times S^1\right\}, \tag{8}$$

more explicitly

$$\begin{aligned}
\mathscr{A}_0(x,\theta) &= \int_{-\infty}^{\infty} f(x+t\theta)e^{-\int_t^\infty a(x+s\theta)ds}dt, && \forall x\in\mathbb{R}^2, \theta\in S^1\\
\mathscr{A}_1(x,\theta) &= C\int_{-\infty}^{\infty} a(x+t\theta)M_\varphi[a,f](x+t\theta,\theta)e^{-\int_t^\infty a(x+s\theta)ds}dt, && \forall x\in\mathbb{R}^2, \theta\in S^1.\\
M_\varphi[a,f](x,\theta) &= \int_{S^1}\varphi(\theta\cdot\theta')\int_{-\infty}^{\infty} f(x+t\theta')e^{-\int_t^\infty a(x+s\theta')ds}dtd\theta'
\end{aligned}$$

We can rewrite these measurements in terms of a new variable $s\in\mathbb{R}$ and $\theta\in S^1$ as follows

$$\begin{aligned}
\mathscr{A}_0(s,\theta) &:= \mathscr{A}_0(s\theta^\perp,\theta)\\
&= \int_{-\infty}^{\infty} f(s\theta^\perp+t\theta)e^{-\int_t^\infty a(s\theta^\perp+\tau\theta)d\tau}dt,\\
&= R_a[f](s,\theta),
\end{aligned} \tag{9}$$

$$\begin{aligned}
\mathscr{A}_1(s,\theta) &:= \mathscr{A}_1(s\theta^\perp,\theta)\\
&= C\int_{-\infty}^{\infty} a(s\theta^\perp+t\theta)M_\varphi[a,f](s\theta^\perp+t\theta,\theta)e^{-\int_t^\infty a(s\theta^\perp+\tau\theta)d\tau}dt,\\
&= CR_a[aM_\varphi[a,f]](s,\theta).
\end{aligned} \tag{10}$$

Inverse Problem: Given $\mathscr{A}_0$ and $\mathscr{A}_1$ for all $s\in\mathbb{R}, \theta\in S^1$, we want to recover $a(x)$ and $f(x)$, $\forall x\in\mathbb{R}^2$. Now we are interested in inverting the operator $\mathscr{A}$. The principal difficulties involved are:

- We cannot calculate R_a^{-1} since a is unknown.
- This is a non-linear problem in a.

In the next subsection, we present the principal results about the inversion of $\mathscr{A}$.

3.2 Linearized Inverse Problem

To study the invertibility of the Albedo operator $\mathscr{A}$ near a known source and attenuation pair $(\breve{a},\breve{f})$ supported in $K = \overline{B}(0,1)$, it is necessary to calculate the differential $D\mathscr{A}[\breve{a},\breve{f}](\delta a,\delta f)$ of the Albedo operator. Multiply by appropriated cut off functions, the idea is to recover $(\delta a,\delta f)$ from $R_{\breve{a}}^{-1}(\mathscr{D}\mathscr{A}[\breve{a},\breve{f}](\delta a,\delta f)) = (L+Q)[\breve{a},\breve{f}](\delta a,\delta f)$, i.e. compute the inverse $\Big((L+Q)[\breve{a},\breve{f}]\Big)^{-1}$, where

$$L[\breve{a},\breve{f}](\delta a,\delta f) := \begin{pmatrix} \delta f + R_{\breve{a}}^{-1} I_{\breve{w}[\breve{a},\breve{f}]}[\delta a] \\ \delta a \cdot \breve{M}_{\varphi} \end{pmatrix}$$

$$Q[\breve{a},\breve{f}](\delta a,\delta f)$$

$$:= \begin{pmatrix} 0 \\ R_{\breve{a}}^{-1} I_{w[\breve{a},\breve{a}\cdot\breve{M}_{\varphi}]}[\delta a] + (\breve{a}\cdot\partial_a \breve{M}_{\varphi}\delta a) + R_{\breve{a}}^{-1} R_{\breve{a}}(\delta a\cdot\breve{M}_{\delta\varphi} + \breve{a}\cdot M_{\varphi}[\breve{a},\delta f]) \end{pmatrix}.$$

Theorems 2 and 3 guarantee that L is invertible and Q is a relatively small, respectively. This will allows us to prove the invertibility of the linear operator $(L+Q)$ which is presented in Theorem 4.

Now, we assume $\breve{a} \in H^2(K)$, $\breve{f} \in C^{\alpha}(K)$, $\breve{f} \geq 0$, $\breve{f}(0) > 1$, $\|\breve{a}\|_{H^2} \leq 1$ and $\|\breve{f}\|_{C^{\alpha}} \leq 1$. In addition, we suppose $\|\delta\varphi\|_{C^{\alpha}(\mathbb{R}\times S^1)}$ sufficiently small (depending on $\breve{f}$).

Theorem 1 $L[\breve{a},\breve{f}], Q[\breve{a},\breve{f}] : L^2(K)\times L^2(K) \to L^2(K)\times L^2(K)$.

Proof The proof is a generalization of [4, Proposition 4].

Theorem 2 *$L[\breve{a},\breve{f}]$ is invertible with*

$$L^{-1}[\breve{a},\breve{f}]\begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} h/\breve{M}_{\varphi} \\ g - R_{\breve{a}}^{-1} I_{w[\breve{a},\breve{f}]}(h/\breve{M}_{\varphi}) \end{pmatrix}$$

and $||L^{-1}[\breve{a},\breve{f}]|| \leq C$ *uniformly.*

Proof The proof is a generalization of [4, Proposition 5].

Theorem 3 $||Q[\breve{a},\breve{f}]|| \leq C||\breve{a}||$.

Proof The proof is a generalization of [4, Proposition 7].

Theorem 4 *If $\breve{a}$ is small enough and smooth, $\breve{f}$ is positive enough and smooth, if $\delta\varphi$ is small enough and smooth, then $(L+Q)[\breve{a},\breve{f}]$ is invertible and the inverse can be written explicitly as a Neumann series.*

Proof The proof is a generalization of [4, Theorem 4].

4 Reconstruction Algorithm

In this section, we solve the inverse problem (8) where $\mathscr{A}_0$ and $\mathscr{A}_1$ are given by (9) and (10), respectively.

4.1 Linear System and General Algorithm

Knowing $\mathscr{A}_0$ and $\mathscr{A}_1$, our problem is to solve for each $s \in \mathbb{R}$ and for each $\theta \in S^1$:

$$R_a[f](s,\theta) = \mathscr{A}_0(s,\theta) := b_0(s,\theta) \tag{11}$$

$$CR_a[aM_\varphi[a,f]](s,\theta) = \mathscr{A}_1(s,\theta) := b_1(s,\theta) \tag{12}$$

Recalling that AtRT is a line integral we are able to represent it as a matrix which in turn will determine a linear system for f and a as follows:

$$A_0 f = b_0,$$

$$A_1 a = b_1,$$

where A_0 is determine by a discretization of $R_a[f]$ and A_1 by $R_a[aM[a,f]]$. We explain these matrices in the next subsection. First, we present the iterative algorithm for recovering f and a simultaneously in Algorithm 2. This algorithm makes use of Algorithm 1 that helps us to solve any linear system $Ax = b$.

Algorithm 1 ART for solving $Ax = b$

```
Given $A \in \mathscr{M}_{m\times n}(\mathbb{R})$ and $b \in \mathbb{R}^m$. Initialize $x^0$ and number of iterations $Niter$ .
$k \leftarrow 1$
**for** $i = 1 :$ Niter **do**
    **for** $k = 0 : m + 1$ **do**
        Choose $r$ randomly (without repetition) in $\{1, 2, \dots, m\}$.
        Take $a_r$ as the $r$-nt row of $A$.
        Take $b_r$ as the $r$-nt component of $b$.
        Calculate
            $$x^{k+1} = x^k + \frac{b_r - \langle a_r, x^k\rangle}{\|a_r\|^2} a_r,$$
    **end for**
**end for**
```

Algorithm 2 Simultaneous reconstruction of f and a

```
Given $b_0$, $b_1$. Initialized $a^0$ and fix the number of iterations $Niter$.
$k \leftarrow 0$
for $k = 0$ : Niter do
    Calculate $A_0^k$ with $a^k$.
    Solve $A_0^k f^k = b_0$ to get $f^k$ using Algorithm 1.
    Calculate $A_1^k$ with $f^k$ and $a^k$.
    Solve $A_1^k a^{k+1} = b_1$ to get $a^{k+1}$ using Algorithm 1.
end for
```

4.2 Discretization and Matrices Construction

Source f and attenuation a are studied as medical images with $N \times N$ pixels. Then, we consider $I = N^2$ pixels and denote $f(x_i, y_i) =: f_i$ and $a(x_i, y_i) = a_i$ with $i = 1, 2, \ldots, I$.

Let M be the number of angles φ_j in which the detector rotates, so $\theta_j = (\cos\varphi_j, \sin\varphi_j)$ for $j = 1, 2, \ldots, J$. Additionally, we consider $s_k \in (-1, 1)$ for $k = 1, 2, \ldots, K$, as the distances to the origin associated to a line, i.e. we define L_{jk} as the line $L(\theta_j, s_k) = \{x \in \mathbb{R}^2 : x = s_k\theta_j^{\perp} + t\theta,\ t \in \mathbb{R}\}$ and we denote by $w_{ijk} = \text{length}(L_{jk} \cap p_i)$.

Finally, we write $b_{jk}^0 = b_0(\theta_j, s_k)$ the measurement $\mathscr{A}_0(\theta_j, s_k)$ over the line L_{jk}. In Fig. 1 all these variables are explained.

4.2.1 Matrix A_0 Construction

Using the notation describe before, we can write the discretization of (11) by

$$R_a[f](\theta_j, s_k) = b_{jk}^0, \qquad j = 1, \ldots, J, \quad k = 1, \ldots, K,$$

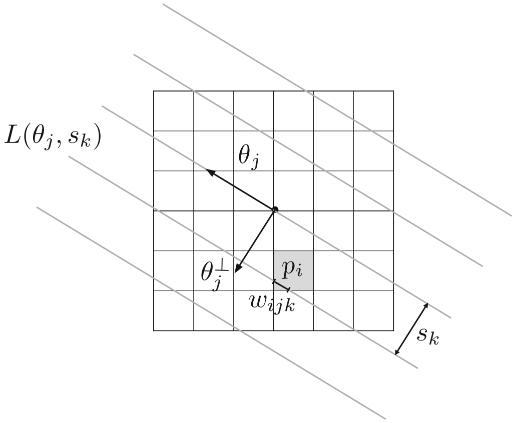


Fig. 1 Line $L_{jk} = L(\theta_j, s_k)$ parametrized by direction θ_j and distance s_k from the origin. For a pixel p_i, w_{ijk} represents the length of $L_{jk} \cap p_i$

where,

$$R_a[f](\theta_j, s_k) \approx \sum_{i=1}^{I} w_{ijk} f_i e^{-D_{ij}a}, \tag{13}$$

with

$$D_{ij}a \approx \int_0^\infty a(p_i + t\theta_j)dt.$$

This last integral is calculated using rotations and sums by columns taking advantage that a is represented by a matrix.

Then, Eq. (13) can be written as the linear system

$$A_0 f = b_0, \qquad A_0 \in \mathscr{M}_{(J\cdot K\times I)},\ f \in \mathbb{R}^I,\ b_0 \in \mathbb{R}^{J\cdot K}$$

with $A^0_{ijk} = w_{ijk}e^{-D_{ij}a}$. Written in this way A is a three dimensional array which is reordered to have $J \cdot K$ rows and I columns. On the other hand, matrices f and a, were reshaped as vectors, from left to right and from top to bottom.

4.2.2 Matrix A_1 Construction

Now we write (12) in a discrete form, as follows:

$$CR_a[aM_\varphi[a,f]](\theta_j, s_k) = b^1_{jk}, \qquad j = 1, \ldots, J, \quad k = 1, \ldots, K.$$

First, let us remember that M_φ is given by

$$M_\varphi[a,f](x,\theta) = \int_{S^1} \varphi(\theta\cdot\theta') \int_0^\infty f(x+t\theta')e^{-\int_x^{x+t\theta'} a(x+\tau\theta')d\tau} dt d\theta',$$

then for each p_i and θ_j, we denote by $M_{ij} := M_\varphi[a,f](p_i,\theta_j)$. This last expression is approximated by

$$M_{ij} = \sum_{j'=1}^{J} u_{ij'}\varphi_{jj'}\Delta\theta'$$

where,

$$u_{ij'} = h\sum_{l\in\mathscr{C}} f(p_i + lh\theta'_j)e^{-(Da(p_i,\theta'_j) - Da(p_i+lh\theta'_j,\theta'_j))}$$

$$\approx \int_0^\infty f(p_i + t\theta_{j'})e^{-\int_{p_i}^{p_i+t\theta_{j'}} a(p_i+\tau\theta_{j'})d\tau} dt$$

with $\mathscr{C} = \{l \in \mathbb{N}\colon lh\theta_{j'} \in \text{sop}f\}$ and h represents the discretization step between two pixels. The point $p_i + lh\theta_{j'}$ is approximated with the nearest pixel p_j. And, $\varphi_{jj'}$ is given by Klein–Nishina's formula.

Combining these approximation for (12) we get

$$CR_a[aM[a,f]](\theta_j, s_k) \approx C\sum_{i=1}^{I} w_{ijk}a_iM_{ij}e^{-D_{ij}a},$$

and a linear system for a is determined by:

$$A_1a = b_1, \qquad A_1 \in \mathscr{M}_{(J\cdot K\times I)},\ a \in \mathbb{R}^I,\ b_1 \in \mathbb{R}^{J\cdot K}.$$

Here, $A^1_{ijk} = Cw_{ijk}M_{ij}e^{-D_{ij}a}$ is reorder to get a bidimensional $J\cdot K\times I$ matrix, and a is reorder as f.

5 Numerical Experiments

In this section we present results obtained with Algorithm 2 implemented in `Matlab`.

We are working in the unit square $[-1,1]^2$ discretized into an equispaced cartesian grid of size $N\times N$ with $N = 128$. The quantities (a,f) are supported inside the unit disc $D = \{x^2 + y^2 < 1\}$.

In all the experiments, we consider three iterations, i.e. $Niter = 3$ in Algorithm 2. These number is enough to reach a good reconstruction of our objects f and a.

We add to our measurements a noise of two natures:

1. First, we added simulated instrumental noise, characterized by an amplitude A so that each pixel value p is replaced by a draw $A\cdot \text{Pois}\left(\frac{p}{A}\right)$,
2. After that we added background noise is added, characterized by a bias value

$$B = \frac{\text{\# added background photons}}{\text{\#photons measured}}.$$

We decided a quantum value q of energy representing one photon, for each additional photon, we add q to a pixel chosen at random with uniform probability among all data pixels.

The experiments with 'low noise' and 'high noise' are carried out with the respective values $(A,B) = (0.2, 0.5)$ and $(A,B) = (0.4, 5)$ (Fig. 2).

The measurements $(\mathscr{A}_0, \mathscr{A}_1)$ are displayed in Fig. 3, and the errors after convergence in all three cases (noiseless, low noise, high noise) are displayed in Fig. 4.

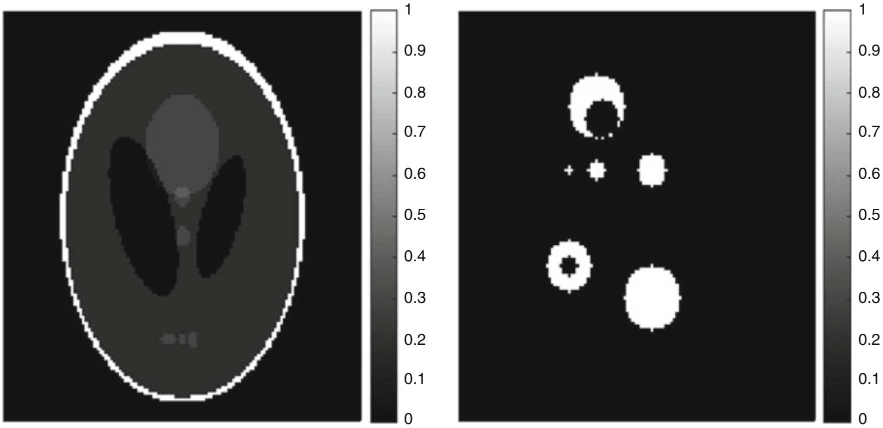


Fig. 2 Examples of discontinuous a (left) and f (right)

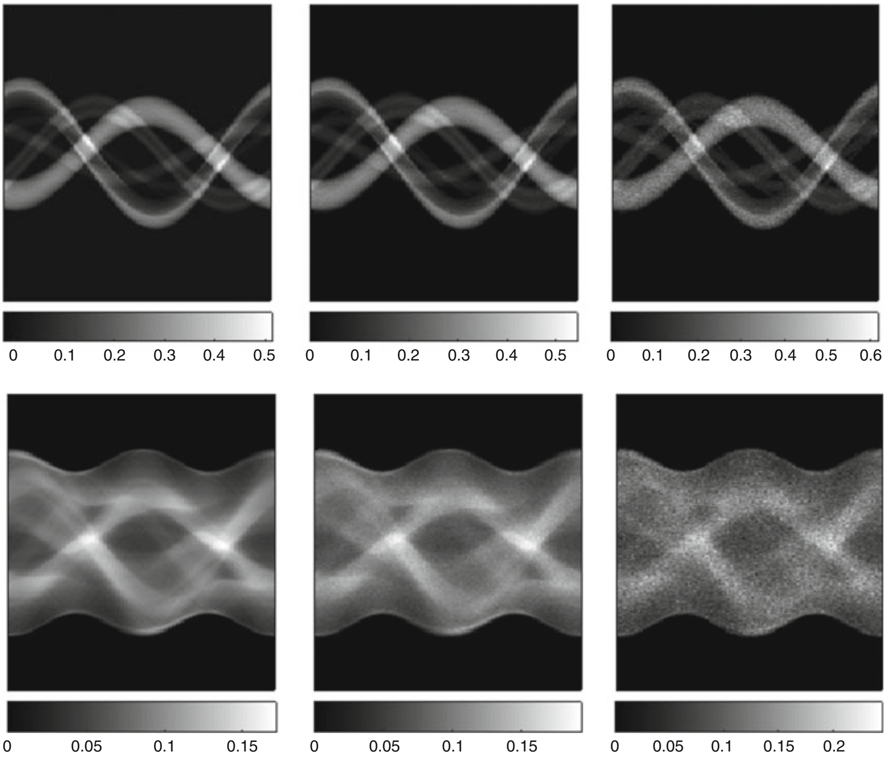


Fig. 3 Forward data $\mathscr{A}_0(a,f)$ (top row) and $\mathscr{A}_1(a,f)$ (bottom row), with (a,f) given in Fig. 2. Left to right: noiseless, low noise, high noise

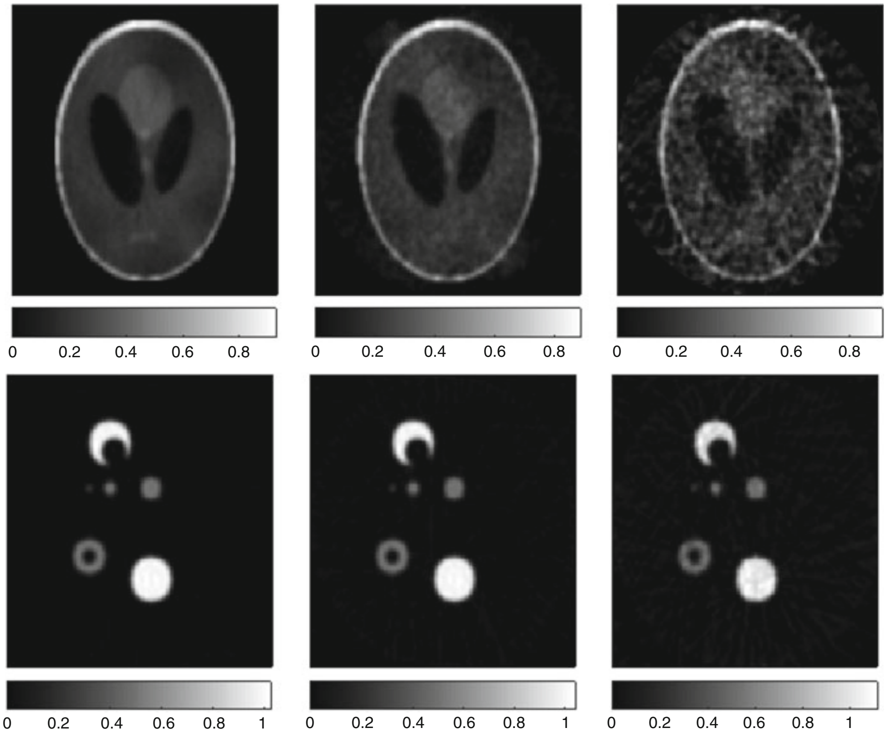


Fig. 4 Reconstructed a (top row) and f (bottom row) after convergence. Left to right: noiseless, low noise, high noise

According to Fig. 4, we can appreciated that reconstructions are satisfactory. Calculation time is reasonable with $N = 128$ ($\sim$ 10 min), the execution time for smaller N, is reduced considerable.

Noise in data makes the algorithm to give negative values to both a and f, although both quantities are physically non-negative. This problem is avoided by including positivity constrains in matrices A_0 and A_1 without affecting its speed. A Gaussian low-pass filter is included after each iteration to avoid the propagation of high frequencies, which appeared in reconstructions of the source and attenuation.

Acknowledgements E.C. was supported by CONICYT-PCHA/Doctorado Nacional/2016-21161721, A.O. was partially supported by CONICYT-Fondecyt grant 1151512. C.T. was partially supported by CONICYT - PIA - Anillo ACT1416. M.C. was partially supported by CONICYT - PIA - Anillo ACT1416 and Fondecyt grant number 1141189. P.I. was partially supported by CONICYT - PIA - Anillo ACT1416.

References

1. A.V. Bronnikov, Numerical solution of the identification problem for the attenuated Radon transform. Inverse Prob. **15**(5), 1315 (1999)
2. A.V. Bronnikov, Reconstruction of attenuation map using discrete consistency conditions. IEEE Trans. Med. Imaging **19**(5), 451–462 (2000)
3. Y. Censor, D.E. Gustafson, A. Lent, H. Tuy, A new approach to the emission computerized tomography problem: simultaneous calculation of attenuation and activity coefficients. IEEE Trans. Nucl. Sci. **26**(2), 2775–2779 (1979)
4. M. Courdurier, F. Monard, A. Osses, F. Romero, Simultaneous source and attenuation reconstruction in SPECT using ballistic and single scattering data. Inverse Prob. **31**(9), 095002 (2015)
5. V. Dicken, A new approach towards simultaneous activity and attenuation reconstruction in emission tomography. Inverse Prob. **15**(4), 931 (1999)
6. D. Gourion, D. Noll, The inverse problem of emission tomography. Inverse Prob. **18**(5), 1435 (2002)
7. S. Luo, J. Qian, P. Stefanov, Adjoint state method for the identification problem in SPECT: recovery of both the source and the attenuation in the attenuated X-ray transform. SIAM J. Imag. Sci. **7**(2), 696–715 (2014)
8. S.H. Manglos, Determination of the attenuation map from SPECT projection data alone. J. Nucl. Med. **35**, 193 (1993)
9. R. Ramlau, R. Clackdoyle, Accurate attenuation correction in SPECT imaging using optimization of bilinear functions and assuming an unknown spatially-varying attenuation distribution. In *Nuclear Science Symposium, 1998. Conference Record. 1998*, vol. 3 (IEEE, New York, 1998), pp. 1684–1688
10. A. Welch, R. Clack, F. Natterer, G.T. Gullberg, Toward accurate attenuation correction in SPECT without transmission measurements. IEEE Trans. Med. Imaging **16**(5), 532–541 (1997)
11. H. Zaidi, B. Hasegawa, Determination of the attenuation map in emission tomography. J. Nucl. Med. **44**(2), 291–315 (2003)

On ℓ^1-Regularization Under Continuity of the Forward Operator in Weaker Topologies

Daniel Gerth and Bernd Hofmann

Abstract Our focus is on the stable approximate solution of linear operator equations based on noisy data by using ℓ^1-regularization as a sparsity-enforcing version of Tikhonov regularization. We summarize recent results on situations where the sparsity of the solution slightly fails. In particular, we show how the recently established theory for weak*-to-weak continuous linear forward operators can be extended to the case of weak*-to-weak* continuity. This might be of interest when the image space is non-reflexive. We discuss existence, stability and convergence of regularized solutions. For injective operators, we will formulate convergence rates by exploiting variational source conditions. The typical rate function obtained under an ill-posed operator is strictly concave and the degree of failure of the solution sparsity has an impact on its behavior. Linear convergence rates just occur in the two borderline cases of proper sparsity, where the solutions belong to ℓ^0, and of well-posedness. For an exemplary operator, we demonstrate that the technical properties used in our theory can be verified in practice. In the last section, we briefly mention the difficult case of oversmoothing regularization where $x^\dagger$ does not belong to ℓ^1.

1 Introduction

We are going to deal with the stable solution of linear operator equations

$$Ax = y \tag{1}$$

with a *bounded linear* operator $A : \ell^1 \to Y$, mapping from the *non-reflexive* infinite dimensional space ℓ^1 of absolutely summable infinite real or complex sequences to an *infinite dimensional Banach space* Y. Instead of the exact right-hand side y from

D. Gerth · B. Hofmann (✉)
Faculty of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany
e-mail: daniel.gerth@mathematik.tu-chemnitz.de

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_4

the range $\mathscr{R}(A)$ of A we assume to have only noisy data $y^\delta \in Y$ available which satisfy the deterministic noise model

$$\|y - y^\delta\|_Y \le \delta \tag{2}$$

with prescribed noise level $\delta > 0$. Our focus for solving Eq. (1) is on the method of ℓ^1-regularization, where for regularization parameters $\alpha > 0$ the minimizers x_α^δ of the extremal problem

$$\frac{1}{p}\|Ax - y^\delta\|_Y^p + \alpha\,\|x\|_{\ell^1} \to \min, \qquad \text{subject to} \quad x \in \ell^1, \tag{3}$$

are used as approximate solutions. This method is a sparsity-enforcing version of Tikhonov regularization, possessing applications in different branches of imaging, natural sciences, engineering and mathematical finance. It was comprehensively analyzed with all its facets and varieties in the last 15 years (cf., e.g., the corresponding chapters in the books [31–33] and the papers [1, 4, 8, 12, 20–22, 25, 28, 29]). We restrict our considerations to *injective* operators A such that, for right-hand sides, the element $x^\dagger = (x_1^\dagger, x_2^\dagger, \ldots) \in \ell^1$ denotes the uniquely determined solution to (1). For assertions concerning the case of non-injective operators A in the context of ℓ^1-regularization, we refer to [9]. In the non-injective case, even the ℓ^1-norm minimizing solutions need not be uniquely determined. As a consequence, very technical conditions must be introduced in order to formulate convergence assertions and rates. In our framework, the Propositions 7 and 11 below would have to be adapted, which however is out of the scope of this paper.

With the paper [5] as starting point and preferably based on variational source conditions first introduced in [24], convergence rates for ℓ^1-regularization of operator equations (1) and modifications like elastic-net

$$\frac{1}{p}\|Ax - y^\delta\|_Y^p + \alpha\left(\frac{1}{2}\|x\|_{\ell^2}^2 + \eta\,\|x\|_{\ell^1}\right) \to \min, \qquad \text{subject to} \quad x \in \ell^1, \tag{4}$$

have been verified under the condition that the sparsity assumption slightly fails (cf. [6, 13, 14]). This means that the solution $x^\dagger \in \ell^1$ is not sparse, abbreviated as $x^\dagger \notin \ell^0$. Most recently in [11], the first author and Jens Flemming have shown that complicated conditions on A, usually supposed for proving convergence rates in ℓ^1-regularization (cf. [5, Assumption 2.2 (c)] and condition (9) below), can be simplified to the requirement of weak*-to-weak continuity of the injective operator A. This seems to be convincing if Y is a reflexive Banach space. The present paper, however, makes assertions also in the case that A is only weak*-to-weak* continuous, which is of interest for non-reflexive Banach spaces Y. Moreover, we complement results from [11], for example with respect to the well-posed situation.

The paper is organized as follows. In Sect. 2 we recall basic properties of ℓ^1-regularization. We proceed in Sect. 3 by discussing the ill-posedness of Eq. (1). We mention that in particular variational source conditions allow us to deal with the ill-posedness and yield convergence rates. For our convergence analysis a particular property of the operator is necessary. In Sect. 4 we show that weak*-to-weak continuity and injectivity imply this property. Interestingly, the same property holds under weak*-to-weak* continuity and injectivity as shown in Sect. 5. There we also derive the convergence rates which hold for both continuity assumptions. Finally, we demonstrate that even the case of a well-posed operator is reflected in our property in Sect. 6. There we also hint at the case of oversmoothing regularization, which occurs when one employs ℓ^1-regularization although the true solution $x^\dagger$ does not belong to ℓ^1.

2 Preliminaries and Basic Propositions

In this paper, we consider the variant (3) of ℓ^1-regularization with some exponent $p > 1$ and with a regularization parameter $\alpha > 0$. Let $y \in \mathscr{R}(A)$. Then, due to the injectivity of A, there exists a uniquely determined solution $x^\dagger \in \ell^1$ to (1). With the following Proposition 1 we recall the assertions of Proposition 2.8 in [5] with respect to existence, stability, convergence and sparsity of the ℓ^1-regularized solutions x_α^δ. The proof ibidem emphasizes the fact that most of these properties follow directly from the general theory of Tikhonov regularization in Banach spaces (cf., e.g., [24, Section 3] and [33, Section 4.1]). Since for $p > 1$ the Tikhonov functional to be minimized in (3) is strictly convex, the regularized solutions x_α^δ, whenever they exist, are uniquely determined for all $\alpha > 0$.

Proposition 1 *Let $A : \ell^1 \to Y$ be weak*-to-weak continuous, i.e., $x_n \rightharpoonup^* x_0$ in ℓ^1 implies that $Ax_n \rightharpoonup Ax_0$ in Y. Then for all $\alpha > 0$ and all $y^\delta \in Y$ there exist uniquely determined minimizers $x_\alpha^\delta \in \ell^1$ of the Tikhonov functional from (3). These regularized solutions are sparse, i.e., $x_\alpha^\delta \in \ell^0$, and they are stable with respect to the data, i.e., small perturbations in y^δ in the norm topology of Y lead only to small changes in x_α^δ with respect to the weak*-topology in ℓ^1. If $\delta_n \to 0$ and if the regularization parameters $\alpha_n = \alpha(\delta_n, y^{\delta_n})$ are chosen such that $\alpha_n \to 0$ and $\frac{\delta_n^p}{\alpha_n} \to 0$ as $n \to \infty$, then $x_{\alpha_n}^{\delta_n}$ converges in the weak*-topology of ℓ^1 to the uniquely determined solution $x^\dagger$ of the operator equation (1). Moreover we have $\lim_{n\to\infty} \|x_{\alpha_n}^{\delta_n}\|_{\ell^1} = \|x^\dagger\|_{\ell^1}$, which, as a consequence of the weak* Kadec-Klee property in ℓ^1 (see, e.g., [3, Lemma 2.2]), implies norm convergence*

$$\lim_{n\to\infty} \|x_{\alpha_n}^{\delta_n} - x^\dagger\|_{\ell^1} = 0.$$

The weak*-to-weak continuity of A in combination with the *stabilizing property* of the penalty functional $\|x\|_{\ell^1}$ in ℓ^1 together with an appropriate choice of the regularization parameter $\alpha > 0$ represent basic assumptions of Proposition 1. In contrast to regularization in reflexive Banach space, where the level sets of the norm functional are weakly compact, we have in ℓ^1 weak* compactness of the corresponding level sets according to the sequential Banach-Alaoglu theorem (cf., e.g., [30, Theorems 3.15 and 3.17]), which we present in form of the following lemma.

Lemma 1 *The closed unit ball of a Banach space X is compact in the weak*-topology if there is a separable Banach space Z (predual space) with dual $Z^* = X$. Then any bounded sequence $\{x_n\}_{n\in\mathbb{N}}$ in X has a weak*-convergent subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ such that $x_{n_k} \rightharpoonup^* x_0 \in X$ as $k \to \infty$.*

The occurring kind of compactness of the level sets with $X = \ell^1$ and predual space $Z = c_0$ ensures the existence of minimizers x_α^δ of the functional (3).

Throughout this paper, we use the terms 'continuous', 'compact' or 'lower semicontinuous' for an operator, a set or a functional always in the sense of 'sequentially continuous', 'sequentially compact' or 'sequentially lower semicontinuous'. As the Lemmas 6.3 and 6.5 from [10] show, there is no reason for a distinction in case of using weak topologies. From Lemma 2.7 and Proposition 2.4 in [5] one can take assertions concerning sufficient conditions for the weak*-to-weak continuity of A, which we summarize in the Proposition 2 below. As also indicated in Proposition 1, for the choice of α, the so-called regularization property

$$\alpha(\delta, y^\delta) \to 0 \qquad \text{and} \qquad \frac{\delta^p}{\alpha(\delta, y^\delta)} \to 0 \qquad \text{as} \quad \delta \to 0, \tag{5}$$

where α tends to zero, but sufficiently slow, plays an important role. In our studies, we consider on the one hand *a priori parameter choices* $\alpha_{APRI} = \alpha(\delta)$ defined as

$$\alpha(\delta) := \frac{\delta^p}{\varphi(\delta)}, \qquad 0 < \delta \le \overline{\delta}, \tag{6}$$

with *concave* index functions φ. In this context, we call $\varphi : [0,\infty) \to [0,\infty)$ an *index function* if φ with $\varphi(0) = 0$ is continuous and strictly increasing. Obviously, an a priori parameter choice α_{APRI} from (6) with an arbitrary concave index function φ satisfies (5) as $\lim_{\delta\to+0}\varphi(\delta) = 0$ is valid for each index function and we have $\frac{\delta^p}{\varphi(\delta)} = \frac{\delta}{\varphi(\delta)}\,\delta^{p-1} \to 0$ as $\delta \to 0$, because δ^{p-1} is an index function for all exponents $p > 1$ in (3) and the factor $\frac{\delta}{\varphi(\delta)}$ is bounded whenever φ is concave.

On the other hand, we consider the *sequential discrepancy principle*, comprehensively analyzed in [2] (see also [23]), as a specific *a posteriori parameter choice* $\alpha_{SDP} = \alpha(\delta, y^\delta)$ for the regularization parameter. For prescribed $\tau > 1,\ 0 < q < 1$, and a sufficiently large value $\alpha_0 > 0$, we let

$$\Delta_q := \{\alpha_j > 0 : \alpha_j = q^j\alpha_0, \quad j = 1, 2, \ldots\}.$$

Given $\delta > 0$ and $y^\delta \in Y$, we choose $\alpha = \alpha_{SDP} \in \Delta_q$ according to the sequential discrepancy principle such that

$$\|Ax_\alpha^\delta - y^\delta\| \le \tau\delta < \|Ax_{\alpha/q}^\delta - y^\delta\|. \tag{7}$$

By Theorem 1 in [2] it has been shown that there is $\overline{\delta} > 0$ such that α_{SDP} is well-defined for $0 < \delta \leq \overline{\delta}$ and satisfies (5) whenever data compatibility in the sense of [2, Assumption 3] takes place.

Consequently, both regularization parameter choices $\alpha = \alpha_{APRI}$ and $\alpha = \alpha_{SDP}$ are applicable for the ℓ^1-regularization in order to get existence, stability and convergence of regularized solutions in the sense of Proposition 1. Now we are going to discuss conditions under which weak*-to-weak continuity of $A : \ell^1 \to Y$ can be obtained. The occurring cross connections are relevant in order to ensure existence, stability and convergence of regularized solutions, but they have also an essential impact on convergence rates which will be discussed in Sect. 4.

Proposition 2 *Let $A : \ell^1 \to Y$ with adjoint operator $A^* : Y^* \to \ell^\infty$ satisfy the condition*

$$\mathscr{R}(A^*) \subseteq c_0, \tag{8}$$

where c_0 is the Banach space of real-valued sequences converging to zero equipped with the supremum norm. Then A is weak-to-weak continuous. In particular, (8) is fulfilled whenever there exist, for all $k \in \mathbb{N}$, source elements $f^{(k)} \in Y^*$ such that the system of source conditions*

$$e^{(k)} = A^* f^{(k)} \tag{9}$$

holds true, where $\{e^{(k)}\}_{k\in\mathbb{N}}$ is the sequence of k-th standard unit vectors which forms a Schauder basis in c_0. Under the condition (9) we even have the equality

$$\overline{\mathscr{R}(A^*)}^{\ell^\infty} = c_0. \tag{10}$$

The paper [1] shows that the condition (9), originally introduced by Grasmair in [19], can be verified for a wide class of applied linear inverse problems. But as also the counterexamples in [12] indicate, it may fail if the underlying basis smoothness is insufficient. However, weak*-to-weak continuity of A can be reformulated in several ways as the following proposition, proven in [9, Lemma 2.1], shows. This proposition brings more order into the system of conditions.

Proposition 3 *The three assertions*

(i) $\{Ae^{(k)}\}_{k\in\mathbb{N}}$ *converges in Y weakly to zero, i.e.* $Ae^{(k)} \rightharpoonup 0$ *as* $k \to \infty$,
(ii) $\mathscr{R}(A^*) \subseteq c_0$,
(iii) *A is weak*-to-weak continuous,*

are equivalent.

As outlined in [5], the operator equation (1) with operator $A : \ell^1 \to Y$ is often motivated by a background operator equation $\tilde{A}\tilde{x} = y$ with an injective and bounded linear operator $\tilde{A}$ mapping from the infinite dimensional Banach space $\tilde{X}$ with uniformly bounded Schauder basis $\{u^{(k)}\}_{k\in\mathbb{N}}$, i.e. $\|u^{(k)}\|_{\tilde{X}} \leq K < \infty$, to the

Banach space Y. Here, following the setting in [19] we take into account a *synthesis operator* $L : \ell^1 \to \tilde{X}$ defined as $Lx := \sum_{k=1}^{\infty} x_k u^{(k)}$ for $x = (x_1, x_2, \ldots) \in \ell^1$, which is a well-defined, injective and bounded linear operator, and so is the composite operator $A = \tilde{A} \circ L : \ell^1 \to Y$. In particular A is always weak*-to-weak continuous if A has a bounded extension to ℓ^p, $1 < p < \infty$, as this yields (i) in Proposition 3. Even more specific, A is weak*-to-weak continuous if $\tilde{X}$ is a Hilbert space. Since this case appears rather often in practice, the continuity property comes "for free" in this situation.

3 Ill-Posedness and Conditional Stability

In this section, we discuss ill-posedness phenomena of the operator equation (1) based on Nashed's definition from [27], which we formulate in the following as Definition 1 for the simplified case of an injective bounded linear operator. Moreover, we draw a connecting line to the phenomenon of conditional well-posedness characterized by conditional stability estimates, which yield for appropriate choices of the regularization parameter convergence rates in Tikhonov-type regularization.

Definition 1 The operator equation $Ax = y$ with an injective bounded linear operator $A : X \to Y$ mapping between infinite dimensional Banach spaces X and Y is called *well-posed* if the range $\mathscr{R}(A)$ of A is a closed subset of Y, otherwise the equation is called *ill-posed*. In the ill-posed case, we call the equation *ill-posed of type I* if $\mathscr{R}(A)$ contains an infinite dimensional closed subspace and otherwise *ill-posed of type II*.

The following proposition taken from [13, Propositions 4.2 and 4.4] and the associated Fig. 1 give some more insight into the different situations distinguished in Definition 1.

Proposition 4 *Consider the operator equation $Ax = y$ from Definition 1. If this equation is well-posed, i.e., $\mathscr{R}(A) = \overline{\mathscr{R}(A)}^{Y}$ and there is some constant $\underline{c} > 0$*

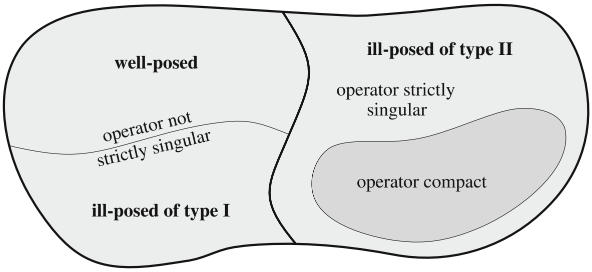


Fig. 1 Properties of A for well-posedness and ill-posedness types of equations from Definition 1

such that $\|Ax\| \geq \underline{c}\,\|x\|$ for all $x \in X$ or the equation is ill-posed of type I, then the operator A is non-compact. Consequently, compactness of A implies ill-posedness of type II. More precisely, for an ill-posed equation $Ax = y$ with injective A and infinite dimensional Banach spaces X and Y, ill-posedness of type II occurs if and only if A is strictly singular. This means that the restriction of A to an infinite dimensional subspace of X is never an isomorphism (linear homeomorphism). If X and Y are both Hilbert spaces and the equation is ill-posed, then ill-posedness of type II occurs if and only if A is compact.

Now we apply the case distinction of Definition 1, verified in detail in Proposition 4, to our situation of Eq. (1) with $X := \ell^1$ and $A : \ell^1 \to Y$. We start with a general observation in Proposition 5, which motivates the use of ℓ^1-regularization for the stable approximate solution of (1), because the equation is mostly ill-posed. Below we enlighten the cross connections a bit more by the discussion of some example situations.

Proposition 5 *If Y is a reflexive Banach space, then the operator equation* (1) *is always ill-posed of type II.*

Proof As a consequence of the theorem from [18] we have that every bounded linear operator $A : \ell^1 \to Y$ is strictly singular if Y is a reflexive Banach space. Hence well-posedness and ill-posedness of type I cannot occur in such case. □

Example 1 Consider that for reflexive Y we have a composition $A = \tilde{A} \circ L$ with forward operator $\tilde{A} : \tilde{X} \to Y$ and synthesis operator $L : \ell^1 \to \tilde{X}$ as mentioned in Sect. 2. Then (1) is ill-posed of type II even if $\tilde{A}$ is continuously invertible and hence the equation $\tilde{A}\tilde{x} = y$ well-posed. This may occur, for example, for Fredholm or Volterra integral equations of the second kind. Similarly, if $\tilde{A}$ as mapping between Hilbert spaces is non-compact with non-closed range and hence $\tilde{A}\tilde{x} = y$ is ill-posed of type I (which occurs, e.g., for multiplication operators mapping in $L^2(0,1)$), (1) is still ill-posed of type II. In the frequent case that $\tilde{X}$ is a separable Hilbert space and $\{u^{(k)}\}_{k\in\mathbb{N}}$ an orthonormal basis, then A is compact whenever $\tilde{A} : \tilde{X} \to Y$ is compact (occurring for example for Fredholm or Volterra integral equations of the first kind).

Example 2 If $A := \mathscr{E}_q$ with $1 \leq q < \infty$ and $Y := \ell^q$ is the embedding operator, then solving Eq. (1) based on noisy data $y^\delta \in \ell^q$ fulfilling (2) is a *denoising problem* (see also [13, Sect. 5] and [14, Example 6.1]). For $1 < q < \infty$ the embedding operator $A = \mathscr{E}_q$ is strictly singular with non-closed range but non-compact. Due to Proposition 5 the equation is ill-posed of type II. Moreover, we have $Ae^{(k)} \rightharpoonup 0$ in ℓ^q, which due to Proposition 3 implies that A is weak*-to-weak continuous and $\mathscr{R}(A^*) \subseteq c_0$. The latter is obvious, because the adjoint A^* is the embedding operator from ℓ^{q^*} to ℓ^∞ with $1/q + 1/q^* = 1$ and $\mathscr{R}(A^*) = \ell^{q^*}$. In particular, the source condition (9) applies with $f^{(k)} = e^{(k)} \in \ell^{q^*} \subset c_0$ for all $k \in \mathbb{N}$.

Example 3 For $q = 1$ in the previous example we have the continuously invertible identity operator $A = Id : \ell^1 \to \ell^1$ with closed range $\mathscr{R}(A) = \ell^1$. Then Eq. (1) is *well-posed*, but we have $Ae^{(k)} \not\rightharpoonup 0$ in ℓ^1 for $k \to \infty$, which due to Proposition 3 indicates that the range $\mathscr{R}(A^*)$ of the adjoint of A does not belong to c_0 and in particular A is not weak*-to-weak continuous. This is evident, because the adjoint of $A = Id$ is the identity $A^* = Id : \ell^\infty \to \ell^\infty$ and $\mathscr{R}(A^*) = \ell^\infty$. We will come back to this example later.

For obtaining error estimates in ℓ^1-regularization on which convergence rates are based, we need some kind of conditional well-posedness in order to overcome the ill-posedness of Eq. (1). Well-posed varieties of Eq. (1) yield stability estimates $\|x - x^\dagger\|_{\ell^1} \le K\|Ax - Ax^\dagger\|_Y$ for all $x \in \ell^1$, which under (2) and for the choice $\alpha = \alpha_{SDP}$ imply the best possible rate

$$\|x_\alpha^\delta - x^\dagger\|_{\ell^1} = O(\delta) \quad \text{as} \quad \delta \to 0\,, \tag{11}$$

which is typical for well-posed situations. We will come back to this in Sect. 6. We say that a *conditional stability estimate* holds true if there is a subset $\mathscr{M} \subset \ell^1$ such that

$$\|x - x^\dagger\|_{\ell^1} \le K(\mathscr{M})\|Ax - Ax^\dagger\|_Y \quad \text{for all} \quad x \in \mathscr{M}\,. \tag{12}$$

Because $\mathscr{M}$ is not known a priori, such kind of stability requires the additional use of regularization for bringing the approximate solutions to $\mathscr{M}$ such that a rate (11) can be verified. This idea was first published in [7] by Cheng and Yamamoto. In the context of ℓ^1-regularization for our Eq. (1), we have estimates of the form (12) if the solution $x^\dagger \in \ell^0$ is *sparse*, i.e. only a finite number of non-zero components occur in the infinite sequence $x^\dagger$. Then $\mathscr{M}$ can be considered as a subset of ℓ^0 with specific properties, and the sparsity of ℓ^1-regularized solutions verified in Proposition 1 ensures that the corresponding approximate solutions belong to $\mathscr{M}$. This implies the rate (11) for $x^\dagger \in \ell^0$, although Eq. (1) is not well-posed.

A similar but different kind of conditional well-posedness estimates are *variational source conditions*, which attain in our setting the form

$$\beta\,\|x - x^\dagger\|_{\ell^1} \le \|x\|_{\ell^1} - \|x^\dagger\|_{\ell^1} + \varphi(\|Ax - Ax^\dagger\|_Y) \quad \text{for all} \quad x \in \ell^1\,, \tag{13}$$

satisfied for a constant $0 < \beta \le 1$ and some concave index function φ. From [23, Theorems 1 and 2] we find directly the convergence rates results of the subsequent proposition.

Proposition 6 *If the variational source condition* (13) *holds true for a constant* $0 < \beta \le 1$ *and some concave index function* φ, *then we have for* ℓ^1*-regularized solutions* x_α^δ *the convergence rate*

$$\|x_\alpha^\delta - x^\dagger\|_{\ell^1} = O(\varphi(\delta)) \quad as \quad \delta \to 0 \tag{14}$$

whenever the regularization parameter α *is chosen either a priori as* $\alpha = \alpha_{APRI}$ *according to* (6) *or a posteriori as* $\alpha = \alpha_{SDP}$ *according to* (7).

Consequently, for the manifestation of convergence rates results in the next section it remains to find constants β, concave index functions φ and sufficient conditions for the verification of corresponding variational inequalities (13).

4 Convergence Rates Results for ℓ^1-Regularization

The first step to derive a variational source condition (13) at the solution point $x^\dagger = (x_1^\dagger, x_2^\dagger, \ldots) \in \ell^1$ was taken by Lemma 5.1 in [5], where the inequality

$$\|x - x^\dagger\|_{\ell^1} \leq \|x\|_{\ell^1} - \|x^\dagger\|_{\ell^1} + 2\left(\sum_{k=n+1}^{\infty} |x_k^\dagger| + \sum_{k=1}^{n} |x_k - x_k^\dagger|\right) \tag{15}$$

was proven for all $x = (x_1, x_2, \ldots) \in \ell^1$ and all $n \in \mathbb{N}$. Then under the source condition (9), valid for all $k \in \mathbb{N}$, one directly finds

$$\sum_{k=1}^{n} |x_k - x_k^\dagger| = \sum_{k=1}^{n} |\langle e^{(k)}, x - x^\dagger\rangle_{\ell^\infty \times \ell^1}| = \sum_{k=1}^{n} |\langle f^{(k)}, A(x - x^\dagger)\rangle_{Y^* \times Y}| \tag{16}$$

and hence from (15) that a function of type

$$\varphi(t) = 2 \inf_{n \in \mathbb{N}} \left(\sum_{k=n+1}^{\infty} |x_k^\dagger| + \gamma_n\, t\right) \tag{17}$$

with $\beta = 1$ and

$$\gamma_n = \sum_{k=1}^{n} \|f^{(k)}\|_{Y^*} \tag{18}$$

provides us with a variational inequality (13). Along the lines of the proof of [5, Theorem 5.2] one can show the assertion of the following lemma.

Lemma 2 *If $\{\gamma_n\}_{n\in\mathbb{N}}$ is a non-decreasing sequence, then φ from (17) is a well-defined and concave index function for all $x^\dagger \in \ell^1$.*

Both the decay rate of $x_k^\dagger \to 0$ as $k \to \infty$ and the behaviour of γ_n as $n \to \infty$ in (17) have impact on the resulting rate function φ. A power-type decay of $x_k^\dagger$ leads to Hölder convergence rates (see [5, Example 5.3] and [13, Example 3.4]), whereas exponential decay of $x_k^\dagger$ leads to near-to-δ rates slowed down by a logarithmic factor (see and [3, Example 3.5] and [13, Example 3.5]). In the case that $x^\dagger \in \ell^0$ is sparse with $x_k^\dagger = 0$ for all $k > n_0$, then the best possible rate (11) is seen. This becomes clear from formula (17), because then φ fulfills the inequality $\varphi(t) \leq 2\gamma_{n_0}\, t$.

From Proposition 6 we have that for all concave index functions φ from (17) a convergence rate (14) for the ℓ^1-regularization takes place in the case of appropriate choices of the regularization parameter α whenever a constant $0 < \beta \leq 1$ exists such that (13) is valid with φ from (17). When the condition (9) is valid, this is the case with $\beta = 1$ and γ_n from (18). Under the same condition the rate was slightly improved in [13] (see also [14]) by showing that γ_n from (18) can be replaced with

$$\gamma_n = \sup_{\substack{a_k \in \{-1,0,1\} \\ k=1,\ldots,n}} \left\| \sum_{k=1}^{n} a_k f^{(k)} \right\|_{Y^*} . \tag{19}$$

However, the condition (9) may fail as was noticed first in [12] for a bidiagonal operator. Therefore, assumption (9) was replaced by a weaker (but not particularly eye-pleasing) one in [12]. Ibid the authors assume, in principle, that for each $n \in \mathbb{N}$ there are elements $f^{(n,k)}$ such that for all $1 \leq i \leq n$

$$[A^* f^{(n,k)}]_i = [e^{(k)}]_i$$

and

$$\left| \sum_{k=1}^{n} [A^* f^{(n,k)}]_i \right| \leq c \quad \text{for all} \quad i > n \quad \text{and} \quad c < 1.$$

This means that each basis vector $e^{(k)}$ can be approximated exactly up to arbitrary position but with a non-zero tail consisting of sufficiently small elements. Later, in [14], a more clearly formulated property was assumed which implies the one from [12]. We give a slightly reformulated version of this property in the following. In this context, we notice that P_n denotes the projection operator applied to elements $x = (x_1, x_2, \ldots, x_n, x_{n+1}, \ldots)$ such that $P_n x = (x_1, x_2, \ldots, x_n, 0, 0, \ldots)$.

Property 1 For arbitrary $\mu \in [0, 1)$, we have a real sequence $\{\gamma_n\}_{n \in \mathbb{N}}$ such that for each $n \in \mathbb{N}$ and each $\xi = \xi(n) \in \ell^\infty$, with

$$\xi_k \begin{cases} \in [-1, 1], & \text{if } k \leq n, \\ = 0, & \text{if } k > n \end{cases} , \tag{20}$$

there exists some $\eta = \eta(\mu, n, \xi) \in Y^*$ satisfying

(a) $P_n A^* \eta = \xi$,
(b) $|[(I - P_n) A^* \eta]_k| \leq \mu \quad$ for all $\; k > n$,
(c) $\|\eta\|_{Y^*} \leq \gamma_n$.

It is important to note that it was a substantial breakthrough in the recent paper [11] to show that Property 1 follows directly from injectivity and weak*-to-weak continuity of the operator A. Namely, the following proposition was proven there.

Note that we changed the definition of the ξ in (20) slightly. By checking the proofs in the original paper one sees the amendments we made are not relevant.

Proposition 7 *Let $A : \ell^1 \to Y$ be bounded, linear and weak*-to-weak continuous. Then the following assertions are equivalent.*

(i) *A is injective,*
(ii) $e^{(k)} \in \overline{\mathscr{R}(A^*)}^{\ell^\infty}$ *for all* $k \in \mathbb{N}$,
(iii) $\overline{\mathscr{R}(A^*)}^{\ell^\infty} = c_0$,
(iv) *Property 1 holds.*

In other words, for such operators there exist appropriate sequences $\{\gamma_n\}_{n\in\mathbb{N}}$ occurring in (17) such that a variational source condition (13) holds for an index function φ from (17) and constant $\beta = \frac{1-\mu}{1+\mu}$ (see Proposition 1 below). Item (b) in Property 1 is a generalization of (9). Namely, the canonical basis vectors $e^{(k)}$ do not necessarily belong to the range of A^* but to its closure. For the proof of Proposition 7 we refer to [11]. Most of the steps are identical or at least similar to the proof of Proposition 11 which we will give later.

5 Extensions to Non-reflexive Image Spaces

If the injective bounded linear operator $A : \ell^1 \to Y$ fails to be weak*-to-weak continuous, then the results of the preceding section do not apply. In case that Y is a non-reflexive Banach space, it makes sense to consider the weaker property of weak*-to-weak* continuity of A. An already mentioned example is the identity mapping $A = Id$ for $Y = \ell^1$. In ℓ^1, weak convergence and norm convergence coincide (Schur property), but there is no coincidence with weak* convergence. Thus, the identity mapping cannot be weak*-to-weak continuous, but it is weak*-to-weak* continuous as the following Proposition 8 shows. It is a modified extension of Proposition 3. Following [10, Lemma 6.5] we formulate this extension and repeat below the relevant proof details.

Proposition 8 *Let Z be a separable Banach space which acts as a predual space for the Banach space $Y = Z^*$. Then the following four assertions are equivalent.*

(i) $\{Ae^{(k)}\}_{k\in\mathbb{N}}$ *converges in Y weakly* to zero,*
(ii) $\mathscr{R}(A^*|_Z) := \{v \in \ell^\infty : v = A^*z$ *for some* $z \in Z \subseteq Y^*\} \subseteq c_0$,
(iii) *A is weak*-to-weak* continuous,*
(iv) *There is a bounded linear operator $S : Z \to c_0$ such that $A = S^*$.*

Proof Let (i) be satisfied. Then for each A^*z from $\mathscr{R}(A^*|_Z)$ we have

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

This yields $A^*z \in c_0$ and hence (ii) is valid.

$$\begin{tikzcd}
Z \subseteq Y^* \arrow[d, "S"'] & Y = Z^* & Y^* = Z^{**} \arrow[d, "A^* = S^{**}"'] \\
c_0 & \ell^1 = (c_0)^* \arrow[u, "A = S^*"] & \ell^\infty = (\ell^1)^* = (c_0)^{**}
\end{tikzcd}$$

Fig. 2 Schematic display of the operators and underlying spaces needed in this section. Top and bottom row: the (separable) Banach spaces under consideration. Middle row: the operators we work with

Now let (ii) be true. If we take a weakly* convergent sequence $x_n \rightharpoonup^* x_0$ in ℓ^1 as $n \to \infty$, then $\langle Ax_n, z\rangle_{Z^* \times Z} = \langle z, Ax_n\rangle_{Y^* \times Y} = \langle A^*z, x_n\rangle_{\ell^\infty \times \ell^1}$ for all z in Z. Because moreover A^*z belongs to c_0 and ℓ^1 is the dual of c_0, we may write this as $\langle A^*z, x_n\rangle_{\ell^\infty \times \ell^1} = \langle x_n, A^*z\rangle_{\ell^1 \times c_0}$. Thus,

$$\lim_{n\to\infty} \langle Ax_n, z\rangle_{Z^* \times Z} = \lim_{n\to\infty} \langle x_n, A^*z\rangle_{\ell^1 \times c_0} = \langle x_0, A^*z\rangle_{\ell^1 \times c_0} = \langle Ax_0, z\rangle_{Z^* \times Z} \quad \text{for all } \; z \in Z,$$

which proves condition (iii). From (iii) and the fact that $e^{(k)} \rightharpoonup^* 0$ in ℓ^1 as $k \to \infty$ we immediately obtain (i). Finally, the equivalence between (iii) and (iv) can be found, e.g., in [26, Theorem 3.1.11]. □

As a consequence of item (iv) in Proposition 8, each weak*-to-weak* continuous linear operator is automatically bounded. Figure 2 illustrates the connection between the different spaces and operators we juggle around in this section.

For the identity mapping $A = Id : \ell^1 \to Y$ with $Y = \ell^1$ and predual $Z = c_0$, property (i) of Proposition 8 is trivially satisfied which yields the weak*-to-weak* continuity of this operator. Note that the case $Y = \ell^1, A = Id$ is only of theoretical interest. Precisely, 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 choice $p = 1$ in (3), where we have $Y = \ell^1$, the ℓ^1-regularized solutions coincide with the data y^δ if $\alpha < 1$ and we have the best possible rate (11).

Main parts of the above mentioned Proposition 1 on existence, stability and convergence of ℓ^1-regularized solutions x_α^δ remain true if $A : \ell^1 \to Y$ is only weak*-to-weak* continuous. The sparsity property $x_\alpha^\delta \in \ell^0$, however, will fail in general (consider the example of the identity as mentioned above). Existence, stability and convergence assertions remain valid, because their proofs basically rely on the fact that the mapping $x \mapsto \|Ax - y^\delta\|_Y$ is a weakly* lower semicontinuous functional. This is the case in both variants, with or without *, since the norm functional is weakly and also weakly* lower semicontinuous. For the existence of regularized solutions (minimizers of the Tikhonov functional (3)) again the Banach-Alaoglu theorem (Lemma 1) is required and yields weakly* compact level sets of the ℓ^1-norm functional.

Our goal is to proof an analogue to Proposition 7 for weak*-to-weak* continuous operators. We start with a first observation.

Proposition 9 *Let $A : \ell^1 \to Y$ be injective and weak*-to-weak* continuous and let $Y = Z^*$ for some Banach space Z. Then*

$$\overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty} = c_0.$$

Proof From item (iv) of Proposition 8 we take the operator $S : Z \to c_0$ with $A = S^*$. As A is injective, i.e., $\mathscr{N}(A) = \{0\}$, it follows $\overline{\mathscr{R}(S)}^{c_0} = \mathscr{N}(S^*)_\perp = \mathscr{N}(A)_\perp = c_0$. There, the subscript denotes the pre-annihilator for a set V, in our situation with $(c_0)^* = \ell^1$ and $V \subseteq \ell^1$ defined as

$$V_\perp := \{x \in c_0 : \langle \zeta, x\rangle_{\ell^1 \times c_0} = 0 \quad \forall \zeta \in V\}.$$

Let $\eta \in Z$ and recall $Y^* = Z^{**}$ (cf. Fig. 2). Then for each $x \in \ell^1$

$$\begin{aligned}\langle A^*\, \eta, x\rangle_{\ell^\infty \times \ell^1} &= \langle \eta, A\,x\rangle_{Y^* \times Y} = \langle \eta, A\,x\rangle_{Z \times Y} = \langle S\,\eta, x\rangle_{c_0 \times \ell^1}\\ &= \langle S\,\eta, x\rangle_{\ell^\infty \times \ell^1},\end{aligned}$$

i.e., $A^*|_Z = S$. Thus $\overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty} = \overline{\mathscr{R}(S)}^{c_0} = c_0$. At this point we emphasize that in both Banach spaces ℓ^∞ and c_0 the same supremum norm applies. □

We will show in Proposition 11 that conversely $\overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty} = c_0$ implies injectivity for weak*-to-weak* continuous operators. Before doing so we need the following Proposition which coincides in principle with [11, Proposition 9].

Proposition 10 *Let A be injective and weak*-to-weak* continuous. Moreover, let $\varepsilon > 0$ and $n \in \mathbb{N}$. Then for each $\xi \in c_0$ there exists $\tilde{\xi} \in \mathscr{R}(A^*)$ such that*

$$\tilde{\xi}_k = \xi_k \quad \textit{for} \quad k \le n \qquad \textit{and} \qquad |\tilde{\xi}_k - \xi_k| \le \varepsilon \quad \textit{for} \quad k > n.$$

Proof We proof the proposition by induction with respect to n. For $\xi \in c_0$ set

$$\xi^+ := (\xi_1 + \varepsilon, \xi_2, \xi_3, \ldots) \qquad \text{and} \qquad \xi^- := (\xi_1 - \varepsilon, \xi_2, \xi_3, \ldots).$$

By Proposition 9 we have that $c_0 = \overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty} \subset \overline{\mathscr{R}(A^*)}^{\ell^\infty}$. Hence we find elements $\tilde{\xi}^+ \in \mathscr{R}(A^*)$ and $\tilde{\xi}^- \in \mathscr{R}(A^*)$ with

$$\|\tilde{\xi}^+ - \xi^+\|_{\ell^\infty} \le \varepsilon \qquad \text{and} \qquad \|\tilde{\xi}^- - \xi^-\|_{\ell^\infty} \le \varepsilon.$$

Consequently, $\tilde{\xi}_1^+ \ge \xi_1 \ge \tilde{\xi}_1^-$ and $|\tilde{\xi}_k^+ - \xi_k| \le \varepsilon$ as well as $|\tilde{\xi}_k^- - \xi_k| \le \varepsilon$ 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 $|\tilde{\xi}_k - \xi_k| \le \varepsilon$ 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 $\xi \in c_0$ 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 $\tilde{\xi}^+ \in \mathscr{R}(A^*)$ and $\tilde{\xi}^- \in \mathscr{R}(A^*)$ with

$$\tilde{\xi}_k^+ = \xi_k = \tilde{\xi}_k^- \quad \text{for } k \le m$$

and

$$|\tilde{\xi}_k^+ - \xi_k^+| \le \varepsilon \quad \text{and} \quad |\tilde{\xi}_k^- - \xi_k^-| \le \varepsilon \quad \text{for } k > m.$$

Consequently, $\tilde{\xi}_{m+1}^+ \ge \xi_{m+1} \ge \tilde{\xi}_{m+1}^-$ and $|\tilde{\xi}_k^+ - \xi_k| \le \varepsilon$ as well as $|\tilde{\xi}_k^- - \xi_k| \le \varepsilon$ 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 $|\tilde{\xi}_k - \xi_k| \le \varepsilon$ for $k > m + 1$, which proves the proposition for $n = m + 1$. □

Now we come to the main result of this section. The proof is similar and in part identical to the one of Proposition 12 in [11].

Proposition 11 *Let $A : \ell^1 \to Y$ be bounded, linear and weak*-to-weak* continuous. Then the following assertions are equivalent.*

(i) A is injective,
(ii) $\overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty} = c_0$,
(iii) $e^{(k)} \in \overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty}$ for all $k \in \mathbb{N}$,
(iv) Property 1 holds.

Proof We show (i) ⇒ (iv) ⇒ (iii) ⇒ (ii) ⇒ (i).

(i)⇒(iv): Fix $\mu \in (0, 1)$, $n \in \mathbb{N}$ and take some ξ as described in Property 1. By Proposition 10 with $\varepsilon := \mu$ there exists some η such that $A^*\eta$ ($= \tilde{\xi}$ in the proposition) satisfies items (a) and (b) in Property 1. In particular we have $\{\eta_k\}_{k\in 1,\ldots,n}$ such that $P_n A^* \eta_k = e^{(k)}$ and $|[(I - P_n)A^*\eta_k]_i| \le \frac{\mu}{n}$ for all $i > n$. Since $\xi \in c_0$ it is

$$\xi = \sum_{k=1}^{n} c_k e^{(k)} = \sum_{k=1}^{n} c_k A^* \eta_k = A^* \left(\sum_{k=1}^{n} c_k \eta_k \right),$$

for coefficients $-1 \le c_k \le 1$, i.e., $\xi = A^*\eta$ with $||\eta|| \le \sum\limits_{k=1}^{n} ||\eta_k||$ as an upper bound for γ_n. By construction this η also fulfills $|[(I-P_n)A^*\eta]_i| \le \sum\limits_{i=1}^{n} |[(I-P_n)A^*\eta_k]_i| \le \mu$.

(iv)⇒(iii): Fix k, fix $n \geq k$, take a sequence $(\mu_m)_{m\in\mathbb{N}}$ in $(0, 1)$ with $\mu_m \to 0$ and choose $\xi := e^{(k)}$ in Property 1. Then for a corresponding sequence $(\eta_m)_{m\in\mathbb{N}}$ from Property 1 we obtain

$$\|e^{(k)} - A^*\eta_m\|_{\ell^\infty} \leq \|e^{(k)} - P_nA^*\eta_m\|_{\ell^\infty} + \|(I - P_n)A^*\eta_m\|_{\ell^\infty}.$$

The first summand is zero by the choice of ξ and the second summand is bounded by μ_m. Thus, $\|e^{(k)} - A^*\eta_m\|_{\ell^\infty} \to 0$ if $m \to \infty$.

(iii)⇒(ii): $(e^{(k)})_{k\in\mathbb{N}}$ is a Schauder basis in c_0. Thus, $c_0 \subseteq \overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty}$. Proposition 8 yields $\overline{\mathscr{R}(A^*|_Z)}^{\ell^\infty} \subseteq c_0$. Hence $\overline{\mathscr{R}(A^*|_Z)} = c_0$.

(ii)⇒(i): One easily shows that $\overline{\mathscr{R}(A^*)}^{\ell^\infty} \subseteq \mathscr{N}(A)^\perp$. Thus, $c_0 \subseteq \mathscr{N}(A)^\perp$. If we have some $x \in \ell^1$ with $Ax = 0$, then for each $u \in c_0 \subseteq \mathscr{N}(A)^\perp$ we obtain

$$\langle x, u\rangle_{\ell^1\times c_0} = \langle u, x\rangle_{\ell^\infty\times\ell^1} = 0,$$

which is equivalent to $x = 0$. □

Since, in the context of both Propositions 7 and 11, the injectivity of A yields Property 1, the consequences with respect to variational source conditions and convergence rate results are identical for a weak*-to-weak and a weak*-to-weak* continuous operator A. We formulate the following theorem and the subsequent corollary and prove the theorem for a weak*-to-weak* continuous operator $A : \ell^1 \to Y$. In particular, the corollary requires the existence of a separable predual space Z of Y in order to apply Lemma 1 and to ensure the stabilizing property of the Tikhonov penalty. However, the proof of the theorem repeats point by point the ideas of the proof from [11, Corollary 11] focused on weak*-to-weak continuous operators A.

Theorem 1 *Let the bounded linear operator $A : \ell^1 \to Y$ be injective and weak*-to-weak* continuous, where we additionally assume that the Banach space Y possesses a separable predual Banach space Z with $Z^* = Y$. Moreover, let $\mu \in [0, 1)$ and $\{\gamma_n\}_{n\in\mathbb{N}}$ be such that Property 1 is fulfilled. Then a variational source condition (13) with the constant $\beta = \frac{1-\mu}{1+\mu} \in [0, 1)$ and the concave index function φ given by (17) is fulfilled.*

Proof Fix $n \in \mathbb{N}$ and $x \in \ell^1$ and let $\xi := \operatorname{sgn} P_n(x - x^\dagger) \in \ell^\infty$ be the sequence of signs of $P_n(x - x^\dagger)$. Then by Property 1 there is some η such that

$$\begin{aligned}\sum_{k=1}^{n} |x_k - x_k^\dagger| &= \langle \xi, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} = \langle P_nA^*\eta, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} \\ &= \langle P_nA^*\eta - A^*\eta, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} + \langle A^*\eta, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} \\ &= -\langle (I - P_n)A^*\eta, (I - P_n)(x - x^\dagger)\rangle_{\ell^\infty\times\ell^1} + \langle A^*\eta, x - x^\dagger\rangle_{\ell^\infty\times\ell^1} \\ &\leq \mu\|(I - P_n)(x - x^\dagger)\|_{\ell^1} + \gamma_n\|Ax - Ax^\dagger\|_Y. \end{aligned} \tag{21}$$

The triangle inequality yields

$$\|P_n(x-x^\dagger)\|_{\ell^1} \le \mu\big(\|(I-P_n)x\|_{\ell^1} + \|(I-P_n)x^\dagger\|_{\ell^1}\big) + \gamma_n\|Ax-Ax^\dagger\|_Y. \qquad (22)$$

Now

$$\begin{aligned}&\beta\|x-x^\dagger\|_{\ell^1} - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1}\\ &= \beta\|P_n(x-x^\dagger)\|_{\ell^1} + \beta\|(I-P_n)(x-x^\dagger)\|_{\ell^1} - \|P_nx\|_{\ell^1} - \|(I-P_n)x\|_{\ell^1}\\ &\quad + \|P_nx^\dagger\|_{\ell^1} + \|(I-P_n)x^\dagger\|_{\ell^1}\end{aligned}$$

together with

$$\beta\|(I-P_n)(x-x^\dagger)\|_{\ell^1} \le \beta\|(I-P_n)x\|_{\ell^1} + \beta\|(I-P_n)x^\dagger\|_{\ell^1}$$

and

$$\|P_nx^\dagger\|_{\ell^1} = \|P_n(x-x^\dagger-x)\|_{\ell^1} \le \|P_n(x-x^\dagger)\|_{\ell^1} + \|P_nx\|_{\ell^1}$$

shows

$$\begin{aligned}&\beta\|x-x^\dagger\|_{\ell^1} - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1}\\ &\quad \le 2\|(I-P_n)x^\dagger\|_{\ell^1} + (1+\beta)\|P_n(x-x^\dagger)\|_{\ell^1}\\ &\qquad - (1-\beta)\big(\|(I-P_n)x\|_{\ell^1} + \|(I-P_n)x^\dagger\|_{\ell^1}\big).\end{aligned}$$

Combining this estimate with the previous estimate (22) and taking into account that $\beta = \frac{1-\mu}{1+\mu}$ and $\mu = \frac{1-\beta}{1+\beta}$ we obtain for all $x \in \ell^1$

$$\begin{aligned}\beta\|x-x^\dagger\|_{\ell^1} - \|x\|_{\ell^1} + \|x^\dagger\|_{\ell^1} &\le 2\|(I-P_n)x^\dagger\|_{\ell^1} + \frac{2}{1+\mu}\gamma_n\|Ax-Ax^\dagger\|_Y\\ &\le 2\|(I-P_n)x^\dagger\|_{\ell^1} + 2\gamma_n\|Ax-Ax^\dagger\|_Y.\end{aligned}$$

Taking the infimum over all $n \in \mathbb{N}$ completes the proof. □

The variational source condition immediately yields convergence rates according to Proposition 6.

Corollary 1 *Under the conditions of Theorem 1 the ℓ^1-regularized solutions x_α^δ as minimizers of* (3) *fulfil*

$$\|x_\alpha^\delta - x^\dagger\|_{\ell^1} = O(\varphi(\delta)) \quad \textit{as} \quad \delta \to 0\,,$$

with the concave index function φ *from* (17)*, whenever the regularization parameter* α *is chosen either a priori as* $\alpha = \alpha_{APRI}$ *according to* (6) *or a posteriori as* $\alpha = \alpha_{SDP}$ *according to* (7)*.*

In order to familiarize the reader with the concepts in this work we will look at a particular operator to exemplify our theory. In particular we verify Property 1.

Example 4 Let $X = Y = \ell^1$ and

$$[Ax]_k = x_k + x_{k+1}, \qquad k \in \mathbb{N}.$$

In other words, A maps $x = (x_1, x_2, x_3, \dots)$ to $Ax = (x_1 + x_2, x_2 + x_3, x_3 + x_4, \dots)$. Clearly A is linear. Observe that $||Ax||_{\ell^1} \leq 2||x||_{\ell^1}$ and hence A is bounded. One easily verifies the adjoint $A^* : \ell^\infty \to \ell^\infty$,

$$A^* y = (y_1, y_2 + y_1, y_3 + y_2, y_3 + y_4, \dots).$$

Both A and A^* are injective. Since $\overline{\mathscr{R}(A)}^{\ell^1} = \mathscr{N}(A^*)_\perp$, where

$$\mathscr{N}(A^*)_\perp := \{x \in \ell^1 : \langle y, x\rangle_{\ell^\infty \times \ell^1} = 0 \ \ \forall y \in \mathscr{N}(A^*)\} = \ell^1$$

we have $\overline{\mathscr{R}(A)}^{\ell^1} = \ell^1$. It is however easy to see that $\mathscr{R}(A) \neq \ell^1$. For example there is no $x \in \ell^1$ such that $Ax = e^{(2)}$. Namely, solving $Ax = e^{(2)}$ for x leads to the system of equations $x_1 = -x_2, x_2 = 1 - x_3, x_4 = -x_3, x_5 = -x_4$, etc. Due to the alternating character again there is no $x \in \ell^1$ that satisfies this system. We have shown that $\overline{\mathscr{R}(A)}^{\ell^1} \neq \mathscr{R}(A)$, i.e., the corresponding operator equation (1) is ill-posed.

Next we prove that A is weak*-to-weak* continuous but not weak*-to-weak continuous. To this end we use the properties (i) in Propositions 3 and 8, respectively. First let $\xi \in c_0$. Then, with

$$[Ae^{(k)}]_i = \begin{cases} 1 & i = k, k-1 \\ 0 & else \end{cases} \qquad \forall i \geq 2$$

it is

$$\langle \xi, Ae^{(k)}\rangle_{c_0 \times \ell^1} = \xi_{k-1} + \xi_k \to 0\,,$$

since $\xi \in c_0$. For $\xi \in \ell^\infty$, however,

$$\langle \xi, Ae^{(k)}\rangle_{\ell^\infty \times \ell^1} = \xi_{k-1} + \xi_k$$

does in general not converge to zero (let, e.g., $\xi \equiv 1$). Summarizing the properties of the forward operator for the present example, we note that A is linear, injective, weak*-to-weak* continuous, but not weak*-to-weak continuous. Moreover its range is not closed such that the corresponding operator equation (1) is ill-posed.

For this particular operator let us investigate Property 1. We will see in the following that this actually holds with $\mu = 0$ and $\gamma_n = n$, i.e., $e^{(k)} \in \mathscr{R}(A^*)$ for all $k \in \mathbb{N}$. We will also show that $e^{(k)} \in \overline{\mathscr{R}(A^*|_{c_0})}^{\ell^\infty}$ according to item (iii) of Proposition 11. Even then we still have $\gamma_n = n$ for arbitrary $0 < \mu < 1$.

Fix an arbitrary $n \in \mathbb{N}$ and let $\xi = (\xi_1, \xi_2, \ldots, \xi_n, 0, \ldots) \in \ell^\infty$ with $\xi_i \in [-1, 1]$. We are looking for an $\eta \in \ell^\infty$ satisfying Property 1. Observe that, by definition of A^* and for given ξ, any η satisfying $P_n A^* \eta = \xi$ has the structure $\eta_1 = \xi_1$, $\eta_2 = \xi_2 - \eta_1$, $\eta_3 = \xi_3 - \eta_2$ and so on until $\eta_n = \xi_n - \eta_{n-1}$. In other words it is $\eta_n = \sum_{i=1}^n (-1)^{n-i} \xi_i$ and

$$||\eta_n||_{\ell^\infty} \leq n,$$

which yields item (c) of Property 1 with $\gamma_n = n$. Now fix arbitrary $0 < \mu < 1$. We have $[A^*\eta]_{n+1} = \eta_{n+1} + \eta_n$ and require $|\eta_{n+1} + \eta_n| \leq \mu$. Thus we may take any η_{n+1} with

$$-\eta_n - \mu \leq \eta_{n+1} \leq -\eta_n + \mu.$$

Analogously we find that in general

$$(-1)^i \eta_n - i\mu \leq \eta_{n+i} \leq (-1)^i \eta_n + i\mu, \quad i = 1, 2, \ldots.$$

Therefore, the choice of the tail of η is ambiguous. A viable pick is $\eta_{n+i} = (-1)^i \eta_n$. Then

$$\eta = (\eta_1, \eta_2, \ldots, \eta_n, -\eta_n, \eta_n, -\eta_n, \ldots) \tag{23}$$

with η_i, $1 \leq i \leq n$, as above and

$$A^*\eta = (\xi_1, \xi_2, \ldots, \xi_n, 0, 0, 0, \ldots).$$

In particular, this means that $e^{(k)} \in \mathscr{R}(A^*)$ (choose $\xi_i = 1$ and $\xi_j = 0$ for $i \neq j$). Note that $\eta \in \ell^\infty$ but $\eta \notin c_0$ in (23). However, we also see that for any ξ and arbitrary $0 < \mu < 1$ there are (infinitely many) choices for the tail of η such that $\eta \in c_0$ and item (b) of Property 1 holds. Independent of μ, all choices satisfy item (c) of Property 1 with $\gamma_n = n$. To set this into relation, we would obtain the same γ_n for a diagonal operator $\tilde{A} : \ell^2 \to \ell^2$ with singular values decaying as $\sigma_i \sim \frac{1}{\sqrt{i}}$, see [17].

Please note that in practice it is not necessary to verify Property 1 in the way we did here. In particular the sequential discrepancy principle (7) does not require the knowledge of any of the parameters from Property 1 in order to guarantee the convergence rates implied by that property.

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

Example 5 If $Y = \ell^1$ and

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

for all k in $\mathbb{N}$ and all x in ℓ^1, then $Ae^{(k)} = e^{(1)} + e^{(k)}$ if $k > 1$. Thus, $Ae^{(k)} \rightharpoonup^* e^{(1)} \neq 0$. 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. Note that, because of the first component, A does not have a bounded extension to any ℓ^p-space with $1 < p < \infty$.

6 The Well-Posed Case and Further Discussions

Proposition 12 *If the operator equation* (1) *is well-posed, i.e.* $\mathscr{R}(A) = \overline{\mathscr{R}(A)}^Y$, *then under the conditions of Theorem* 1 *the* ℓ^1*-regularized solutions* x_α^δ *as minimizers of* (3) *fulfil*

$$\|x_\alpha^\delta - x^\dagger\|_{\ell^1} = O(\delta) \quad as \quad \delta \to 0$$

whenever the regularization parameter α *is chosen either a priori as* $\alpha = \alpha_{APRI} \sim \delta^{p-1}$ *or a posteriori as* $\alpha = \alpha_{SDP}$ *according to* (7).

Proof The well-posedness condition $\mathscr{R}(A) = \overline{\mathscr{R}(A)}^Y$ implies $\mathscr{R}(A^*) = \overline{\mathscr{R}(A^*)}^{\ell^\infty}$ (cf. [26, Theorem 3.1.21]) and hence $V := \mathscr{R}(A^*)$ is a closed subspace of ℓ^∞, which can be considered as a Banach space with the same supremum norm as ℓ^∞. Then, for the injective operator $A : \ell^1 \to Y$, Banach's theorem concerning the continuity of the inverse operator ensures that the linear operator $(A^*)^{-1} : V \to Y^*$ is bounded. Moreover, the elements $\tilde{\xi} \in \mathscr{R}(A^*)$ in Proposition 10 associated to ξ from Property 1 satisfy the inequality $\|\tilde{\xi}\|_{\ell^\infty} \leq 1 + \varepsilon$, and with $\tilde{\xi} = A^*\eta$ we have $\|\eta\|_{Y^*} \leq \|(A^*)^{-1}\|_{V \to Y^*} (1 + \varepsilon) \leq K < \infty$. Hence, the sequence $\{\gamma_n\}_{n \in \mathbb{N}}$ in Property 1 is uniformly bounded by the finite positive constant K. Taking into account the proof of Theorem 1 we have with $\beta = \frac{1-\mu}{1+\mu}$ and for all $x \in \ell^1$

$$\beta \|x - x^\dagger\|_{\ell^1} \leq \|x\|_{\ell^1} - \|x^\dagger\|_{\ell^1} + 2\|(I - P_n)x^\dagger\|_{\ell^1} + 2K\|Ax - Ax^\dagger\|_Y,$$

i.e. the variational inequality (13) with

$$\varphi(t) = 2 \inf_{n\in\mathbb{N}} \left(\sum_{k=n+1}^{\infty} |x_k^\dagger| + K\,t \right) = K\,t.$$

This, however, yields by Proposition 6 the rate (11) and completes the proof of the proposition. □

Property 1 enables us to show convergence rates for ℓ^1-regularization for ill-posed and well-posed problems with sparse and non-sparse solutions. It has been shown in [17] that the rate function φ in (14) does in general not saturate. Even more, the rate is always obtainable either with an a priori or an a posteriori choice of the regularization parameter. This stands in sharp contrast to classical Tikhonov regularization, i.e., (3) with $p = 2$ and $||x||^2_{\ell^2}$ as penalty, which is known to admit convergence rates up to a maximum of $\delta^{2/3}$ for a suitable a priori choice of the regularization parameter and only a rate of $\delta^{1/2}$ under the discrepancy principle. Since the smoothness of the solution is typically unknown, this makes ℓ^1-regularization more attractive from the viewpoint of regularization theory than its ℓ^2 counterpart. One simply uses the discrepancy principle and no longer has to care about the smoothness of the solution. However, one may run into trouble when the solution does not belong to ℓ^1 but only to $\ell^2\backslash\ell^1$ such that $||x^\dagger||_{\ell^1} = \infty$. In such a case we call the regularization method (3) *oversmoothing*.

There are promising results showing that even in the situation of oversmoothing, ℓ^1-regularization may lead to convergence rates in a weaker norm. Again, an a priori choice or the discrepancy principle for the choice of α would lead to the optimal rates. Preliminary results have been shown in the preprint [15]. There, a strategy is shown to derive convergence rates in the ℓ^2-norm for ℓ^1-regularization for every $x^\dagger \in \ell^2$. The proof of a theorem analogously to Proposition 6 unfortunately was incomplete. It revolves around approximating $x^\dagger$ with $P_n x^\dagger$ and letting $n = n(\delta) \to \infty$ as $\delta \to 0$ with a specific choice of $n = n(\delta)$. The open problem was to show that the support of the approximate solutions is not larger than this $n(\delta)$. It appears that such a statement is possible by using item (c) in Property 1 and the necessary optimality condition for a minimizer of (3), where the latter provides us with the norm $||\eta||_{Y*}$ in Property 1 corresponding to a $\xi = A^*\eta \in \partial||x_\alpha^\delta||_{\ell^1}$. Since we are not able to use a variational source condition when $||x^\dagger||_{\ell^1} = \infty$ we need to use a different approach to show convergence rates. To this end we seem to have a chance to adapt the strategy of [17] based on elementary steps. Consequently, we hope to complete the detailed proof in an upcoming paper [16].

Acknowledgements The research of the authors was financially supported by Deutsche Forschungsgemeinschaft (DFG-grant HO 1454/10-1).

References

1. S.W. Anzengruber, B. Hofmann, R. Ramlau, On the interplay of basis smoothness and specific range conditions occurring in sparsity regularization. Inverse Prob. **29**, 125002, 21 (2013)
2. S.W. Anzengruber, B. Hofmann, P. Mathé, Regularization properties of the sequential discrepancy principle for Tikhonov regularization in Banach spaces. Appl. Anal. **93**, 1382–1400 (2014)
3. R.I. Boţ, B. Hofmann, The impact of a curious type of smoothness conditions on convergence rates in ℓ^1-regularization. Eurasian J. Math. Comput. Appl. **1**, 29–40 (2013)
4. K. Bredies, D.A. Lorenz, Regularization with non-convex separable constraints. Inverse Prob. **25**, 085011, 14 (2009)
5. M. Burger, J. Flemming, B. Hofmann, Convergence rates in ℓ^1-regularization if the sparsity assumption fails. Inverse Prob. **29**, 025013, 16 (2013)
6. D. Chen, B. Hofmann, J. Zou, Elastic-net regularization versus ℓ^1-regularization for linear inverse problems with quasi-sparse solutions. Inverse Prob. **33**(1), 015004, 17 (2017)
7. J. Cheng, M. Yamamoto, One new strategy for a priori choice of regularizing parameters in Tikhonov's regularization. Inverse Prob. **16**(4), L31–L38 (2000)
8. I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math. **57**(11), 1413–1457 (2004)
9. J. Flemming, Convergence rates for ℓ^1-regularization without injectivity-type assumptions. Inverse Prob. **32**(9), 095001, 19 (2016)
10. J. Flemming, Quadratic inverse problems and sparsity promoting regularization – two subjects, some links between them and an application in laser optics. Habilitation thesis, Technische Universität Chemnitz, Germany, 2017
11. J. Flemming, D. Gerth, Injectivity and weak*-to-weak continuity suffice for convergence rates in ℓ^1-regularization. J. Inv. Ill-Posed Prob. (2017, Published ahead of print). https://doi.org/10.1515/jiip-2017-0008
12. J. Flemming, M. Hegland, Convergence rates in ℓ^1-regularization when the basis is not smooth enough. Appl. Anal. **94**, 464–476 (2015)
13. J. Flemming, B. Hofmann, I. Veselić, On ℓ^1-regularization in light of Nashed's ill-posedness concept. Comput. Methods Appl. Math. **15**(3), 279–289 (2015)
14. J. Flemming, B. Hofmann, I. Veselić, A unified approach to convergence rates for ℓ^1-regularization and lacking sparsity. J. Inverse Ill-Posed Prob. **24**(2), 139–148 (2016)
15. D. Gerth, Tikhonov regularization with oversmoothing penalties. Preprint (2016). https://www.tu-chemnitz.de/mathematik/preprint/2016/PREPRINT_08.pdf
16. D. Gerth, Convergence rates for ℓ^1-regularization when the exact solution is not in ℓ^1. Article in preparation (2017)
17. D. Gerth, Convergence rates for ℓ^1-regularization without the help of a variational inequality. Electron. Trans. Numer. Anal. **46**, 233–244 (2017)
18. S. Goldberg, E. Thorp, On some open questions concerning strictly singular operators. Proc. Am. Math. Soc. **14**, 334–336 (1963)
19. M. Grasmair, Well-posedness and convergence rates for sparse regularization with sublinear l^q penalty term. Inverse Prob. Imaging **3**(3), 383–387 (2009)
20. M. Grasmair, Generalized Bregman distances and convergence rates for non-convex regularization methods. Inverse Prob. **26**, 115014, 16 (2010)
21. M. Grasmair, M. Haltmeier, O. Scherzer, Sparse regularization with l^q penalty term. Inverse Prob. **24**(5), 055020, 13 (2008)
22. M. Grasmair, M. Haltmeier, O. Scherzer, Necessary and sufficient conditions for linear convergence of ℓ^1-regularization. Comm. Pure Appl. Math. **64**, 161–182 (2011)
23. B. Hofmann, P. Mathé, Parameter choice in Banach space regularization under variational inequalities. Inverse Prob. **28**, 104006, 17 (2012)
24. B. Hofmann, B. Kaltenbacher, C. Pöschl, O. Scherzer, A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Prob. **23**, 987–1010 (2007)

25. D.A. Lorenz, Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. J. Inverse Ill-Posed Prob. **16**, 463–478 (2008)
26. R.E. Megginson, *An Introduction to Banach Space Theory*. Graduate Texts in Mathematics, vol. 183 (Springer, New York, 1998)
27. M.Z. Nashed, A new approach to classification and regularization of ill-posed operator equations, in *Inverse and Ill-Posed Problems (Sankt Wolfgang, 1986)*. Notes and Reports in Mathematics in Science and Engineering, vol. 4 (Academic, Boston, MA, 1987), pp. 53–75
28. R. Ramlau, Regularization properties of Tikhonov regularization with sparsity constraints. Electron. Trans. Numer. Anal. **30**, 54–74 (2008)
29. R. Ramlau, E. Resmerita, Convergence rates for regularization with sparsity constraints. Electron. Trans. Numer. Anal. **37**, 87–104 (2010)
30. W. Rudin, *Functional Analysis*. International Series in Pure and Applied Mathematics, 2nd edn. (McGraw-Hill, New York, 1991)
31. O. Scherzer (Ed.), *Handbook of Mathematical Methods in Imaging*, 2nd edn. (Springer, New York, 2011)
32. O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, F. Lenzen, *Variational Methods in Imaging*. Applied Mathematical Sciences, vol. 167 (Springer, New York, 2009)
33. T. Schuster, B. Kaltenbacher, B. Hofmann, K.S. Kazimierski, *Regularization Methods in Banach Spaces*. Radon Series on Computational and Applied Mathematics, vol. 10 (Walter de Gruyter, Berlin/Boston, 2012)

On Self-regularization of Ill-Posed Problems in Banach Spaces by Projection Methods

Uno Hämarik and Urve Kangro

Abstract We consider ill-posed linear operator equations with operators acting between Banach spaces. For the stable solution of ill-posed problems regularization is necessary, and for using computers discretization is necessary. In some cases discretization may also be used as regularization method with discretization parameter as regularization parameter, additional regularization is not needed. Regularization by discretization is called self-regularization. We consider self-regularization by projection methods, giving necessary and sufficient conditions for self-regularization by a priori choice of the dimension of subspaces as the regularization parameter. Convergence conditions are also given for the choice of the dimension by the discrepancy principle, without the requirement that the projection operators are uniformly bounded.

1 Introduction

Consider an ill-posed linear operator equation

$$Au = f, \quad f \in \mathscr{R}(A) \tag{1}$$

where $A \in L(E, F)$ is a linear injective mapping between nontrivial Banach spaces E and F. In practice only noisy data f^δ will be given. We assume here that the noise level δ satisfying

$$\|f^\delta - f\| \leq \delta \tag{2}$$

is known. For the stable solution of problem (1) it will be regularized to guarantee the convergence of regularized solutions to an exact solution u_* of (1) as δ goes to zero (see [9, 34]). Often ill-posed problems are formulated in infinite-dimensional

U. Hämarik · U. Kangro (✉)
University of Tartu, Tartu, Estonia
e-mail: uno.hamarik@ut.ee; urve.kangro@ut.ee

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_5

space, but for using computers the problem will be discretized, leading to some finite-dimensional (n-dimensional) problem. Typically discretization and regularization are used as separate procedures (see [14] for error estimates in regularized projection methods). However, if the data are exact, the successful discretization can lead to well posed problem with unique solution, which may converge to the solution of the original infinite-dimensional problem, if the dimensions of the discretized problems tend to infinity (see [20] for convergence conditions of projection methods). In this situation the self-regularization is possible: if data are noisy with known noise level δ, then by proper choice of $n = n(\delta)$ the solutions of discretized equations with noisy data converge to the solution of the original problem with exact data.

Self-regularization is probably the oldest regularization method. It is folklore of numerics that in numerical differentiation of a given noisy function by finite difference scheme, the discretization step h as the regularization parameter should be chosen in dependence of the noise level (see e.g. [9, 26]). From 1972 it is known (see [2]) that the quadrature formula method is a self-regularization method for the solution of the Volterra integral equation of the first kind; the rules for choice of the discretization step $h = h(\delta)$ as the regularization parameter in dependence of noise in the kernel and in the right-hand were given in [2] (see also [1]).

In this paper we consider projection methods. Let $E_n \subset E$, $Z_n \subset F^*$, $n \in \mathbb{N}$, be finite-dimensional nontrivial subspaces which have the role of approximating the spaces E and F^*, respectively. The general projection method defines a finite-dimensional approximation u_n to u_* by

$$u_n \in E_n \text{ and } \forall z_n \in Z_n \, : \, \langle z_n, Au_n \rangle_{F^*,F} = \langle z_n, f^\delta \rangle_{F^*,F}. \tag{3}$$

We also consider the least squares method (the “least residual” method would be a more natural name)

$$u_n \in \operatorname{argmin}\{ \|A\tilde{u}_n - f^\delta\|_F \, : \, \tilde{u}_n \in E_n\} \tag{4}$$

and the least error method

$$u_n \in \operatorname{argmin}\{ \|\tilde{u}\|_E \, : \, \forall z_n \in Z_n \, : \, \langle z_n, A\tilde{u} \rangle_{F^*,F} = \langle z_n, f^\delta \rangle_{F^*,F} \}. \tag{5}$$

The name “least error” method is justified by the fact that the obtained approximation u_n satisfies in case of exact data the inequalities

$$\|u_* - u_n\| \le \|u_* - v_n\|, \quad D(u_*, u_n) \le D(u_*, v_n) \quad \forall v_n \in E_n$$

in Hilbert and Banach spaces respectively (see [16, 32]), where $D(u_*, u_n)$ is the Bregman distance and $E_n \subset E$ is a certain subspace. It means that u_n is respectively the orthogonal projection or Bregman projection of u_* onto E_n. This method is called dual least-squares method in [3, 9, 21, 24, 26] and moment method in [21]. If E, F are Hilbert spaces, the least squares and least error methods are characterized by the equalities $Z_n = AE_n$ and $E_n = A^*Z_n$ respectively. If $E = F$ is a Hilbert space and $A = A^* \geq 0$, Galerkin method $E_n = F_n$ also can be used. Approximate solutions of the least squares and least error methods are found from a system of equations which is linear in Hilbert spaces and unfortunately nonlinear in Banach spaces.

In the collocation method, $Z_n = \text{span}\{\delta(t - t_i), i = 1, \ldots, n\}$ consists of linear combinations of Dirac's δ-functions $\delta(t-t_i)$ with support at collocation points t_i, $i = 1, \ldots, n$. Then (3) are the collocation conditions

$$u_n \in E_n, \qquad Au_n(t_i) = f^\delta(t_i), i = 1, \ldots, n \tag{6}$$

for finding u_n from arbitrary fixed subspace E_n.

Use of $Z_n = \text{span}\{\delta(t - t_i), i = 1, \ldots, n\}$ in the least error method (5) gives the least error collocation method, called also least-squares collocation [8, 9, 25] or moment collocation [21]. This method uses also collocation conditions (6), but the approximate set E_n is not arbitrary, it results from the condition that u_n is a minimum-norm solution of Eq. (6). If E is Hilbert space, then E_n is a subspace of E, but if E is a Banach space, then E_n is not necessarily a linear space.

Self-regularization by projection method was studied in [26], where the error estimates were given in Banach space formulation, convergence conditions were given for the collocation method, in Hilbert space formulation also for least squares and least error methods. The error estimates there (in Sobolev space formulation for least squares and Galerkin method also in [27]) allow to formulate a priori rules for the choice of dimension $n = n(\delta)$. For operator equations with noisy operator and noisy right-hand side the least squares, least error and Galerkin method were studied with a priori parameter choice in [12] and with a posteriori choice via discrepancy principle in [13]. Necessary and sufficient conditions for regularization by general projection methods in Hilbert spaces were given in [32], applications to mentioned methods and to class of integral equations of the first kind with Green type kernels were given. Convergence of the least error collocation method in case of exact data was proved for the space $E = L_2$ in [8, 21, 25], for Sobolev space $E = W_2^m$ in [33], for a priori choice $n = n(\delta)$ in [8], for a posteriori choice $n = n(\delta)$ by the monotone error rule in [15]. In the least error method in Hilbert spaces, a posteriori choice by the monotone error rule was studied in [10, 15], by the balancing principle in [3] (these both rules need weaker assumptions than the discrepancy principle). In Banach spaces the discrepancy principle was studied in [21] for a method close to the least squares method, in [16] for the general projection method and for the least squares method. Error estimates in Sobolev and Hölder-Zygmund norms of various discretization methods in certain boundary integral equations with a priori choice of $n = n(\delta)$ were given in [4]. Convergence of collocation method in case of exact data was analysed in [5–7], convergence by choice of $n = n(\delta)$ by discrepancy principle was proved in [16]. See also other works [11, 17, 18, 22–24] about self-regularization.

In this paper we consider in Sect. 2 the general projection method. The necessary and sufficient conditions for self-regularization by a priori choice $n = n(\delta)$ are given. Our approach is similar to [21], instead of a projector we use operator Q_n defined by (7). In previous treatments of the a posteriori choice of $n = n(\delta)$ by the discrepancy principle it was required that the projection operators are uniformly bounded. We modify the discrepancy principle so that this requirement is not needed. In Sect. 3 we consider the least squares method, in Sect. 4 the collocation method, where also numerical examples are given.

2 The General Projection Method

Let Q_n be the linear operator defined by

$$Q_n : F \to Z_n^* \quad \forall g \in F\,, z_n \in Z_n \;:\; \langle Q_n g, z_n\rangle_{Z_n^*, Z_n} = \langle z_n, g\rangle_{F^*, F} \tag{7}$$

which allows us to write (3) as

$$u_n \in E_n \quad \text{and} \quad Q_n A u_n = Q_n f^\delta. \tag{8}$$

The norm of Q_n equals one since

$$\|Q_n\| = \sup_{g\in F,\, \|g\|_F=1} \|Q_n g\|_{Z_n^*} = \sup_{g\in F,\, \|g\|_F=1, z_n\in Z_n,\, \|z_n\|_{F^*}=1} \langle Q_n g, z_n\rangle_{Z_n^*, Z_n} =$$

$$= \sup_{g\in F,\, \|g\|_F=1, z_n\in Z_n,\, \|z_n\|_{F^*}=1} \langle z_n, g\rangle_{F^*, F} = 1.$$

In the following lemma from [16] we give conditions under which the operator $A_n := Q_n A|_{E_n} : E_n \to Z_n^*$ has an inverse, the quantities

$$\kappa_n := \sup_{v_n\in E_n} \frac{\|v_n\|_E}{\|A v_n\|_F}, \quad \breve{\kappa}_n := \|A_n^{-1} Q_n\|, \quad \tilde{\kappa_n} := \|A_n^{-1}\| = \sup_{v_n\in E_n} \frac{\|v_n\|_E}{\|Q_n A v_n\|_F}, \tag{9}$$

$$\tau_n := \sup_{v_n\in E_n, v_n\neq 0} \frac{\|A v_n\|_F}{\|Q_n A v_n\|_{Z_n^*}}. \tag{10}$$

are finite and u_n from (3) is well-defined.

Lemma 1 *Let*

$$\dim(E_n) = \dim(Z_n) \tag{11}$$

and

$$\mathscr{N}(Q_n A) \cap E_n = \{0\} \tag{12}$$

hold. Then the operator A_n has an inverse and (3) is uniquely solvable for any $f^\delta \in F$. We have the inequalities

$$\kappa_n \le \breve{\kappa}_n \le \tilde{\kappa}_n \le \tau_n \kappa_n. \tag{13}$$

If

$$\exists \tau < \infty : \quad \tau_n \le \tau \quad \forall n \in \mathbb{N} \tag{14}$$

then also

$$\tilde{\kappa}_n \le \tau \kappa_n,$$

i.e. the quantities κ_n, $\breve{\kappa}_n$ and $\tilde{\kappa}_n$ are all equivalent as $n \to \infty$.

Remark 1 If $\mathscr{R}(A) \ne \overline{\mathscr{R}(A)}$ and the subspaces E_n satisfy the condition

$$\inf_{v_n \in E_n} \|v_n - v\| \to 0 \quad \forall v \in E \text{ as } n \to \infty, \tag{15}$$

then A^{-1} is unbounded and $\kappa_n \to \infty$ as $n \to \infty$.

2.1 Convergence with A Priori Choice of n

Theorem 1 *Let the operator A be injective. Let for $n \ge n_0$ the assumptions (11), (12) be satisfied. Then for $n \ge n_0$ the projection method (3) defines the unique approximation u_n, and the following error estimate holds:*

$$\begin{aligned} \|u_n - u_*\|_E &\le \min_{v_n \in E_n} [\,\|u_* - v_n\|_E + \|A_n^{-1} Q_n A(u_* - v_n)\|_E] + \breve{\kappa}_n \delta \qquad (16) \\ &\le (1 + \|A_n^{-1} Q_n A\|)\, dist(u_*, E_n) + \breve{\kappa}_n \delta. \end{aligned}$$

In case of exact data ($\delta = 0$) the convergence

$$\|u_n - u_*\|_E \to 0 \textit{ as } n \to \infty \tag{17}$$

holds if and only if there exists a sequence of approximations $(\hat{u}_n)_{n \in \mathbb{N}}$, $\hat{u}_n \in E_n$, satisfying the convergence conditions

$$\|u_* - \hat{u}_n\|_E \to 0 \textit{ as } n \to \infty \tag{18}$$

and

$$\|A_n^{-1} Q_n A(u_* - \hat{u}_n)\| \to 0 \textit{ as } n \to \infty. \tag{19}$$

If these conditions hold and the data are noisy, then choosing $n = n(\delta)$ according to a priori rule

$$n(\delta) \to \infty \text{ and } \breve{\kappa}_{n(\delta)}\delta \to 0 \text{ as } \delta \to 0 \tag{20}$$

we have convergence

$$\|u_{n(\delta)} - u_*\|_E \to 0 \text{ as } \delta \to 0. \tag{21}$$

Proof For any $v_n \in E_n$ we have, due to linearity of A,

$$\|u_n - u_*\|_E \le \|u_* - v_n\|_E + \|u_n - v_n\|_E = \|u_* - v_n\|_E + \|A_n^{-1}Q_n(f^\delta - Av_n)\|_E =$$

$$= \|u_* - v_n\|_E + \|A_n^{-1}Q_n[A(u_* - v_n) + f^\delta - f]\|_E \le$$

$$\le \|u_* - v_n\|_E + \|A_n^{-1}Q_nA(u_* - v_n)\|_E + \breve{\kappa}_n\delta,$$

hence the convergence estimate (16) holds.

If (18), (19) hold, then estimate (16) with $v_n = \hat{u}_n$ and our assumptions on the choice of $n(\delta)$ give convergence in both cases $\delta = 0$ and $\delta > 0$.

To show the necessity of (18), (19), note that if $\delta = 0$ and the convergence (17) holds, then $\hat{u}_n = u_n$ satisfies (18) and (19) (then $A_n^{-1}Q_nA(u_* - \hat{u}_n) = u_n - \hat{u}_n = 0$). □

According to the previous theorem in case $\delta = 0$ convergence (17) may hold due to sufficient smoothness of the solution. From this theorem we get in the following theorem conditions for convergence for every $f \in \mathscr{R}(A)$ (i.e. for every $u_* \in E$ without additional smoothness requirements).

Theorem 2 *Let the operator A be injective. Let for $n > n_0$ the assumptions (11), (12) be satisfied. Then in case of exact data ($\delta = 0$) the convergence (17) holds for every $f \in \mathscr{R}(A)$ if and only if the subspaces E_n satisfy condition (15) and the projectors $A_n^{-1}Q_nA : E \to E_n$ are uniformly bounded, i.e.,*

$$\|A_n^{-1}Q_nA\| \le M \tag{22}$$

for all $n \ge n_0$ and some constant M.

The last two conditions are necessary and sufficient for existence of relations $n = n(\delta)$ for convergence (21) for every $f \in \mathscr{R}(A)$ given approximately as arbitrary f^δ with $\|f^\delta - f\| \le \delta$.

Proof At first we show that conditions (15), (22) are sufficient for convergence of u_n. If condition (22) holds, then the error estimate (16) is of the form

$$\|u_n - u_*\|_E \le (1 + M) \min_{v_n \in E_n} \|u_* - v_n\|_E + \breve{\kappa}_n\delta, \tag{23}$$

this together with (15) guarantees convergence (17) for $\delta = 0$ and with parameter choice (20) also for $\delta \to 0$.

To show necessity of conditions (15), (22) for convergence of u_n, note that convergence (21) for every f^δ with $\|f^\delta - f\| \le \delta$ implies convergence (17) for f (i.e. convergence (17) for $\delta = 0$). Let $\delta = 0$ and $u_n \to u_*$ for all $u_* \in E$ as $n \to \infty$. Then (15) holds. But in case $u_n = A_n^{-1}Q_nAu_* \to u_*$ we have that $A_n^{-1}Q_nA \to I$ pointwise on E. By the uniform boundedness principle (Banach–Steinhaus theorem) this implies that $A_n^{-1}Q_nA$ must be uniformly bounded, which is condition (22). □

Remark 2 The boundedness property (22) holds, if uniformly bounded operators $\{P_n : E \to E_n, n \in N\}$ exist, satisfying

$$\breve{\kappa}_n \|A(I - P_n)\| \le M. \tag{24}$$

Namely condition (22) is equivalent to the condition

$$\|A_n^{-1}Q_nA(I - P_n)\| \le M', \tag{25}$$

while $A_n^{-1}Q_nA(I - P_n) = A_n^{-1}Q_nA - A_n^{-1}Q_nAP_n$ and the operator $A_n^{-1}Q_nAP_n = P_n$ is bounded. If (24) holds then using equality $\breve{\kappa}_n = \|A_n^{-1}Q_n\|$ we get (25).

For the convergence analysis in case of exact data we can choose different image spaces, particularly such that the equation becomes well-posed. But for noisy data the image space is determined by the data.

The following theorem (about the case of the exact data) shows, that convergence for one equation implies convergence also for certain other equations.

Theorem 3 *Let the operator A be injective. Let conditions (11), (12), (18) hold for $n \ge n_0$. Let the operator $A : E \to F$ have the form $A = S + K$, where $S : E \to W \subset F$ is invertible, W is a Banach space with continuous imbedding and $K : E \to W$ is compact. Let the operator $S_n := Q_nS|_{E_n} : E_n \to Z_n^*$ be invertible and $\|S_n^{-1}Q_nS\| \le M$ for some constant M. Then the projection equation $Q_nAu_n = Q_nf$ has for n large enough a unique solution $u_n \in E_n$, and $u_n \to u_*$ as $n \to \infty$.*

Proof From compactness of K follows the compactness of operator $S^{-1}K$. Denote $S_n = Q_nS|_{E_n}$. Since $S^{-1} : W \to E$ is bounded, the pointwise convergence $S_n^{-1}Q_nS \to I$ on W as $n \to \infty$ implies the pointwise convergence $S_n^{-1}Q_n \to S^{-1}$ as $n \to \infty$. From the pointwise convergence $S_n^{-1}Q_n \to S^{-1}$ and the compactness of K follows the norm convergence

$$\|(I + S_n^{-1}Q_nK) - (I + S^{-1}K)\| \to 0 \text{ as } n \to \infty.$$

Therefore the inverse operator $[I + S_n^{-1}Q_nK]^{-1} : E_n \to E_n$ exists and is uniformly bounded for large n. Due to equality $Q_nA = Q_nS[I + (Q_nS)^{-1}Q_nK]$ the operator Q_nA on E_n is invertible for large n with the inverse

$$(Q_nA)^{-1} = [I + (Q_nS)^{-1}Q_nK]^{-1}(Q_nS)^{-1}.$$

The equality

$$(Q_nA)^{-1}Q_nA = [I + (Q_nS)^{-1}Q_nK]^{-1}(Q_nS)^{-1}Q_nS(I + S^{-1}K)$$

allows to estimate

$$\|(Q_nA)^{-1}Q_nA\| \leq \|[I + (Q_nS)^{-1}Q_nK]^{-1}\|M\|I + S^{-1}K\| =: M_K.$$

This estimate may be rewritten in the form $\|A_n^{-1}Q_nA\| \leq M_K$, where the constant M_K depends on the operator K. Therefore condition (22) is satisfied and Theorem 2 guarantees convergence. □

For considering the influence of the noisy data, the behaviour of the quantities $\breve{\kappa}_n$ is essential. For estimating these quantities we introduce operators $\Pi_n : Z_n^* \to F$ such that the equality $Q_n\Pi_nQ_n = Q_n$ holds. Then the operator Π_nQ_n is a projector in F. Let $F_n = \mathscr{R}(\Pi_n)$. We assume that $F_n \subset W$ and let $W_n = F_n$, equipped with the norm of W. Let I_n be the identity operator, considered as acting from F_n to W_n.

Theorem 4 *Let conditions (11), (12), (22) hold for $n \geq n_0$. Let the operator $A : E \to W$ be invertible. Assume the projectors Π_nQ_n are uniformly bounded in F. Then*

$$\breve{\kappa}_n \leq C\|I_n\|_{F_n \to W_n}, \qquad n \geq n_0. \tag{26}$$

Proof We have

$$\breve{\kappa}_n = \|A_n^{-1}Q_n\|_{F\to E_n} = \|A_n^{-1}Q_nI_n\Pi_nQ_n\|_{F\to E_n} = \|A_n^{-1}Q_nAA^{-1}I_n\Pi_nQ_n\|_{F\to E_n} \leq$$

$$\leq \|A_n^{-1}Q_nA\|_{E\to E}\|A^{-1}\|_{W\to E}\|I_n\|_{F_n\to W_n}\|\Pi_nQ_n\|_{F\to F_n}.$$

This implies (26), since the other multipliers besides $\|I_n\|_{F_n\to W_n}$ are bounded. □

We point out that the choice of operators Π_n is quite arbitrary and is not determined by the method itself. For example, in collocation methods Π_nQ_n should be an interpolation projector, but it can be interpolation by splines, or polynomial interpolation or trigonometric interpolation or maybe something else, which may suit the particular problem. The only conditions are that the result is smooth enough (it must belong to the space W) and Π_nQ_n are uniformly bounded.

Estimates for $\|I_n\|_{F_n\to W_n}$ can be found using inverse properties of approximation subspaces (estimating elements of F_n via their norm in W_n). Splines are often useful here, because their inverse properties (estimates of the derivatives in terms of the splines themselves) are well known. Estimates for operators $\|A(I - P_n)\|$ in condition (24) can be derived from the approximation properties of the approximation subspaces. Often the norm $\|(I - P_n^*)A^*\| = \|A(I - P_n)\|$ is easier to estimate.

2.2 *Convergence with A Posteriori Choice of n: The Discrepancy Principle*

For the discrepancy principle, in previous works the assumption (14) about uniform boundedness of τ_n was required. For collocation methods this is the uniform boundedness of the interpolation projector onto the subspace $AE_n \subset F$. If $F = C^m$, (14) does not hold in general. In the next two theorems we consider two versions of the discrepancy principle, condition (14) is assumed only in the first version.

Theorem 5 *Let the assumptions of Lemma 1 be satisfied for $n \geq n_0$, and let u_n be defined by the projection method (3). Let the convergence*

$$\breve{\kappa}_{n+1} dist(f, AE_n) \to 0 \; as \; n \to \infty \tag{27}$$

holds. We also assume that there exists a sequence of approximations $(\hat{u}_n)_{n\in\mathbb{N}}$, $\hat{u}_n \in E_n$, satisfying (18) and (19). Let condition (14) holds. Let $b > \tau + 1$ be fixed and for $\delta > 0$, let $n = n_{DP}(\delta)$ be the first index such that

$$\|Au_n - f^\delta\|_F \leq b\delta. \tag{28}$$

Then $n_{DP}(\delta)$ is finite and

$$\|u_{n_{DP}(\delta)} - u_*\|_E \to 0 \; as \; \delta \to 0. \tag{29}$$

Proof For any n let $v_n \in E_n$ be such that $\|f^\delta - Av_n\| = \text{dist}(f^\delta, AE_n)$. We have

$$\|Au_n - f^\delta\|_F \leq \|A(u_n - v_n)\|_F + \|Av_n - f^\delta\|_F \leq$$

$$\leq \tau_n \|Q_n A(u_n - v_n)\| + \|Av_n - f^\delta\|_F = \tau_n \|Q_n(f^\delta - Av_n)\| + \|Av_n - f^\delta\|_F \leq$$

$$\leq (\tau_n + 1)\,\text{dist}(f^\delta, AE_n) \leq (\tau_n + 1)\,(\delta + \text{dist}(f, AE_n)). \tag{30}$$

This inequality together with (14) and relation

$$\text{dist}(f, AE_n) \leq \|A(u_* - \hat{u}_n)\| \to 0 \quad \text{as } n \to \infty \tag{31}$$

imply that n_{DP} is finite.

If for some $\delta_k \to 0\,(k \to \infty)$ the discrepancy principle gives $n_{DP}(\delta_k) \leq N$ with $N \geq 0$, then the sequence $u_{n_{DP}(\delta_k)}$ lies in a finite-dimensional subspace — the linear hull of E_n, $n = 1, \ldots, N$. Since

$$\|Au_{n_{DP}(\delta_k)} - f^{\delta_k}\|_F \leq b\delta_k, \tag{32}$$

then $Au_{n_{DP}(\delta_k)} \to f$ as $k \to \infty$. This implies convergence $u_{n_{DP}(\delta_k)} \to u_*$ as $k \to \infty$, since the operator A has the bounded inverse on finite-dimensional subspaces.

Consider now the general case $n_{DP}(\delta) \to \infty$ as $\delta \to 0$. Let $m = n_{DP}(\delta) - 1 \geq 0$. For $n = m$ the inequality (28) does not hold, and (30) with (14) gives

$$b\delta < \|Au_m - f^\delta\|_F \leq (\tau + 1)(\delta + \operatorname{dist}(f, AE_m)), \tag{33}$$

therefore also

$$\frac{(b - 1 - \tau)\delta}{\tau + 1} < \operatorname{dist}(f, AE_m). \tag{34}$$

The convergence (27) implies

$$\breve{\kappa}_{n_{DP}}\delta < \frac{\tau + 1}{b - 1 - \tau}\breve{\kappa}_{n_{DP}}\operatorname{dist}(f, AE_{n_{DP}-1}) \to 0 \text{ as } n_{DP} \to \infty. \tag{35}$$

Therefore the second term in estimate (16) converges as $n = n_{DP} \to \infty$. Convergence of the first term there follows as in the proof of Theorem 1, using $v_n = \hat{u}_n$, $n = n_{DP}(\delta)$ and assumptions (18), (19). □

Theorem 6 *Let the assumptions of Theorem 5 be satisfied without requirement (14). Let the sequence*

$$b_n > (1 + \tau_n)(1 + \varepsilon) \tag{36}$$

be fixed with some fixed $\varepsilon > 0$ and $n = n_{DP}(\delta)$ be chosen as the first index such that

$$\|Au_n - f^\delta\|_F \leq b_n\delta. \tag{37}$$

Then $n_{DP}(\delta)$ is finite and the convergence (29) holds.

Proof The proof is similar to the proof of the previous theorem. Condition (36) gives the inequality $(\tau_n + 1)\delta \leq b_n\delta - \varepsilon(\tau_n + 1)\delta$, and the estimate (30) can be continued as follows:

$$\|Au_n - f^\delta\|_F \leq b_n\delta + (\tau_n + 1)(\operatorname{dist}(f, AE_n) - \varepsilon\delta).$$

Due to convergence (31) the second summand here will be negative for sufficiently large n, therefore $n_{DP}(\delta)$ will be finite. If for some $\delta_k \to 0$ $(k \to \infty)$ the discrepancy principle gives $n_{DP}(\delta_k) \leq N$, the proof of convergence $u_{n_{DP}(\delta_k)} \to u_*$ as $k \to \infty$ is the same as in the previous theorem with the exception, that the inequality $\|Au_{n_{DP}(\delta_k)} - f^{\delta_k}\|_F \leq b_N\delta_k$ is used instead of (32). The proof of convergence (29) in case $n_{DP}(\delta) \to \infty$ as $\delta \to 0$ is the same as in the previous theorem, only in the inequalities (33)–(35) the quantities b, τ and $\dfrac{\tau + 1}{b - 1 - \tau}$ are replaced by b_m, τ_m and $\varepsilon^{-1} < \dfrac{\tau_m + 1}{b_m - 1 - \tau_m}$, respectively. □

3 The Least Squares Method

In the least squares method (4) we use the condition

$$\mathcal{N}(A) \cap E_n = \{0\} \tag{38}$$

instead of the requirement of the injectivity of the operator A. In [16] the following result is proved.

Theorem 7 *Let condition (38) be satisfied for all $n \in \mathbb{N}$. Then an approximation u_n according to the least squares method (4) exists and the error estimate*

$$\|u_n - u_*\| \leq \inf_{v_n \in E_n} \{ \|u_* - v_n\|_E + 2\kappa_n \|Au_* - Av_n\|_F \} + 2\kappa_n \delta$$

holds. If there exists a sequence of approximations $(\hat{u}_n)_{n\in\mathbb{N}}$, $\hat{u}_n \in E_n$, satisfying (18) and

$$\kappa_n \|A(u_* - \hat{u}_n)\|_F \to 0 \text{ as } n \to \infty, \tag{39}$$

then we have in case of exact data convergence $\|u_n - u_*\|_E \to 0$ as $n \to \infty$, and in case of noisy data with the choice of $n = n(\delta)$ according to

$$n(\delta) \to \infty \text{ and } \kappa_{n(\delta)}\delta \to 0 \text{ as } \delta \to 0$$

convergence

$$\|u_{n(\delta)} - u_*\|_E \to 0 \text{ as } \delta \to 0\,. \tag{40}$$

If in addition to convergences (18), (39) also $\kappa_{n+1} \|A(u_* - \hat{u}_n)\|_F \to 0$ as $n \to \infty$ holds, then convergence (40) holds also with the choice of $n(\delta)$ by the discrepancy principle: for fixed $b > 1$ choose $n(\delta)$ as the first index such that $\|Au_n - f^\delta\| < b\delta$.

The discrepancy principle fits better to the least squares method than to other projection methods in the sense that there is no need to calculate or estimate the quantities τ, τ_n which may be a hard task.

4 Application: Collocation Method for Volterra Integral Equations

We consider collocation method for Volterra integral equations. In the first two examples these equations are cordial integral equations studied in [19, 28–31]. We give properties of these equations in Sect. 4.1 and consider the collocation method in Sect. 4.2.

4.1 Cordial Integral Equations

Consider cordial integral equations of the first kind

$$\int_0^t \frac{1}{t}a(t,s)\phi(\frac{s}{t})u(s)ds = f(t), \quad 0 \le t \le T, \tag{41}$$

where $\phi \in L^1(0,1)$ is called the core of the cordial integral operator, and a,f are given smooth enough functions. Define the cordial integral operators

$$(V_\phi u)(t) = \int_0^t \frac{1}{t}\phi(\frac{s}{t})u(s)ds, \quad (V_{\phi,a}u)(t) = \int_0^t \frac{1}{t}a(t,s)\phi(\frac{s}{t})u(s)ds.$$

Denote $\Delta_T = \{(s,t) : t \in [0,T],\ s \in [0,t]\}$. The following results are proven in [28–31].

Theorem 8 *Let $\phi \in L^1(0,1)$, $a \in C^m(\Delta_T)$. Then $V_{\phi,a} \in \mathscr{L}(C^m[0,T])$ and*

$$\|V_{\phi,a}\|_{C^m[0,T]} \le C\|\phi\|_{L^1(0,1)}\|a\|_{C^m(\Delta_T)}.$$

Theorem 9 *Let $\phi \in L^1(0,1)$ and let $\lambda \in \mathbf{C}$ with $\operatorname{Re}\lambda > 0$. Then t^λ is an eigenfunction of V_ϕ in $C[0,T]$, and the corresponding eigenvalue is $\hat{\phi}(\lambda) = \int_0^1 \phi(x)x^\lambda dx$. If $\operatorname{Re}\lambda > m$, then the eigenfunction belongs to $C^m[0,T]$.*

Theorem 10 *Let $\phi \in L^1(0,1)$, $a \in C^m(\Delta_T)$. Then the spectrum of $V_{\phi,a}$ in $C^m[0,T]$ is given by $\sigma_m(V_{\phi,a}) = \{0\} \cup \{a(0,0)\hat{\phi}(k),\ k = 0,\dots,m\} \cup \{a(0,0)\hat{\phi}(\lambda),\ \operatorname{Re}\lambda > m\}$.*

Theorem 11 *Let $\phi \in L^1(0,1)$, $x(1-x)\phi'(x) \in L^1(0,1)$, $\int_0^1 \phi(x)dx > 0$ and there exists $\beta < 1$ such that $(x^\beta\phi(x))' \ge 0$ for $x \in (0,1)$. Assume also that $a \in C^{m+1}(\Delta_T)$ and $a(t,t) \ne 0$. Then $V_{\phi,a}$ is injective in $C[0,T]$, $C^{m+1}[0,T] \subset V_{\phi,a}(C^m[0,T]) \subset C^m[0,T]$, and $V_{\phi,a}^{-1} \in \mathscr{L}(C^{m+1}[0,T], C^m[0,T])$.*

Corollary 1 *Let the assumptions of Theorem 11 be satisfied and let $f \in C^{m+1}[0,T]$ be given. Then Eq. (41) is uniquely solvable in $C[0,T]$ and its solution is in $C^m[0,T]$.*

4.2 Polynomial Collocation Method for Cordial Integral Equations, Numerical Results

According to Theorem 9, functions $t^k, k \in \mathbb{N}$ are eigenfunctions of the cordial integral operator V_ϕ, therefore the polynomial collocation method is well adapted for these equations. We look for solutions in the form $u_n(s) = \sum_{j=0}^n c_j s^j$. In the collocation method we choose the collocation points $t_k \in [0,T]$, $k = 0,\dots,n$ and find c_k, $k = 0, \dots, n$ from the collocation equations

$$\sum_{j=0}^{n} c_j \int_0^{t_k} \frac{1}{t_k} a(t_k, s)\phi(\frac{s}{t_k}) s^j ds = f(t_k), \quad k = 0, \dots, n.$$

To set up the system, one has to calculate exactly or "well enough" the integrals

$$\int_0^t \frac{1}{t_k} a(t_k, s)\phi(\frac{s}{t_k}) s^j ds.$$

For theoretical results it is convenient to use the basis $\{s^j\}$ for polynomials; for practical calculations though, this results in very badly conditioned systems. So for larger N one has to use a better basis, for example the (scaled) Chebyshev polynomials $T_p(t) = \cos\big(p \arccos\big(\frac{2t}{T} - 1\big)\big)$. In fact, it may be simpler to make first the change of variables $t = \frac{T}{2}(1 - \cos y)$ and then work with trigonometric polynomials in y instead.

In the following Examples 1, 2, $E = F = C[0, T]$, E_n is the space of polynomials of order up to n and Z_n is the linear span of δ-functions with supports t_k, $k = 0, \dots, n$. Let $a(t, s) \equiv 1$. Then $V_\phi : E_n \to E_n$ and τ_n is simply the norm of the interpolation projector from C to C with the interpolation nodes t_k, $k = 0, \dots, n$. If t_k are the Chebyshev nodes, then $\tau_n \approx \frac{2}{\pi} \ln(n + 1) + 1$.

In Examples 1, 2 certain noise levels were chosen and the noise was generated by random numbers with uniform distribution at the collocation nodes, and on nine times denser mesh for calculating the discrepancy. We also found the optimal number n_{opt} and the corresponding error $e_{opt} = \min_{n \in \mathbb{N}} \|u_n - u_*\|_E = \|u_{n_{opt}} - u_*\|_E$. The discrepancy principle was used for finding proper $n = n(\delta)$. The condition (14) is not satisfied in Examples 1, 2. According to the discrepancy principle from Theorem 6 we found the first $n = n_{DP}$ satisfying the inequality $\|Au_n - f^\delta\|_F \le b_n \delta$ with $b_n = 1.001(1 + \tau_n)$. We denote the corresponding error by $e_{DP} = \|u_{n_{DP}} - u_*\|$. The optimal errors and the errors obtained by using the discrepancy principle are presented in the following Tables 1 and 2. In these tables also $b_{n_{DP}}$ are presented.

Example 1 Consider the cordial integral equation (here $\phi(x) = \frac{1}{\sqrt{x}}$)

$$\int_0^t \frac{u(s)ds}{\sqrt{st}} = \frac{1}{t^2 + 1}, \quad t \in [0, T] \tag{42}$$

with exact solution $u(s) = \frac{1 - 3s^2}{2(s^2+1)^2}$. For this equation κ_n can be estimated using Markoff's inequality, by Cn^2. Since the right-hand side of the equation is analytic, $\mathrm{dist}(f, AE_n)$ converges to zero exponentially, hence the assumptions of Theorem 6 are satisfied.

We took $T{=}10$ and used noisy data with noise levels $\delta{=}10^{-4}, 10^{-6}, \dots, 10^{-14}$. The number of collocation nodes was $10, 15, 20, \dots, 110$.

Table 1 Optimal errors with the corresponding n_{opt} and errors obtained by using the discrepancy principle with $b_{n_{DP}}$ for Eq. (42)

δ	e_{opt}	n_{opt}	e_{DP}	n_{DP}	$b_{n_{DP}}$
10^{-4}	$6 \cdot 10^{-2}$	25	$8 \cdot 10^{-2}$	20	3.94
10^{-6}	$1.01 \cdot 10^{-3}$	40	$2.4 \cdot 10^{-3}$	30	4.19
10^{-8}	$1.51 \cdot 10^{-5}$	40	$1.51 \cdot 10^{-5}$	40	4.36
10^{-10}	$1.8 \cdot 10^{-7}$	50	$1.8 \cdot 10^{-7}$	50	4.56
10^{-12}	$4.69 \cdot 10^{-9}$	75	$9.58 \cdot 10^{-9}$	60	4.62
10^{-14}	$7.04 \cdot 10^{-11}$	105	$7.57 \cdot 10^{-11}$	70	4.71

Table 2 Optimal errors with the corresponding n_{opt} and errors obtained by using the discrepancy principle with $b_{n_{DP}}$ for Eq. (43)

δ	e_{opt}	n_{opt}	e_{DP}	n_{DP}	$b_{n_{DP}}$
10^{-3}	$1.5 \cdot 10^{-1}$	10	$1.5 \cdot 10^{-1}$	10	3.53
10^{-4}	$5 \cdot 10^{-2}$	40	$1.1 \cdot 10^{-1}$	30	3.94
10^{-5}	$5.24 \cdot 10^{-3}$	20	$2 \cdot 10^{-2}$	50	4.5
10^{-6}	$6.13 \cdot 10^{-4}$	40	$5.16 \cdot 10^{-3}$	100	4.94
10^{-7}	$9.17 \cdot 10^{-5}$	90	$5.77 \cdot 10^{-3}$	230	5.46

Example 2 Consider the equation

$$\int_0^t \frac{u(s)ds}{\sqrt{st}} = t^{3/2}(2-t)^{5/2}, \quad t \in [0,2]. \tag{43}$$

The exact solution is $u(s) = 2s^{3/2}(2-s)^{5/2} - \frac{5}{2}s^{5/2}(2-s)^{3/2}$. Since the integral operator is the same as in Example 1, κ_n is the same. The distance $\mathrm{dist}(f, AE_n)$ can be estimated by Cn^{-3}, hence the assumptions of Theorem 6 are satisfied.

We used noisy data with noise levels $\delta = 10^{-3}, 10^{-4}, \ldots, 10^{-7}$. The number of collocation nodes was 10, 20, 30, ..., 300.

4.3 Spline-Collocation for Volterra Integral Equation, Numerical Results

We consider a Volterra integral equation of the first kind

$$(Au)(t) := \int_0^t K(t,s)u(s)\,ds = f(t), \ t \in [0,1] \tag{44}$$

with the operator $A \in L(L^p(0,1), C[0,1]),\ 1 \le p \le \infty$. The approximation space is $E_n = S_{k-1}^{(-1)}(I_\Delta)$, the space of discontinuous piecewise polynomials of order $k-1$ with mesh Δ. In the collocation method we find u_n from the spline space E_n such that

$$Au_n(t_{i,j}) = f^\delta(t_{i,j}), \quad i = 1, \ldots, n, \ j = 1, \ldots, k$$

Table 3 Optimal errors and errors obtained by using the discrepancy principle for Eq. (45); left with $q = 3/2$ and right with $q = 5/2$

δ	e_{opt}	n_{opt}	e_{DP}	n_{DP}	e_{opt}	n_{opt}	e_{DP}	n_{DP}
10^{-1}	$2.5 \cdot 10^{-1}$	1	$2.5 \cdot 10^{-1}$	1	$2.9 \cdot 10^{-1}$	1	$2.9 \cdot 10^{-1}$	1
10^{-2}	$6.8 \cdot 10^{-2}$	2	$6.8 \cdot 10^{-2}$	2	$5.4 \cdot 10^{-2}$	2	$5.4 \cdot 10^{-2}$	2
10^{-3}	$1.3 \cdot 10^{-2}$	8	$1.8 \cdot 10^{-2}$	5	$9 \cdot 10^{-3}$	6	$1.1 \cdot 10^{-2}$	5
10^{-4}	$3.2 \cdot 10^{-3}$	24	$3.3 \cdot 10^{-3}$	20	$1.7 \cdot 10^{-3}$	15	$3 \cdot 10^{-3}$	8
10^{-5}	$7.6 \cdot 10^{-4}$	72	$8.4 \cdot 10^{-4}$	86	$3.5 \cdot 10^{-4}$	32	$6.2 \cdot 10^{-4}$	18
10^{-6}	$1.9 \cdot 10^{-4}$	128	$3.3 \cdot 10^{-4}$	512	$6.8 \cdot 10^{-5}$	72	$9.9 \cdot 10^{-5}$	46
10^{-7}	$4.5 \cdot 10^{-5}$	512	$1.2 \cdot 10^{-4}$	2048	$1.5 \cdot 10^{-5}$	128	$1.5 \cdot 10^{-5}$	128

where $t_{i,j} = (i - 1 + c_j)h \in [0, 1]$, $i = 1, \dots, n$, $j = 1, \dots, k$ are collocation nodes and $0 < c_1 < \dots < c_k \leq 1$ are collocation parameters whose choice is essential.

Example 3 Consider the equation

$$Au(t) = \int_0^t u(s)ds = \frac{t^q}{q}, \quad t \in [0, 1], \quad q \in \{3/2, 5/2\} \tag{45}$$

with operator $A : L^1(0, 1) \to C[0, 1]$. The exact solution is $u(s) = s^{q-1}$. We used for E_n the space of discontinuous linear splines with uniform mesh ih, $i = 0, \dots, n$, where $h = 1/n$. The collocation points are $t_{i1} = (i - 1 + c)h$, $t_{i2} = ih$, $c \in (0, 1)$. For this problem $\check{\kappa}_n$ can be estimated using Theorem 4. Here we can take for F_n the space of continuous linear splines and the inverse property of these splines gives

$$\forall w_n \in F_n, \quad \|w_n'\| \leq Cn\|w_n\|,$$

hence $\check{\kappa}_n \leq Cn$. The distance $\operatorname{dist}(f, AE_n)$ can be estimated by Cn^{-q}.

It can be shown that here

$$\tau = \begin{cases} 1 + \frac{c^2}{2(1-c)}, & \text{if } c \geq \frac{1}{2}, \\ 1 + \frac{(1-c)^2}{2c} & \text{if } c \leq \frac{1}{2}. \end{cases}$$

The quantity τ is minimal for $c = \frac{1}{2}$, then $\tau = 1.25$. In this example $\tau_n = \tau$ holds, i.e. τ_n does not depend on n. We used $c = \frac{1}{2}$ and for satisfying the condition $b > \tau + 1$ in Theorem 5 we actually took $b = 1.01 + \tau = 2.26$ for the discrepancy principle.

The noisy data were generated by the formula $f^\delta(t_{i,j}) = f(t_{i,j}) + \delta\theta_{i,j}$, where $\delta = 10^{-m}, m \in \{2, \dots, 7\}$ and $\theta_{i,j}$ are random numbers with normal distribution, normed after being generated: $\max_{i,j} |\theta_{i,j}| = 1$ (Table 3).

We can conclude that for these model problems the discrepancy principle gave reasonable results.

Acknowledgements The authors are supported by institutional research funding IUT20-57 of the Estonian Ministry of Education and Research.

References

1. A. Apartsyn, *Nonclassical Linear Volterra Equations of the First Kind*. Inverse and Ill-Posed Problems Series (De Gruyter, Berlin, 2003)
2. A.S. Apartsin, A.B. Bakushinskii, Approximate solution of Volterra integral equations of the first kind by the method of quadratic sums (Russian), in *Differential and Integral Equations*, vol. 1 (Irkutsk Gos. Univ., Irkutsk, 1972), pp. 248–258
3. G. Bruckner, S. Pereverzev, Self-regularization of projection methods with a posteriori discretization level choice for severely ill-posed problems. Inverse Probl. **19**, 147–156 (2003)
4. G. Bruckner, S. Prössdorf, G. Vainikko, Error bounds of discretization methods for boundary integral equations with noisy data. Appl. Anal. **63**, 25–37 (1996)
5. H. Brunner, *Collocation Methods for Volterra Integral and Related Functional Equations* (Cambridge University Press, Cambridge, 2004)
6. P. Eggermont, Collocation for Volterra integral equations of the first kind with iterated kernel. SIAM J. Numer. Anal. **20**, 1032–1048 (1983)
7. P. Eggermont, Stability and robustness of collocation methods for Abel-type integral equations. Numer. Math. **45**, 431–445 (1983)
8. H.W. Engl, On least-squares collocation for solving linear integral equations of the first kind with noisy right-hand side. Boll. Geodesia Sci. Aff. **41**, 291–313 (1982)
9. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*. Mathematics and Its Applications, vol. 375 (Kluwer, Dordrecht, 1996)
10. A. Ganina, U. Hämarik, U. Kangro, On the self-regularization of ill-posed problems by the least error projection method. Math. Model. Anal. **19**, 299–308 (2014)
11. C.W. Groetsch, A. Neubauer, Convergence of a general projection method for an operator equation of the first kind. Houston J. Math. **14**, 201–208 (1988)
12. U. Hämarik, Projection methods for the regularization of linear ill-posed problems. Proc. Comput. Center Tartu Univ. **50**, 69–90 (1983)
13. U. Hämarik, Discrepancy principle for choice of dimension in solution of ill-posed problems by projection methods (Russian). Acta Comment. Univ. Tartuensis **672**, 27–34 (1984)
14. U. Hämarik, On the discretization error in regularized projection methods with parameter choice by discrepancy principle, in *Ill-Posed Problems in Natural Sciences*, ed. by A.N. Tikhonov, A.S. Leonov, A.I. Prilepko, I.A. Vasin, V.A. Vatutin, A.G. Yagola (VSP, Utrecht, Moscow, 1992), pp. 24–28
15. U. Hämarik, E. Avi, A. Ganina, On the solution of ill-posed problems by projection methods with a posteriori choice of the discretization level. Math. Model. Anal. **7**(2), 241–252 (2002)
16. U. Hämarik, B. Kaltenbacher, U. Kangro, E. Resmerita, Regularization by discretization in Banach spaces. Inverse Probl. **32** (3), 035004 (2016)
17. B. Hofmann, P. Mathe, S.V. Pereverzev, Regularization by projection: approximation theoretic aspects and distance functions. J. Inv. Ill-Posed Problems **15**, 527–545 (2007)
18. B. Kaltenbacher, Regularization by projection with a posteriori discretization level choice for linear and nonlinear ill-posed problems. Inverse Probl. **16**(5), 1523–1539 (2000)
19. U. Kangro, Cordial Volterra integral equations and singular fractional integro-differential equations in spaces of analytic functions. Math. Model. Anal. **22**(4), 548–567 (2017)
20. S. Kindermann, Projection methods for ill-posed problems revisited. Comput. Methods Appl. Math. **16**(2), 257–276 (2016)
21. R. Kress, *Linear Integral Equations* (Springer, Berlin, 2014)
22. P.K. Lamm, A survey of regularization methods for first-kind Volterra equations, in *Surveys on Solution Methods for Inverse Problems* (Springer, Vienna, 2000), pp. 53–82

23. P. Mathe, S.V. Pereverzev, Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. SIAM J. Numer. Anal. **38**(6), 1999–2021 (2001)
24. P. Mathe, N. Schöne, Regularization by projection in variable Hilbert scales. Appl. Anal. **87**, 201–219 (2008)
25. M.Z. Nashed, G. Wahba, Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind. Math. Comp. **28**, 69–80 (1974)
26. F. Natterer, Regularisierung schlecht gestellter Probleme durch Projektionsverfahren. Numer. Math. **28**, 329–341 (1977)
27. R.G. Richter, Numerical solution of integral equations of the first kind with nonsmooth kernels. SIAM J. Numer. Anal. **15**, 511–522 (1978)
28. G. Vainikko, Cordial Volterra integral equations 1. Numer. Funct. Anal. Optim. **30**, 1145–1172 (2009)
29. G. Vainikko, Cordial Volterra integral equations 2. Numer. Funct. Anal. Optim. **31**, 191–219 (2010)
30. G. Vainikko, First kind cordial Volterra integral equations 1. Numer. Funct. Anal. Optim. **33**(6), 680–704 (2012)
31. G. Vainikko, First kind cordial Volterra integral equations 2. Numer. Funct. Anal. Optim. **35**, 1607–1637 (2014)
32. G. Vainikko, U. Hämarik, Projection methods and self-regularization in ill-posed problems. Izvestiya Vysshikh Uchebnykh Zavedenii Matematika **10**, 3–17 (1985) (in Russian); Soviet Math. **29**, 1–20 (1985)
33. G. Vainikko, U. Hämarik, Self-regularization solving ill-posed problems by projection methods, in *Models and Methods in Operational Research*, ed. by B.A. Beltyukov, V.P. Bulatov (Nauka, Novosibirsk, 1988), pp. 157–164
34. G.M. Vainikko, A.Yu. Veretennikov, *Iteration Procedures in Ill- Posed Problems* (Nauka, Moscow, 1986)

Monotonicity-Based Regularization for Phantom Experiment Data in Electrical Impedance Tomography

Bastian Harrach and Mach Nguyet Minh

Abstract In electrical impedance tomography, algorithms based on minimizing the linearized-data-fit residuum have been widely used due to their real-time implementation and satisfactory reconstructed images. However, the resulting images usually tend to contain ringing artifacts. In this work, we shall minimize the linearized-data-fit functional with respect to a linear constraint defined by the monotonicity relation in the framework of real electrode setting. Numerical results of standard phantom experiment data confirm that this new algorithm improves the quality of the reconstructed images as well as reduce the ringing artifacts.

1 Introduction

Electrical Impedance Tomography (EIT) is a recently developed non-invasive imaging technique, where the inner structure of a reference object can be recovered from the current and voltage measurements on the object's surface. It is fast, inexpensive, portable and requires no ionizing radiation. For these reasons, EIT qualifies for continuous real time visualization right at the bedside.

In clinical EIT applications, the reconstructed images are usually obtained by minimizing the linearized-data-fit residuum [3, 7]. These algorithms are fast and simple. However, to the best of the authors' knowledge, there is no rigorous global convergence results that have been proved so far. Moreover, the reconstructed images usually tend to contain ringing artifacts.

B. Harrach (✉)
Department of Mathematics, Goethe University Frankfurt, Frankfurt, Germany
e-mail: harrach@math.uni-frankfurt.de

M. N. Minh
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
e-mail: minh.mach@helsinki.fi

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_6

Recently, Seo and one of the author have shown in [19] that a single linearized step can give the correct shape of the conductivity contrast. This result raises a question that whether to regularize the linearized-data-fit functional such that the corresponding minimizer yields a good approximation of the conductivity contrast. An affirmative answer has been proved in [18] for the continuum boundary data. In the present paper, we shall apply this new algorithm to the real electrode setting and test with standard phantom experiment data. Numerical results later on show that this new algorithm helps to improve the quality of the reconstructed images as well as reduce the ringing artifacts. It is worth to mention that our new algorithm is non-iterative, hence, it does not depend on an initial guess and does not require expensive computation. Other non-iterative algorithms, for example, the Factorization Method [14, 16] and the Monotonicity-based Method [4, 20, 30, 31], on the other hand, are much more sensitive to measurement errors than our new algorithm when phantom data or real data are applied [5, 10, 22, 33].

The paper is organized as follows. In Sect. 2 we introduce the mathematical setting, describe how the measured data can be collected and set up a link between the mathematical setting and the measured data. Section 3 presents our new algorithm and the numerical results were shown in Sect. 4. We conclude this paper with a brief discussion in Sect. 5.

2 Mathematical Setting

Let $\Omega \subseteq \mathbb{R}^n, n \geq 2$ describe the imaging subject and $\sigma : \Omega \to \mathbb{R}$ be the unknown conductivity distribution inside Ω. We assume that Ω is a bounded domain with smooth boundary $\partial\Omega$ and that the function σ is real-valued, strictly positive and bounded. Electrical Impedance Tomography (EIT) aims at recovering σ using voltage and current measurements on the boundary of Ω. There are several ways to inject currents and measure voltages. We shall follow the *Neighboring Method* (aka Adjacent Method) which was suggested by Brown and Segar in [6] and is still widely being used by practitioners. In this method, electrodes are attached on the object's surface, and an electrical current is applied through a pair of adjacent electrodes whilst the voltage is measured on all other pairs of adjacent electrodes excluding those pairs containing at least one electrode with injected current. Figure 1 illustrates the first and second current patterns for a 16-electrode EIT system. At the first current pattern (Fig. 1a), small currents of intensity $I_1^{(1)}$ and $I_2^{(1)} = -I_1^{(1)}$ are applied through electrodes E_1 and E_2 respectively, and the voltage differences $U_3^{(1)}, U_4^{(1)}, \ldots, U_{15}^{(1)}$ are measured successively on electrode pairs $(E_3, E_4), (E_4, E_5), \ldots, (E_{15}, E_{16})$. In general, for a L-electrode EIT system, at the k-th current pattern, by injecting currents $I_k^{(k)}$ and $I_{k+1}^{(k)} = -I_k^{(k)}$ to electrodes E_k and E_{k+1} respectively, one gets $L-3$ voltage measurements $\{U_l^{(k)}\}$, where $l \in \{1, 2, \ldots, L\}$ and $|k-l| > 1$. Note that here and throughout the paper, the

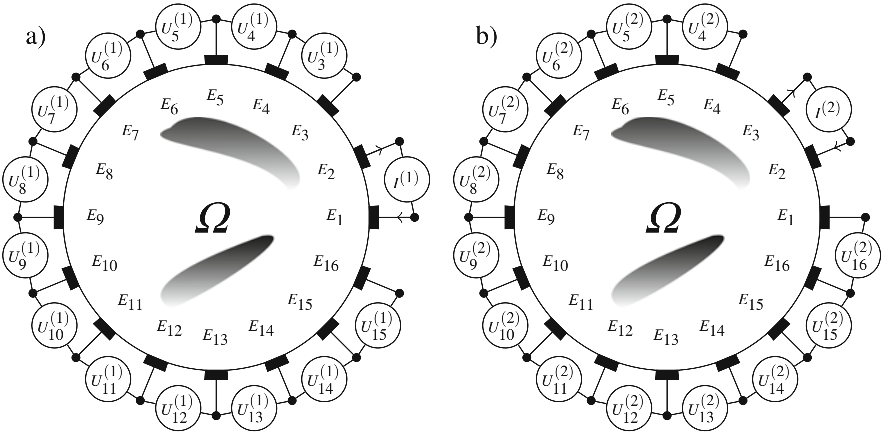

a)
b)
Ω
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10} E_{11} E_{12} E_{13} E_{14} E_{15} E_{16}
$I^{(1)}$
$U_3^{(1)}$ $U_4^{(1)}$ $U_5^{(1)}$ $U_6^{(1)}$ $U_7^{(1)}$ $U_8^{(1)}$ $U_9^{(1)}$ $U_{10}^{(1)}$ $U_{11}^{(1)}$ $U_{12}^{(1)}$ $U_{13}^{(1)}$ $U_{14}^{(1)}$ $U_{15}^{(1)}$
$I^{(2)}$
$U_4^{(2)}$ $U_5^{(2)}$ $U_6^{(2)}$ $U_7^{(2)}$ $U_8^{(2)}$ $U_9^{(2)}$ $U_{10}^{(2)}$ $U_{11}^{(2)}$ $U_{12}^{(2)}$ $U_{13}^{(2)}$ $U_{14}^{(2)}$ $U_{15}^{(2)}$ $U_{16}^{(2)}$

Fig. 1 The Neighboring Method: (**a**) first current pattern, (**b**) second current pattern

electrode index is always considered modulo L, i.e. the index $L + 1$ also refers to the first electrode, etc.

Assuming that the electrodes E_l are relatively open and connected subsets of $\partial\Omega$, that they are perfectly conducting and that contact impedances are negligible, the resulting electric potential $u^{(k)}$ at the k-th current pattern obeys the following mathematical model (the so-called *shunt model* [8]):

$$\begin{array}{lll} \nabla \cdot (\sigma \nabla u) & = 0 & \text{in } \Omega, \\ \int_{E_l} \sigma \partial_\nu u \,\mathrm{d}s & = I_l^{(k)} & \text{for } l = 1, \ldots, L, \\ \sigma \partial_\nu u & = 0 & \text{on } \partial\Omega \setminus \bigcup_{l=1}^{L} E_l, \\ u|_{E_l} & = \text{const.} & \text{for } l = 1, \ldots, L. \end{array} \tag{1}$$

Here ν is the unit normal vector on $\partial\Omega$ pointing outward and $I_l^{(k)} := (\delta_{k,l} - \delta_{k+1,l})I$ describes the k-th applied current pattern where a current of strength $I > 0$ is driven through the k-th and $(k + 1)$-th electrode. Notice that $\{I_l^{(k)}\}$ satisfy the conservation of charge $\sum_{l=1}^{L} I_l^{(k)} = 0$, and that the electric potential $u^{(k)}$ is uniquely determined by (1) only up to the addition of a constant. The voltage measurements are given by

$$U_l^{(k)} := u^{(k)}|_{E_l} - u^{(k)}|_{E_{l+1}}. \tag{2}$$

The herein used shunt model ignores the effect of contact impedances between the electrodes and the imaging domain. This is only valid when voltages are measured on small (see [15]) and current-free electrodes, so that (1) correctly models only the measurements $U_l^{(k)}$ with $|k - l| > 1$. For difference measurements, the missing elements $U_l^{(k)}$ with $|k - l| \leq 1$, on the other hand, can be calculated by interpolation taking into account reciprocity, conservation of voltages and the geometry-specific smoothness of difference EIT data, cf. [17]. For an imaging subject with unknown conductivity σ, one thus obtains a full matrix of measurements $U(\sigma) = (U_l^{(k)})_{k,l=1,\dots,L}$.

3 Monotonicity-Based Regularization

3.1 Standard One-Step Linearization Methods

In difference EIT, the measurements $U(\sigma)$ are compared with measurements $U(\sigma_0)$ for some reference conductivity distribution σ_0 in order to reconstruct the conductivity difference $\sigma - \sigma_0$. This is usually done by a single linearization step

$$U'(\sigma_0)(\sigma - \sigma_0) \approx U(\sigma) - U(\sigma_0).$$

where $U'(\sigma_0) : L^\infty(\Omega) \to \mathbb{R}^{L\times L}$ is the Fréchet derivative of the voltage measurements

$$U'(\sigma_0) : \kappa \mapsto \left(- \int_\Omega \kappa \nabla u_{\sigma_0}^{(k)} \cdot \nabla u_{\sigma_0}^{(l)} \,\mathrm{d}x \right)_{1\le k,l \le L}$$

We discretize the reference domain $\overline{\Omega} = \cup_{j=1}^P \overline{P}_j$ into P disjoint open pixels P_j and make the piecewise-constant Ansatz

$$\kappa(x) = \sum_{j=1}^P \kappa_j \chi_{P_j}(x).$$

This approach leads to the linear equation

$$\mathbf{S}\boldsymbol{\kappa} = \mathbf{V} \tag{3}$$

where $\mathbf{V}$ and the columns of the *sensitivity matrix* $\mathbf{S}$ contain the entries of the measurements $U(\sigma) - U(\sigma_0)$ and the discretized Fréchet derivative, resp., written as long vectors, i.e.,

$$\begin{aligned}
\boldsymbol{\kappa} &= (\kappa_j)_{j=1}^P \in \mathbb{R}^P, \\
\mathbf{V} &= (V_i)_{i=1}^{L^2} \in \mathbb{R}^{L^2}, \quad \text{with } V_{(l-1)L+k} = U_l^{(k)}(\sigma) - U_l^{(k)}(\sigma_0), \\
\mathbf{S} &= (S_{i,j}) \in \mathbb{R}^{L^2,P}, \quad \text{with } S_{(l-1)L+k,j} = - \int_{P_j} \nabla u_{\sigma_0}^{(k)} \cdot \nabla u_{\sigma_0}^{(l)} \,\mathrm{d}x.
\end{aligned}$$

Most practically used EIT algorithms are based on solving a regularized variant of (3) to obtain an approximation κ to the conductivity difference $\sigma - \sigma_0$. The popular algorithms NOSER [7] and GREIT [3] use (generalized) Tikhonov regularization and minimize

$$\|\mathbf{S}\boldsymbol{\kappa} - \mathbf{V}\|_{\text{res}}^2 + \alpha \,\|\boldsymbol{\kappa}\|_{\text{pen}}^2 \to \text{min!}$$

with (heuristically chosen) weighted Euclidian norms $\|\cdot\|_{\text{res}}$ and $\|\cdot\|_{\text{pen}}$ in the residuum and penalty term.

3.2 Monotonicity-Based Regularization

It has been shown in [19] that shape information in EIT is invariant under linearization. Thus one-step linearization methods are principally capable of reconstructing the correct (outer) support of the conductivity difference even though they ignore the non-linearity of the EIT measurement process. In [18] the authors developed a monotonicity-based regularization method for the linearized EIT equation for which (in the continuum model) it can be guaranteed that the regularized solutions converge against a function that shows the correct outer shape. In this section, we formulate and analyze this new method for real electrode measurements, and in the next section we will apply it to real data from a phantom experiment and compare it with the GREIT method.

The main idea of monotonicity-based regularization is to minimize the residual of the linearized equation (3)

$$\|\mathbf{S}\boldsymbol{\kappa} - \mathbf{V}\|^2 \to \text{min!}$$

with constraints on the entries of $\boldsymbol{\kappa}$ that are obtained from monotonicity tests.

For the following, we assume that the background is homogeneous and that all anomalies are more conductive, or all anomalies are less conductive than the background, i.e., σ_0 is constant, and either

$$\sigma(x) = \sigma_0 + \gamma(x)\chi_D(x), \qquad \text{or} \qquad \sigma(x) = \sigma_0 - \gamma(x)\chi_D(x).$$

D is an open set denoting the conductivity anomalies, and $\gamma : D \to \mathbb{R}$ is the contrast of the anomalies. We furthermore assume that we are given a lower bound $c > 0$ of the anomaly contrast, i.e. $\gamma(x) \geq c$.

For the monotonicity tests it is crucial to consider the measurements and the columns of the sensitivity matrix $\mathbf{S}$ as matrices and compare them in terms of matrix definiteness, cf. [9, 17, 21] for the origins of this sensitivity matrix based approach. Let $V := U(\sigma) - U(\sigma_0) \in \mathbb{R}^{L\times L}$ denote the EIT difference measurements written as $L \times L$-matrix, and $S_k \in \mathbb{R}^{L\times L}$ denote the k-th column of the sensitivity matrix written as $L \times L$-matrix, i.e. the (j,l)-th entry of S_k is given by

$$-\int_{P_k} \nabla u_{\sigma_0}^{(j)} \cdot \nabla u_{\sigma_0}^{(l)} \,\mathrm{d}x.$$

We then define for each pixel P_k

$$\beta_k := \max\{\alpha \geq 0 : \alpha S_k \geq -|V|\}, \tag{4}$$

where $|V|$ denotes the matrix absolute value of V, and the comparison $\alpha S_k \geq -|V|$ is to be understood in the sense of matrix definiteness, i.e. $\alpha S_k \geq -|V|$ holds if and only if all eigenvalues of $\alpha S_k + |V|$ are non-negative.

Following [18] we then solve the linearized EIT equation (3) using the monotonicity constraints β_k. We minimize the Euclidean norm of the residuum

$$\|\mathbf{S}\boldsymbol{\kappa} - \mathbf{V}\|^2 \to \min! \tag{5}$$

under the constraints that

(C1) in the case $\sigma \geq \sigma_0$: $0 \leq \kappa_k \leq \min(a_+, \beta_k)$, and
(C2) in the case $\sigma \leq \sigma_0$: $0 \geq \kappa_k \geq -\min(a_-, \beta_k)$,

where $a_+ := \sigma_0 - \frac{\sigma_0^2}{\sigma_0 + c}$, and $a_- := c$.

For noisy data V^δ with $\|V^\delta - V\| \leq \delta$ this approach can be regularized by replacing β_k with

$$\beta_k^\delta := \max\{\alpha \geq 0 : \alpha S_k \geq -|V^\delta| - \delta I\}, \tag{6}$$

where $I \in \mathbb{R}^{L \times L}$ is the identity matrix. For the implementation of β_k^δ see Sect. 4.

For the continuum model, and under the assumption that D has connected complement, the authors [18] showed that for exact data this monotonicity-constrained minimization of the linearized EIT residuum admits a unique solution and that the support of the solution agrees with the anomalies support D up to the pixel partition. Moreover, [18] also shows that for noisy data and using the regularized constraints β_k^δ, minimizers exist and that, for $\delta \to 0$, they converge to the minimizer with the correct support. Since practical electrode measurements can be regarded as an approximation to the continuum model, we therefore expect that the above approach will also well approximate the anomaly support for real electrode data.

In the continuum model, the constraints β_k will be zero outside the support of the anomaly and positive for each pixel inside the anomaly. The first property relies on the existence of localized potentials [11] and is only true in the limit of infinitely many, infinitely small electrodes. The latter property is however true for any number of electrodes as the following result shows:

Theorem 1 *If $P_k \subseteq D$, then*

(a) in the case $\sigma \geq \sigma_0$ the constraint β_k fulfills $\beta_k \geq a_+ > 0$, and
(b) in the case $\sigma \leq \sigma_0$ the constraint β_k fulfills $\beta_k \geq a_- > 0$.

Proof If $P_k \subseteq D$ and $\sigma \geq \sigma_0$ then

$$\frac{\sigma_0}{\sigma}(\sigma - \sigma_0) = \sigma_0 - \frac{\sigma_0^2}{\sigma} \geq \left(\sigma_0 - \frac{\sigma_0^2}{\sigma_0 + c}\right)\chi_{P_k} = a_+ \chi_{P_k},$$

and if $P_k \subseteq D$ and $\sigma \leq \sigma_0$ then

$$\sigma_0 - \sigma \geq c\chi_{P_k} = a_- \chi_{P_k}.$$

Hence, it suffices to show that $\alpha S_k \geq -|V|$ holds for all $\alpha > 0$ that fulfill

(a) $\alpha \chi_{P_k} \leq \frac{\sigma_0}{\sigma}(\sigma - \sigma_0)$, or
(b) $\alpha \chi_{P_k} \leq \sigma_0 - \sigma$.

We use the following monotonicity relation from [23, Lemma 3.1] (see also [24, 25] for the origin of this estimate): For any vector $g = (g_j)_{j=1}^L \in \mathbb{R}^L$ we have that

$$\int_\Omega \frac{\sigma_0}{\sigma}(\sigma_0 - \sigma) \left|\nabla u_{\sigma_0}^{(g)}\right|^2 \mathrm{d}x \geq g^\top V g \geq \int_\Omega (\sigma_0 - \sigma) \left|\nabla u_{\sigma_0}^{(g)}\right|^2 \mathrm{d}x, \qquad (7)$$

with $u_{\sigma_0}^{(g)} = \sum_{j=1}^L g_j \nabla u_{\sigma_0}^{(j)}$.

If $\alpha \chi_{P_k} \leq \frac{\sigma_0}{\sigma}(\sigma - \sigma_0)$, then

$$0 \geq g^\top(\alpha S_k)g = -\int_{P_k} \alpha \left|\nabla u_{\sigma_0}^{(g)}\right|^2 \geq \int_\Omega \frac{\sigma_0}{\sigma}(\sigma_0 - \sigma) \left|\nabla u_{\sigma_0}^{(g)}\right|^2 \geq g^\top V g,$$

which shows that $|V| = -V \geq -\alpha S_k$.

If $\alpha \chi_{P_k} \leq \sigma_0 - \sigma$, then

$$0 \leq g^\top(-\alpha S_k)g = \int_{P_k} \alpha \left|\nabla u_{\sigma_0}^{(g)}\right|^2 \leq \int_\Omega (\sigma_0 - \sigma) \left|\nabla u_{\sigma_0}^{(g)}\right|^2 \leq g^\top V g,$$

which shows that $|V| = V \geq -\alpha S_k$. □

4 Numerical Results

In this section, we will test our algorithm on the data set `iirc_data_2006` measured by Professor Eung Je Woo's EIT research group in Korea [26, 27, 29, 32]. `iirc` stands for Impedance Imaging Research Center. The data set is publicly available as part of the open source software framework EIDORS [2] (Electrical Impedance and Diffused Optical Reconstruction Software). Since the data set `iirc_data_2006` is also frequently used in the EIDORS tutorials, we believe that this is a good benchmark example to test our new algorithm.

4.1 Experiment Setting

The data set `iirc_data_2006` was collected using the 16-electrode EIT system KHU Mark1 (see [28] for more information of this system). The reference object was a Plexiglas tank filled with saline. The tank was a cylinder of diameter 0.2 m with 0.01 m diameter round electrodes attached on its boundary. Saline was filled to about 0.06 m depth. Inside the tank, one put a Plexiglas rod of diameter 0.02 m. The conductivity of the saline was 0.15 S/m and the Plexiglas rod was basically non-conductive. Data acquisition protocol was adjacent stimulation, adjacent measurement with data acquired on all electrodes.

The data set `iirc_data_2006` contains the voltage measurements for both homogeneous and non-homogeneous cases. Measurements for the homogeneous case were obtained when the Plexiglas rod was taken away (reference conductivity in this case is 0.15 S/m). In the non-homogeneous case, 100 different voltage measurements were measured corresponding to 100 different positions of the Plexiglas rod.

4.2 Numerical Implementation

EIDORS [2] (Electrical Impedance and Diffused Optical Reconstruction Software) is an open source software that is widely used to reconstruct images in electrical impedance tomography and diffuse optical tomography. To reconstruct images with EIDORS, one first needs to build an EIDORS model that fits with the measured data. In this paper, we shall use the same EIDORS model described in the EIDORS tutorial web-page:

`http://eidors3d.sourceforge.net/tutorial/EIDORS_basics/tutorial110.shtml`

Figure 2 shows the reconstructed images of the 9th-inhomogeneous measurements with different regularization parameters using the EIDORS built-in command `inv_solve`, which follows the algorithm proposed in [1]. We emphasize that, Fig. 2b (regularization parameter is chosen as 0.03 by default) was considered at the EIDORS tutorial web-page, we show them here again in order to easily compare them with the reconstructed images using our new method later on.

4.3 Minimizing the Residuum

In the EIDORS model suggested in the EIDORS tutorial web-page, the reference body was chosen by default as a disk of diameter 1 m and the default reference conductivity was 1 S/m. However, in the experiment setting, the reference body was a cylinder of diameter 0.2 m and the reference conductivity was 0.15 S/m. Hence, an appropriate scaling factor should be applied to the measurements, to make sure

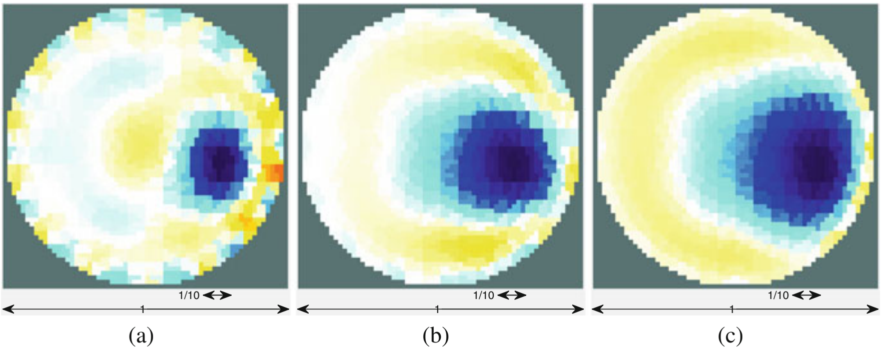


Fig. 2 Reconstructed images for the 9th-inhomogeneous voltage measurements with different regularization parameters. (**a**) Parameter = 0.003. (**b**) Parameter = 0.03. (**c**) Parameter = 0.3

that the EIDORS model fits with these measurements. In the EIDORS tutorial web-page, the measurements were scaled by multiplying by a factor 10^{-4}. In this paper, to increase the precision of the model, we shall find the best scaling factor that minimizes the error between the measured data and the data generated by the EIDORS model. More precisely, let call `vh` the measured data for homogeneous case and `vh_model` the homogeneous data generated by the EIDORS model, the best scaling factor is a minimizer of the following problem

$$\min_{c \in \mathbb{R}} \|c * \texttt{vh} - \texttt{vh\_model}\|_2$$

For this experiment setting, the best factor is $2.49577 * 10^{-5}$. From now on, by measured data we always refer to scaled measured data with respect to this best factor.

The next step is to recover the missing measurements on the driving electrodes. We shall follow the result in [17] to obtain an approximation for these missing measurements using interpolation.

Now we are in a position to minimize the problem (5) under the linear constraint (C1) or (C2). To do this, we need to clarify a_+, a_-, and β_k in the linear constraints. After scaling, the reference conductivity is $\sigma_0 = 1\,\text{S/m}$, and D still denotes the Plexiglas rod with conductivity $\sigma = 0\,\text{S/m}$. Thus, $\gamma = 1$, $a_- = \inf_D \gamma = 1$ and β_k is calculated using (4). In practice, there is no way to obtain the exact value of the matrix V in (4). Indeed, what we know is just the measured data $V^\delta = U^\delta(\sigma) - U^\delta(\sigma_0)$, where δ denotes the noise level. When replacing $|V|$ by the noisy version $|V^\delta|$, it may happen that there is no $\alpha > 0$ so that the matrix $|V^\delta| + \alpha S_k$ is still positive semi-definite. Therefore, instead of using (4), we shall calculate β_k from

$$\beta_k = \max\{\alpha \geq 0 \ : \ |V^\delta| + \alpha S_k \geq -\delta I\}.$$

Here, I represents the identity matrix, and δ is chosen as the absolute value of the smallest eigenvalue of V^δ. Notice that, in the presence of noise, $|V^\delta| + \delta I$ plays the role of the positive semi-definite matrix $|V|$. We shall follow the argument in [18] to calculate β_k. Let L be the lower triangular Cholesky decomposition matrix of $|V^\delta| + \delta I$, and let $\lambda_s(L^{-1}S_k(L^*)^{-1})$ be the smallest eigenvalue of the matrix $L^{-1}S_k(L^*)^{-1}$. Since S_k is negative semi-definite, so is $L^{-1}S_k(L^*)^{-1}$. Thus, $\lambda_s(L^{-1}S_k(L^*)^{-1}) \leq 0$. Arguing in the same manner as in [18], we get

$$\beta_k = -\frac{1}{\lambda_s(L^{-1}S_k(L^*)^{-1})} \geq 0.$$

The minimizer of (5) is then obtained using two different approaches: one employs `cvx` (Fig. 3a), a package for specifying and solving convex programs [12, 13], the other (Fig. 3b) uses the `MATLAB` built-in function `quadprog`

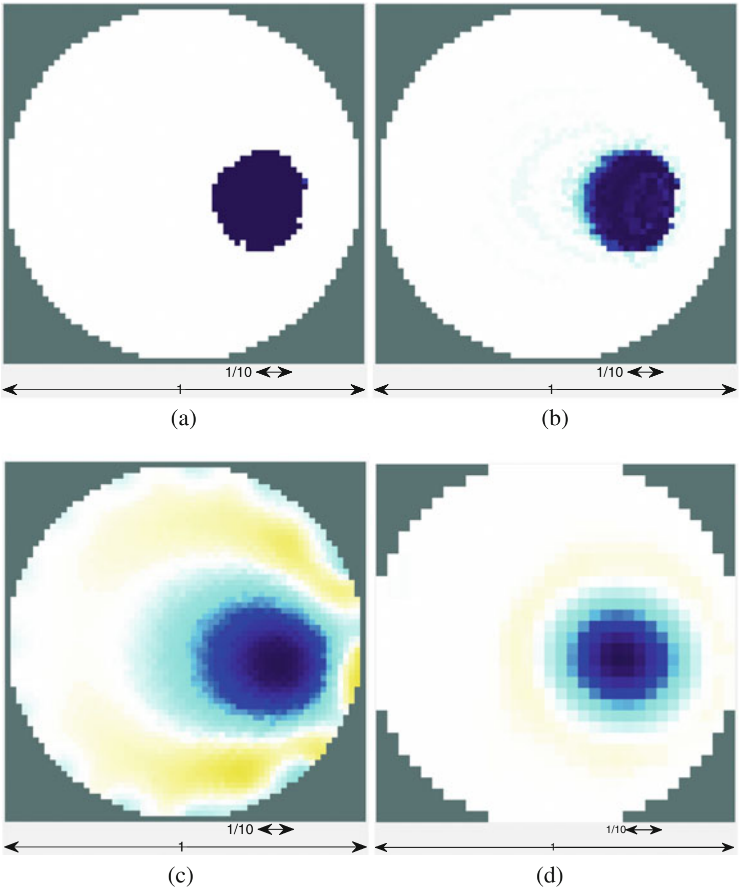


Fig. 3 Reconstructed images for the 9th-inhomogeneous voltage measurements with different algorithms (after scaling the measured data w.r.t the best scaling factor). (**a**) `cvx`. (**b**) `quadprog`. (**c**) EIDORS `inv_solve`. (**d**) GREIT

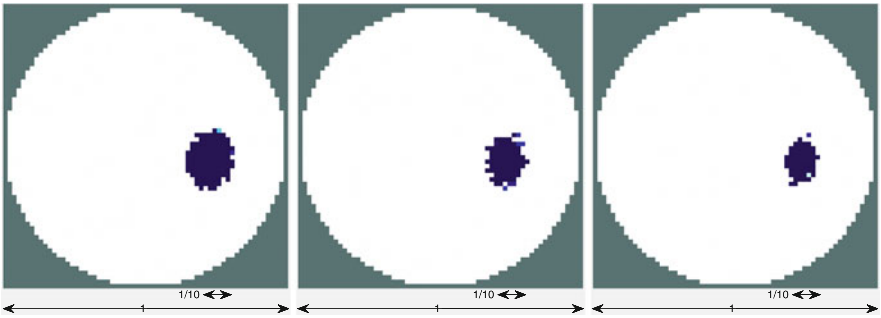
(trust-region-reflective Algorithm). We also show the reconstructed result using the built-in function inv_solve of EIDORS [1] (Fig. 3c) with the default regularization parameter 0.03 and with GREIT algorithm [3] (Fig. 3d) to see that scaling the measured data with the best scaling factor will improve a little bit the reconstructed image. Notice that reconstructed images are highly affected by the choice of the minimization algorithms (see Table 1 for their runtime), and we will see from Fig. 3 that the images obtained by cvx has less artifacts than the others.

It is worth to emphasize that although each EIDORS model is assigned to a default regularization parameter, when using the EIDORS built-in function inv_solve [1], in order to obtain a good reconstruction (Fig. 2) one has to manually choose a regularization parameter, whilst the regularization parameters a_- and β_k in our method are known a-priori provided the information of the conductivity σ and the reference conductivity σ_0 exists. Besides, if we manually choose the parameters $\min(a_-, \beta_k)$, we even get much better reconstructed images (Fig. 4).

Last but not least, our new method proves its advantage when there are more than one inclusions (Fig. 5).

Table 1 Runtime of pictures in Fig. 3

Algorithm	Runtime (s)
cvx	839.3892
quadprog (trust-region-reflective)	5.4467
EIDORS (inv_solve)	0.0231
GREIT	0.0120



(a) (b) (c)

Fig. 4 Reconstructed images for the 9th-inhomogeneous voltage measurements with monotonicity-based algorithm and different choices of lower constraint. (**a**) $\min(2, \beta_k)$. (**b**) $\min(3, \beta_k)$. (**c**) $\min(4, \beta_k)$

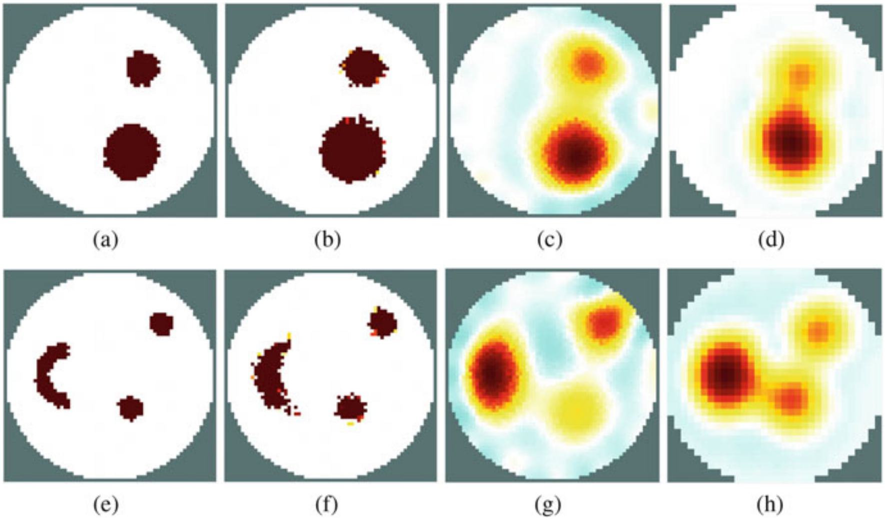


Fig. 5 Reconstructed images for simulated data with 0.1% noise. (From left to right) First column: True conductivity change, Second column: our new method (with cvx), Third column: EIDORS (inv_solve), Last column: GREIT

5 Conclusions

In this paper, we have presented a new algorithm to reconstruct images in EIT in the real electrode setting. Numerical results show that this new algorithm helps to reduce the ringing artifacts in the reconstructed images. Global convergence result of this algorithm has been proved in [18] for the Continuum Model. In future works, we shall aim to prove global convergence result for the Shunt Model setting as well as reduce the runtime to fit with real-time applications.

Acknowledgements The authors thank Professor Eung Je Woo's EIT research group for the iirc phantom data set. MNM thanks the Academy of Finland (Finnish Center of Excellence in Inverse Problems Research) for financial support of the project number 273979. During part of the preparation of this work, MNM worked at the Department of Mathematics of the Goethe University Frankfurt, Germany.

References

1. A. Adler, R. Guardo, Electrical impedance tomography: regularized imaging and contrast detection. IEEE Trans. Med. Imaging **15**(2), 170–179 (1996)
2. A. Adler, W.R. Lionheart, Uses and abuses of EIDORS: an extensible software base for EIT. Physiol. Meas. **27**(5), S25 (2006)
3. A. Adler, J.H. Arnold, R. Bayford, A. Borsic, B. Brown, P. Dixon, T.J. Faes, I. Frerichs, H. Gagnon, Y. Gärber, et al., GREIT: a unified approach to 2D linear EIT reconstruction of lung images. Physiol. Meas. **30**(6), S35 (2009)

4. R.G. Aykroyd, M. Soleimani, W.R. Lionheart, Conditional Bayes reconstruction for ERT data using resistance monotonicity information. Meas. Sci. Technol. **17**(9), 2405 (2006)
5. M. Azzouz, M. Hanke, C. Oesterlein, K. Schilcher, The factorization method for electrical impedance tomography data from a new planar device. Int. J. Biomed. Imaging **2007**, 83016 (2007)
6. B. Brown, A. Seagar, The Sheffield data collection system. Clin. Phys. Physiol. Meas. **8**(4A), 91 (1987)
7. M. Cheney, D. Isaacson, J. Newell, S. Simske, J. Goble, NOSER: an algorithm for solving the inverse conductivity problem. Int. J. Imaging Syst. Technol. **2**(2), 66–75 (1990)
8. K.S. Cheng, D. Isaacson, J. Newell, D.G. Gisser, Electrode models for electric current computed tomography. IEEE Trans. Biomed. Eng. **36**(9), 918–924 (1989)
9. M.K. Choi, B. Harrach, J.K. Seo, Regularizing a linearized EIT reconstruction method using a sensitivity-based factorization method. Inverse Probl. Sci. Eng. **22**(7), 1029–1044 (2014)
10. H. Garde, S. Staboulis, Convergence and regularization for monotonicity-based shape reconstruction in electrical impedance tomography. arXiv preprint arXiv:1512.01718 (2015)
11. B. Gebauer, Localized potentials in electrical impedance tomography. Inverse Probl. Imaging **2**(2), 251–269 (2008)
12. M. Grant, S. Boyd, Graph implementations for nonsmooth convex programs, in *Recent Advances in Learning and Control*, ed. by V. Blondel, S. Boyd, H. Kimura. Lecture Notes in Control and Information Sciences (Springer, Berlin, 2008), pp. 95–110
13. M. Grant, S. Boyd, CVX: matlab software for disciplined convex programming, version 2.1 (2014), http://cvxr.com/cvx
14. M. Hanke, A. Kirsch, Sampling methods, in *Handbook of Mathematical Models in Imaging*, ed. by O. Scherzer (Springer, Berlin, 2011), pp. 501–550
15. M. Hanke, B. Harrach, N. Hyvönen, Justification of point electrode models in electrical impedance tomography. Math. Models Methods Appl. Sci. **21**(06), 1395–1413 (2011)
16. B. Harrach, Recent progress on the factorization method for electrical impedance tomography. Comput. Math. Methods Med. **2013**, 425184 (2013)
17. B. Harrach, Interpolation of missing electrode data in electrical impedance tomography. Inverse Probl. **31**(11), 115008 (2015)
18. B. Harrach, M.N. Minh, Enhancing residual-based techniques with shape reconstruction features in electrical impedance tomography. Inverse Probl. **32**(12), 125002 (2016)
19. B. Harrach, J.K. Seo, Exact shape-reconstruction by one-step linearization in electrical impedance tomography. SIAM J. Math. Anal. **42**(4), 1505–1518 (2010)
20. B. Harrach, M. Ullrich, Monotonicity-based shape reconstruction in electrical impedance tomography. SIAM J. Math. Anal. **45**(6), 3382–3403 (2013)
21. B. Harrach, M. Ullrich, Resolution guarantees in electrical impedance tomography. IEEE Trans. Med. Imaging **34**(7), 1513–1521 (2015)
22. B. Harrach, J.K. Seo, E.J. Woo, Factorization method and its physical justification in frequency-difference electrical impedance tomography. IEEE Trans. Med. Imaging **29**(11), 1918–1926 (2010)
23. B. Harrach, E. Lee, M. Ullrich, Combining frequency-difference and ultrasound modulated electrical impedance tomography. Inverse Probl. **31**(9), 095003 (2015)
24. M. Ikehata, Size estimation of inclusion. J. Inverse Ill-Posed Probl. **6**(2), 127–140 (1998)
25. H. Kang, J.K. Seo, D. Sheen, The inverse conductivity problem with one measurement: stability and estimation of size. SIAM J. Math. Anal. **28**(6), 1389–1405 (1997)
26. T.I. Oh, K.H. Lee, S.M. Kim, H. Koo, E.J. Woo, D. Holder, Calibration methods for a multi-channel multi-frequency EIT system. Physiol. Meas. **28**(10), 1175 (2007)
27. T.I. Oh, E.J. Woo, D. Holder, Multi-frequency EIT system with radially symmetric architecture: KHU Mark1. Physiol. Meas. **28**(7), S183 (2007)
28. T.I. Oh, E.J. Woo, D. Holder, Multi-frequency EIT system with radially symmetric architecture: KHU Mark1. Physiol. Meas. **28**, S183–S196 (2007)
29. T.I. Oh, H. Wi, D.Y. Kim, P.J. Yoo, E.J. Woo, A fully parallel multi-frequency EIT system with flexible electrode configuration: KHU Mark2. Physiol. Meas. **32**(7), 835 (2011)

30. A. Tamburrino, Monotonicity based imaging methods for elliptic and parabolic inverse problems. J. Inverse Ill-Posed Probl. **14**(6), 633–642 (2006)
31. A. Tamburrino, G. Rubinacci, A new non-iterative inversion method for electrical resistance tomography. Inverse Probl. **18**(6), 1809 (2002)
32. H. Wi, H. Sohal, A.L. McEwan, E.J. Woo, T.I. Oh, Multi-frequency electrical impedance tomography system with automatic self-calibration for long-term monitoring. IEEE Trans. Biomed. Circuits Syst. **8**(1), 119–128 (2014)
33. L. Zhou, B. Harrach, J.K. Seo, Monotonicity-based electrical impedance tomography lung imaging. Preprint (2015)

An SVD in Spherical Surface Wave Tomography

Ralf Hielscher, Daniel Potts, and Michael Quellmalz

Abstract In spherical surface wave tomography, one measures the integrals of a function defined on the sphere along great circle arcs. This forms a generalization of the Funk–Radon transform, which assigns to a function its integrals along full great circles. We show a singular value decomposition (SVD) for the surface wave tomography provided we have full data.

Since the inversion problem is overdetermined, we consider some special cases in which we only know the integrals along certain arcs. For the case of great circle arcs with fixed opening angle, we also obtain an SVD that implies the injectivity, generalizing a previous result for half circles in Groemer (Monatsh Math 126(2):117–124, 1998). Furthermore, we derive a numerical algorithm based on the SVD and illustrate its merchantability by numerical tests.

1 Introduction

While the famous two-dimensional Radon transform assigns to a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ all its line integrals, its spherical generalization, the Funk–Radon transform $\mathscr{F}\colon C(\mathbb{S}^2) \to C(\mathbb{S}^2)$, assigns to a function on the two-dimensional sphere $\mathbb{S}^2 = \{\boldsymbol{\xi} \in \mathbb{R}^3 : |\boldsymbol{\xi}| = 1\}$ its integrals

$$\mathscr{F}f(\boldsymbol{\xi}) = \frac{1}{2\pi} \int_{\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle = 0} f(\boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\eta}, \qquad \boldsymbol{\xi} \in \mathbb{S}^2,$$

along all great circles $\{\boldsymbol{\eta} \in \mathbb{S}^2 : \boldsymbol{\eta} \perp \boldsymbol{\xi}\}$, $\boldsymbol{\xi} \in \mathbb{S}^2$. The investigation of the Funk–Radon transform dates back to the work of Funk [11], who showed the injectivity of the operator $\mathscr{F}$ for even functions. In other publications, the operator $\mathscr{F}$ is

R. Hielscher · D. Potts · M. Quellmalz (✉)
Faculty of Mathematics, Chemnitz University of Technology, Reichenhainer Straße 39, 09126 Chemnitz, Germany
e-mail: ralf.hielscher@mathematik.tu-chemnitz.de; daniel.potts@mathematik.tu-chemnitz.de; michael.quellmalz@mathematik.tu-chemnitz.de

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_7

also known as Funk transform, Minkowski–Funk transform or spherical Radon transform.

Similar to the Radon transform, the Funk–Radon transform plays an important role in imaging. Motivated by specific imaging modalities, the Funk–Radon transform has been generalized further to other paths of integration, namely circles with fixed diameter [34, 38], circles containing the north pole [1, 5, 20, 35], circles perpendicular to the equator [12, 23, 44], and nongeodesic hyperplane sections of the sphere [30, 31, 33, 37]. The integrals along half great circles have been investigated in [13, 18, 36]. Interestingly, some of these generalizations lead to injective operators.

In this paper, we replace the great circles as paths of integration in the Funk–Radon transform by great circle arcs with arbitrary opening angle. Let $\boldsymbol{\xi}$, $\boldsymbol{\zeta} \in \mathbb{S}^2$ be two points on the sphere that are not antipodal and denote by $\gamma(\boldsymbol{\xi}, \boldsymbol{\zeta})$ the shortest geodesic connecting both points. Then we aim at recovering $f\colon \mathbb{S}^2 \to \mathbb{C}$ from the integrals

$$g(\boldsymbol{\xi}, \boldsymbol{\zeta}) = \int_{\gamma(\boldsymbol{\xi}, \boldsymbol{\zeta})} f(\boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\eta}, \quad \boldsymbol{\xi}, \boldsymbol{\zeta} \in \mathbb{S}^2, \; \boldsymbol{\xi} \neq -\boldsymbol{\zeta}. \tag{1}$$

The study of this problem is motivated by spherical surface wave tomography. There, one measures the time a seismic wave travels along the Earth's surface from an epicenter to a receiver. Knowing the traveltimes of such waves between many pairs of epicenters and receivers, one wants to recover the local phase velocity. A common approach is the great circle ray approximation, where it is assumed that a wave travels along the arc of the great circle connecting epicenter and receiver. Then the traveltime of the wave equals the integral of the "slowness function" along the great circle arc connecting the epicenter and the receiver, where the slowness function is defined as one over the local phase velocity [29, 40, 43]. Hence, recovering the local phase velocity as a real-valued spherical function from its mean values along certain arcs of great circles is modeled by (1), see [3].

Although (1) uses a very intuitive parametrization of great circle arcs on the sphere, it is not well suited for analyzing the underlying operator since the arc length is restricted to $[0, \pi)$ and, even for continuous f, the function g has no continuous extension to a function on $\mathbb{S}^2 \times \mathbb{S}^2$. Therefore, we parameterize the manifold of all great circle arcs by the arclength $2\psi \in [0, 2\pi]$ and the rotation $Q \in \mathrm{SO}(3)$ that maps the arc of integration to the equator such that the midpoint of the arc is mapped to $(1, 0, 0)^\top$. Using this parametrization, the arc transform is defined in Sect. 3.1 as an operator

$$\mathscr{A}\colon C(\mathbb{S}^2) \to C(\mathrm{SO}(3) \times [0, \pi]).$$

In Theorem 3.3, we derive a singular value decomposition of $\mathscr{A}$, which involves spherical harmonics in $L^2(\mathbb{S}^2)$ and Wigner-D functions in $L^2(\mathrm{SO}(3) \times [0, \pi])$. Furthermore, we give upper and lower bounds for the singular values.

Since the function f lives on a two-dimensional manifold but the transformed function $\mathscr{A}f$ lives on a four-dimensional manifold, the inverse problem is highly overdetermined. For this reason, we consider in Sect. 4 specific subsets of arcs that still allow for the reconstruction of the function f. Most notably, we investigate in Sect. 4.3 the restriction of the arc transform to arcs of constant opening angle. This restriction includes as special cases the ordinary Funk–Radon transform as well as the half circle transform [18]. For the restricted operator, we prove in Theorem 4.4 a singular value decomposition and show that the singular values decay as $(n+\frac{1}{2})^{-\frac{1}{2}}C(\psi,n)$. While for opening angles $2\psi < \pi$ the constant $C(\psi,n)$ is independent of n, it converges to zero for odd n and $2\psi \to 2\pi$.

Finally, we present in Sect. 5 a numerical algorithm for the arc transform with fixed opening angle, which is based on the nonequispaced fast spherical Fourier transform [26] and the nonequispaced fast SO(3) Fourier transform [32].

2 Fourier Analysis on $\mathbb{S}^2$ and SO(3)

In this section, we present some basic facts about harmonic analysis on the sphere $\mathbb{S}^2$ and the rotation group SO(3) and introduce the notation we will use later on.

2.1 *Harmonic Analysis on the Sphere*

In this section, we are going to summarize some basic facts about harmonic analysis on the sphere as it can be found, e.g., in [7, 10, 28]. We denote by $\mathbb{Z}$ the set of integers and with $\mathbb{N}_0$ the nonnegative integers.

We define the two-dimensional sphere $\mathbb{S}^2 = \{\boldsymbol{\xi} \in \mathbb{R}^3 : |\boldsymbol{\xi}| = 1\}$ as the set of unit vectors $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^\top$ in the three-dimensional Euclidean space and make use of its parametrization in terms of the spherical coordinates

$$\boldsymbol{\xi}(\varphi,\vartheta) = (\cos\varphi\,\sin\vartheta, \sin\varphi\,\sin\vartheta, \cos\vartheta)^\top, \quad \varphi \in [0,2\pi),\ \vartheta \in [0,\pi].$$

Let $f\colon \mathbb{S}^2 \to \mathbb{C}$ be some measurable function. With respect to spherical coordinates, the surface measure $\mathrm{d}\boldsymbol{\xi}$ on the sphere reads as

$$\int_{\mathbb{S}^2} f(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \int_0^\pi \int_0^{2\pi} f(\boldsymbol{\xi}(\varphi,\vartheta))\,\sin\vartheta\,\mathrm{d}\varphi\,\mathrm{d}\vartheta.$$

The Hilbert space $L^2(\mathbb{S}^2)$ is the space of all measurable functions $f\colon \mathbb{S}^2 \to \mathbb{C}$ whose norm $\|f\|_{L^2(\mathbb{S}^2)} = (\langle f,f\rangle_{L^2(\mathbb{S}^2)})^{1/2}$ is finite, where $\langle f,g\rangle_{L^2(\mathbb{S}^2)} = \int_{\mathbb{S}^2} f(\boldsymbol{\xi})\,\overline{g(\boldsymbol{\xi})}\,\mathrm{d}\boldsymbol{\xi}$ denotes the usual L^2–inner product.

We define the associated Legendre functions

$$P_n^k(t) = \frac{(-1)^k}{2^n n!} \left(1-t^2\right)^{k/2} \frac{\mathrm{d}^{n+k}}{\mathrm{d}t^{n+k}} \left(t^2-1\right)^n, \quad t \in [-1,1],$$

of degree $n \in \mathbb{N}_0$ and order $k = 0, \ldots, n$. We define the normalized associated Legendre functions by

$$\widetilde{P}_n^k = \sqrt{\frac{2n+1}{4\pi} \frac{(n-k)!}{(n+k)!}} P_n^k \tag{2}$$

and

$$\widetilde{P}_n^{-k} = (-1)^k \widetilde{P}_n^k,$$

where the factor $(-1)^k$ is called Condon–Shortley phase, which is omitted by some authors.

An orthonormal basis in the Hilbert space $L^2(\mathbb{S}^2)$ of square integrable functions on the sphere is formed by the spherical harmonics

$$Y_n^k(\boldsymbol{\xi}(\varphi,\vartheta)) = \widetilde{P}_n^k(\cos\vartheta)\,\mathrm{e}^{\mathrm{i}k\varphi} \tag{3}$$

of degree $n \in \mathbb{N}_0$ and order $k = -n, \ldots, n$. Accordingly, any function $f \in L^2(\mathbb{S}^2)$ can be expressed by its spherical Fourier series

$$f = \sum_{n=0}^{\infty} \sum_{k=-n}^{n} \hat{f}_n^k \, Y_n^k$$

with the spherical Fourier coefficients

$$\hat{f}_n^k = \int_{\mathbb{S}^2} f(\boldsymbol{\xi}) \, \overline{Y_n^k(\boldsymbol{\xi})} \, \mathrm{d}\boldsymbol{\xi}, \quad n \in \mathbb{N}_0, \ k = -n, \ldots, n.$$

We define the space of spherical polynomials of degree up to $N \in \mathbb{N}_0$ by

$$\mathscr{P}_N = \operatorname{span}\left\{ Y_n^k : n = 0, \ldots, N, \ k = -n, \ldots, n \right\}.$$

2.2 Rotational Harmonics

We state some facts about functions on the rotation group SO(3). This introduction is based on [19], rotational Fourier transforms date back to Wigner, 1931, see [42]. The rotation group SO(3) consists of all orthogonal 3×3-matrices with determinant one equipped with the matrix multiplication as group operation. Every rotation $Q \in \mathrm{SO}(3)$ can be expressed in terms of its Euler angles α, β, γ by

$$Q(\alpha, \beta, \gamma) = R_3(\alpha)R_2(\beta)R_3(\gamma), \quad \alpha, \gamma \in [0, 2\pi),\ \beta \in [0, \pi],$$

where $R_i(\alpha)$ denotes the rotation of the angle α about the $\boldsymbol{\xi}_i$-axis, i.e.,

$$R_3(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad R_2(\beta) = \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}.$$

Note that we use this *zyz*-convention of Euler angles throughout this paper. The integral of a function $g \colon \mathrm{SO}(3) \to \mathbb{C}$ on the rotation group is given by

$$\int_{\mathrm{SO}(3)} g(Q)\,\mathrm{d}Q = \int_0^{2\pi} \int_0^{\pi} \int_0^{2\pi} g(Q(\alpha, \beta, \gamma)) \sin(\beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta\,\mathrm{d}\gamma.$$

We define the rotational harmonics or Wigner D-functions $D_n^{k,j}$ of degree $n \in \mathbb{N}_0$ and orders $k, j \in \{-n, \dots, n\}$ by

$$D_n^{k,j}(Q(\alpha, \beta, \gamma)) = \mathrm{e}^{-\mathrm{i}k\alpha} d_n^{k,j}(\cos\beta)\mathrm{e}^{-\mathrm{i}j\gamma},$$

where the Wigner d-functions are given by [41, p. 77]

$$d_n^{k,j}(t) = \frac{(-1)^{n-j}}{2^n} \sqrt{\frac{(n+k)!(1-t)^{j-k}}{(n-j)!(n+j)!(n-k)!(1+t)^{j+k}}} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{n-k} \frac{(1+t)^{n+j}}{(1-t)^{-n+j}}.$$

The Wigner d-functions satisfy the orthogonality relation

$$\int_{-1}^{1} d_n^{k,j}(t)\, d_{n'}^{k,j}(t)\,\mathrm{d}t = \frac{2\delta_{n,n'}}{2n+1}.$$

We define the space of square-integrable functions $L^2(\mathrm{SO}(3))$ with inner product $\langle f, g\rangle_{L^2(\mathrm{SO}(3))} = \int_{\mathrm{SO}(3)} f(Q)\,\overline{g(Q)}\,\mathrm{d}Q$. By the Peter–Weyl theorem, the rotational harmonics $D_n^{k,j}$ are complete in $L^2(\mathrm{SO}(3))$ and satisfy the orthogonality relation

$$\left\langle D_n^{k,j}, D_{n'}^{k',j'} \right\rangle_{L^2(\mathrm{SO}(3))} = \int_{\mathrm{SO}(3)} D_n^{k,j}(Q)\,\overline{D_{n'}^{k',j'}(Q)}\,\mathrm{d}Q = \frac{8\pi^2}{2n+1}\,\delta_{n,n'}\delta_{k,k'}\delta_{j,j'}. \tag{4}$$

We define the rotational Fourier coefficients of $g \in L^2(\mathrm{SO}(3))$ by

$$\hat{g}_n^{k,j} = \frac{2n+1}{8\pi^2} \left\langle g, D_n^{k,j} \right\rangle_{L^2(\mathrm{SO}(3))}, \quad n \in \mathbb{N}_0,\ k, j = -n, \dots, n. \tag{5}$$

Then the rotational Fourier expansion of g holds

$$g = \sum_{n=0}^{\infty} \sum_{k,j=-n}^{n} \hat{g}_n^{k,j} \, D_n^{k,j}.$$

The rotational Fourier transform is also known as SO(3) Fourier transform (SOFT) or Wigner D-transform.

The rotational harmonics $D_n^{k,j}$ are eigenfunctions of the Laplace–Beltrami operator on SO(3) with the corresponding eigenvalues $-n(n+1)$. The rotational harmonics $D_n^{j,k}$ are the matrix entries of the left regular representations of SO(3), see [21, 41]. In particular, the rotation of a spherical harmonic satisfies

$$Y_n^k(Q^{-1}\boldsymbol{\xi}) = \sum_{j=-n}^{n} D_n^{j,k}(Q) \, Y_n^j(\boldsymbol{\xi}). \tag{6}$$

2.3 Singular Value Decomposition

Let $\mathscr{K}\colon X \to Y$ be a compact linear operator between the separable Hilbert spaces X and Y. A singular system $\{(u_n, v_n, \sigma_n) : n \in \mathbb{N}_0\}$ consists of an orthonormal basis $\{u_n\}_{n=0}^{\infty}$ of X, an orthonormal basis $\{v_n\}_{n=0}^{\infty}$ in the closed range of $\mathscr{K}$ and singular values $\sigma_n \to 0$ such that operator $\mathscr{K}$ can be diagonalized as

$$\mathscr{K}x = \sum_{n=0}^{\infty} \sigma_n \, \langle x, u_n \rangle \, v_n, \qquad x \in X.$$

If all singular values σ_n are nonzero, the operator $\mathscr{K}$ is injective and for $y = \mathscr{K}x$, we have

$$x = \sum_{n=0}^{\infty} \frac{\langle y, v_n \rangle}{\sigma_n} u_n.$$

The instability of an inverse problem can be characterized by the decay of the singular values. The problem of solving $\mathscr{K}x = y$ for x is called mildly ill-posed of degree $\alpha > 0$ if $\sigma_n \in \mathscr{O}(n^{-\alpha})$, cf. [8, Sec. 2.2].

3 Circle Arcs

For any two points $\boldsymbol{\xi}, \boldsymbol{\zeta}$ on the sphere $\mathbb{S}^2$ that are not antipodal, there exists a shortest geodesic $\gamma(\boldsymbol{\xi}, \boldsymbol{\zeta})$ between $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$. This geodesic is an arc of the great circle that contains $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$.

The manifold of all great circle arcs is four-dimensional since they are determined by two points $\boldsymbol{\xi}, \boldsymbol{\zeta} \in \mathbb{S}^2$ and only coincide when $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are interchanged.

3.1 The Arc Transform

A great circle arc $\gamma(\boldsymbol{\xi}, \boldsymbol{\zeta})$ can be parameterized by its length $2\psi = \arccos(\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle)$ and a rotation $Q \in \mathrm{SO}(3)$ which is defined as follows. Let

$$\boldsymbol{e}_\varphi = (\cos\varphi,\ \sin\varphi,\ 0)^\top \in \mathbb{S}^2$$

be the point on the equator of $\mathbb{S}^2$ with latitude $\varphi \in \mathbb{R}$. Then there exists a unique rotation $Q \in \mathrm{SO}(3)$ such that $Q(\boldsymbol{\xi}) = \boldsymbol{e}_{-\psi}$ and $Q(\boldsymbol{\zeta}) = \boldsymbol{e}_\psi$. Such an arc γ and its rotation are depicted in Fig. 1. With this definition, the integral over the arc $\gamma(\boldsymbol{\xi}, \boldsymbol{\zeta})$ may be rewritten as

$$\int_{\gamma(\boldsymbol{\xi},\boldsymbol{\zeta})} f(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} = \int_{Q\gamma(\boldsymbol{\xi},\boldsymbol{\zeta})} f(Q^{-1}\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} = \int_{-\psi}^{\psi} f \circ Q^{-1}(\boldsymbol{e}_\varphi)\,\mathrm{d}\varphi.$$

This motivates the following definition of the arc transform

$$\begin{aligned} &\mathscr{A}\colon C(\mathbb{S}^2) \to C(\mathrm{SO}(3) \times [0,\pi]), \\ &\mathscr{A}f(Q,\psi) = \int_{-\psi}^{\psi} f \circ Q^{-1}(\boldsymbol{e}_\varphi)\,\mathrm{d}\varphi. \end{aligned} \tag{7}$$

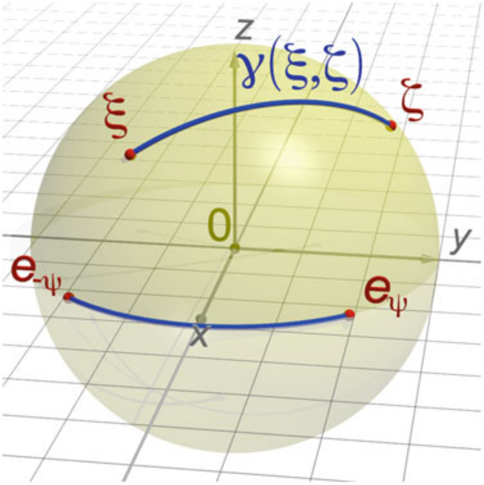


Fig. 1 Visualization of the arc $\gamma(\boldsymbol{\xi}, \boldsymbol{\zeta})$

The great circle arcs $\gamma(\boldsymbol{\xi}, \boldsymbol{\zeta})$ and $\gamma(\boldsymbol{\zeta}, \boldsymbol{\xi})$ are identical. This symmetry also holds for the operator $\mathscr{A}$. Using Euler angles, we have the identity

$$\mathscr{A}f(Q(\alpha, \beta, \gamma), \psi) = \mathscr{A}f(Q(2\pi - \alpha, \pi - \beta, \gamma + \pi), \psi),$$

where we assume the Euler angle γ as 2π-periodic.

3.2 Singular Value Decomposition of the Arc Transform

In the following, we use double factorials defined by $n!! = n(n-2)\cdots 2$ for n even or $n!! = n(n-2)\cdots 1$ for n odd and $0!! = (-1)!! = 1$. The next theorem shows how the arc transform $\mathscr{A}$ acts on spherical harmonics Y_n^k. The corresponding result for the parametrization in terms of the endpoints of an arc is found in [6, Appx. C], see also [3].

Theorem 3.1 *Let $n \in \mathbb{N}_0$ and $k \in \{-n, \dots, n\}$. Then*

$$\mathscr{A}Y_n^k(Q, \psi) = \sum_{j=-n}^{n} \widetilde{P}_n^j(0)\, D_n^{j,k}(Q)\, s_j(\psi), \tag{8}$$

where

$$s_j(\psi) = \begin{cases} 2\psi, & j = 0 \\ \frac{2\sin(j\psi)}{j}, & j \neq 0 \end{cases} \tag{9}$$

and

$$\widetilde{P}_n^j(0) = \begin{cases} (-1)^{\frac{n+j}{2}} \sqrt{\frac{2n+1}{4\pi} \frac{(n-j-1)!!(n+j-1)!!}{(n-j)!!(n+j)!!}}, & n + j \text{ even} \\ 0, & n + j \text{ odd.} \end{cases} \tag{10}$$

Proof By (6), we obtain

$$\mathscr{A}Y_n^k(Q, \psi) = \int_{-\psi}^{\psi} Y_n^k(Q^{-1}(\boldsymbol{e}_\varphi)) = \sum_{j=-n}^{n} D_n^{j,k}(Q) \int_{-\psi}^{\psi} Y_n^j(\boldsymbol{e}_\varphi)\, \mathrm{d}\varphi.$$

By the definition (3) of the spherical harmonics, we see that

$$\int_{-\psi}^{\psi} Y_n^j(\boldsymbol{e}_\varphi)\, \mathrm{d}\varphi = \widetilde{P}_n^j(0) \int_{-\psi}^{\psi} \mathrm{e}^{\mathrm{i}j\varphi}\, \mathrm{d}\varphi = \widetilde{P}_n^j(0)\, s_j(\psi).$$

Hence,

$$\mathscr{A} Y_n^k(Q,\psi) = \sum_{j=-n}^{n} D_n^{j,k}(Q)\,\widetilde{P}_n^j(0)\,s_j(\psi).$$

Now we calculate $\widetilde{P}_n^j(0)$. By Hielscher and Quellmalz [23], $P_n^j(0) = 0$ if $n+j$ is odd and otherwise

$$P_n^j(0) = (-1)^{\frac{n+j}{2}} \frac{(n+j-1)!!}{(n-j)!!}.$$

Hence, we obtain by (2)

$$\begin{aligned}\widetilde{P}_n^j(0) &= \sqrt{\frac{2n+1}{4\pi}\frac{(n-j)!}{(n+j)!}}\,(-1)^{\frac{n+j}{2}}\frac{(n+j-1)!!}{(n-j)!!}\\ &= (-1)^{\frac{n+j}{2}}\sqrt{\frac{2n+1}{4\pi}\frac{(n-j-1)!!(n+j-1)!!}{(n-j)!!(n+j)!!}}\end{aligned}$$

if $n+j$ is even, which implies (10). ■

Lemma 3.2 *Let $n \in \mathbb{N}_0$ and $j \in \{-n,\ldots,n\}$. If $n+j$ is odd, then $\widetilde{P}_n^j(0) = 0$. Otherwise, we have*

$$\frac{2n+1}{2\pi^2\sqrt{(n+1)^2-j^2}} \le \left|\widetilde{P}_n^j(0)\right|^2 \le \frac{2n+1}{4\pi\sqrt{(n+1)^2-j^2}}. \tag{11}$$

Furthermore, for $j \in \mathbb{N}_0$,

$$\lim_{\substack{n\to\infty\\ n+j\ even}} \left|\widetilde{P}_n^j(0)\right| = \frac{1}{\pi}. \tag{12}$$

Proof We first show that for $m \in \mathbb{N}$,

$$\sqrt{\frac{2}{\pi(2m+1)}} \le \frac{(2m-1)!!}{(2m)!!} \le \frac{1}{\sqrt{2m+1}}. \tag{13}$$

With the definition

$$u(m) = \left(\frac{(2m)!!}{(2m-1)!!}\right)^2 \frac{1}{2m+1}, \qquad m \in \mathbb{N}_0,$$

we see that $u(0) = 1$ and u is increasing because of $m \geq 1$ and

$$\frac{u(m)}{u(m-1)} = \frac{(2m)^2}{(2m-1)^2} \frac{2m-1}{2m+1} = \frac{(2m)^2}{(2m)^2-1} > 1.$$

That implies the right inequality of (13). Furthermore, Wallis' product states the convergence

$$u(m) = \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}\cdots\frac{2m}{2m-1}\frac{2m}{2m+1} \longrightarrow \frac{\pi}{2} \tag{14}$$

for $m \to \infty$, see also [4]. This shows the left inequality of (13).

By (10) and (13), we obtain the upper bound

$$\left|\widetilde{P}_n^j(0)\right|^2 = \frac{2n+1}{4\pi} \frac{(n-j-1)!!}{(n-j)!!} \frac{(n+j-1)!!}{(n+j)!!} \leq \frac{2n+1}{4\pi} \frac{1}{\sqrt{n-j+1}} \frac{1}{\sqrt{n+j+1}}.$$

The lower bound follows analogously. Moreover, we have

$$\left|\widetilde{P}_{j+2m}^j(0)\right|^2 = \frac{2(j+2m)+1}{4\pi} \frac{(2m-1)!!}{(2m)!!} \frac{(2m+2j-1)!!}{(2m+2j)!!}.$$

Hence, Wallis product (14) shows that for $j \in \mathbb{N}_0$

$$\begin{aligned}\lim_{m\to\infty} \left|\widetilde{P}_{j+2m}^j(0)\right|^2 &= \lim_{m\to\infty} \frac{2(j+2m)+1}{4\pi} \frac{2}{\pi} \frac{1}{\sqrt{2m+1}\sqrt{2m+2j+1}} \\ &= \lim_{m\to\infty} \frac{2m+j+\frac{1}{2}}{\pi^2\sqrt{(2m+j+1)^2-j^2}} = \frac{1}{\pi^2},\end{aligned}$$

which proves the assertion. ■

Next, we derive a singular value decomposition for the spherical arc transform. To this end, we define for $n \in \mathbb{N}_0$ and $k = -n, \ldots, n$ the functions $E_n^k \in L^2(\mathrm{SO}(3) \times [0, \pi])$ by

$$E_n^k(Q, \psi) = \sum_{j=-n}^{n} D_n^{j,k}(Q)\, \widetilde{P}_n^j(0)\, s_j(\psi), \qquad Q \in \mathrm{SO}(3),\ \psi \in [0, \pi]. \tag{15}$$

Theorem 3.3 *The operator* $\mathscr{A}\colon L^2(\mathbb{S}^2) \to L^2(\mathrm{SO}(3) \times [0, \pi])$ *is compact with the singular value decomposition*

$$\left\{\left(Y_n^k,\ \widetilde{E}_n^k,\ \sigma_n\right) : n \in \mathbb{N}_0,\ k \in \{-n, \ldots, n\}\right\},$$

with the singular values

$$\sigma_n = \left\| E_n^k \right\|_{L^2(\mathrm{SO}(3)\times[0,\pi])} = \sqrt{\frac{32\pi^3}{2n+1}} \sqrt{\frac{\pi^2}{3}\left|\widetilde{P}_n^0(0)\right|^2 + \sum_{j=1}^{n} \frac{1}{j^2}\left|\widetilde{P}_n^j(0)\right|^2} \tag{16}$$

satisfying

$$\sqrt{\frac{16}{3}\pi^3} \le \sigma_n \sqrt{n+1} \le \sqrt{\frac{8}{3}\pi^4 + 4\pi^2}, \qquad n \text{ even}, \tag{17}$$

$$4\sqrt{\pi} \le \sigma_n \sqrt{n+1} \le 2\pi \sqrt{\frac{4}{\sqrt{3}} + 1}, \qquad n \text{ odd}, \tag{18}$$

and the orthonormal function system $\widetilde{E}_k^n = \sigma_n^{-1} E_n^k$*,* $n \in \mathbb{N}_0$*,* $k \in \{-n, \dots, n\}$ *in* $L^2(\mathrm{SO}(3) \times [0, \pi])$*.*

Proof By the orthogonality (4) of the rotational harmonics, we have

$$\begin{aligned}
&\left\langle E_n^k, E_{n'}^{k'} \right\rangle_{L^2(\mathrm{SO}(3)\times[0,\pi])} \\
&= \sum_{j=-n}^{n} \sum_{j'=-n'}^{n'} \widetilde{P}_n^j(0)\,\widetilde{P}_{n'}^{j'}(0) \int_{\mathrm{SO}(3)} D_n^{j,k}(Q)\,\overline{D_{n'}^{j',k'}(Q)}\,\mathrm{d}Q \int_0^\pi s_j(\psi)\,s_{j'}(\psi)\,\mathrm{d}\psi \\
&= \sum_{j=-n}^{n} \sum_{j'=-n'}^{n'} \frac{8\pi^2}{2n+1} \delta_{nn'}\,\delta_{kk'}\,\delta_{jj'}\,\widetilde{P}_n^j(0)\,\widetilde{P}_{n'}^{j'}(0) \int_0^\pi s_j(\psi)\,s_{j'}(\psi)\,\mathrm{d}\psi \\
&= \delta_{nn'}\,\delta_{kk'} \sum_{j=-n}^{n} \frac{8\pi^2}{2n+1} \left|\widetilde{P}_n^j(0)\right|^2 \int_0^\pi s_j(\psi)^2\,\mathrm{d}\psi \\
&= \delta_{nn'}\,\delta_{kk'} \frac{8\pi^2}{2n+1} \sum_{j=-n}^{n} \left|\widetilde{P}_n^j(0)\right|^2 \begin{cases} \frac{4\pi^3}{3}, & j = 0 \\ \frac{2\pi}{j^2}, & j \neq 0. \end{cases}
\end{aligned}$$

This shows that the functions $\mathscr{A} Y_n^k$ are orthogonal in the space $L^2(\mathrm{SO}(3) \times [0, \pi])$ and have the norm

$$\begin{aligned}
\left\| E_n^k \right\|_{L^2(\mathrm{SO}(3)\times[0,\pi])}^2 &= \frac{8\pi^2}{2n+1} \sum_{j=-n}^{n} \left|\widetilde{P}_n^j(0)\right|^2 \begin{cases} \frac{4\pi^3}{3}, & j = 0 \\ \frac{2\pi}{j^2}, & j \neq 0. \end{cases} \\
&= \frac{16\pi^3}{2n+1} \left(\frac{2\pi^2}{3} \left|\widetilde{P}_n^0(0)\right|^2 + 2 \sum_{j=1}^{n} \frac{1}{j^2} \left|\widetilde{P}_n^j(0)\right|^2 \right),
\end{aligned}$$

where we used that $\left|\widetilde{P}_n^j(0)\right| = \left|\widetilde{P}_n^{-j}(0)\right|$. In order to prove that $\mathscr{A}$ is compact, we show that the singular values σ_n decay for $n \to \infty$. We have by Lemma 3.2 for $n = 2m$ even

$$\sigma_{2m}^2 \leq 4\pi^2 \left(\frac{2\pi^2}{3} \frac{1}{2m+1} + 2 \sum_{j=1}^{m} \frac{1}{(2j)^2} \frac{1}{\sqrt{(2m+1)^2 - (2j)^2}} \right).$$

Replacing the sum by an integral, we estimate for n even

$$\begin{aligned} 2 \sum_{j=1}^{m} \frac{1}{(2j)^2} \frac{1}{\sqrt{(2m+1)^2 - (2j)^2}} &\leq 2 \int_{1/2}^{m+1/2} \frac{1}{(2j)^2} \frac{1}{\sqrt{(2m+1)^2 - (2j)^2}} \,\mathrm{d}j \\ &= 2 \left[-\frac{\sqrt{(2m+1)^2 - (2j)^2}}{2j(2m+1)^2} \right]_{1/2}^{m+1/2} \\ &= 2 \frac{\sqrt{m^2+m}}{(2m+1)^2} \leq \frac{1}{2m+1}, \end{aligned}$$

where we made use of the convexity of the integrand. Hence,

$$\sigma_{2m}^2 \leq 4\pi^2 \left(\frac{2\pi^2}{3} \frac{1}{2m+1} + \frac{1}{2m+1} \right) = 4\pi^2 \left(\frac{2\pi^2}{3} + 1 \right) \frac{1}{2m+1}.$$

For odd $n = 2m - 1$, we proceed analogously. We have

$$\sigma_{2m-1}^2 \leq 8\pi^2 \sum_{j=1}^{m} \frac{1}{(2j-1)^2} \frac{1}{\sqrt{(2m)^2 - (2j-1)^2}}.$$

Note that, for the estimation of the sum by an integral, we extract the summand for $j = 1$

$$\begin{aligned} \sigma_{2m-1}^2 &\leq 8\pi^2 \left(\frac{1}{\sqrt{(2m)^2 - 1}} + \int_{1}^{m+1/2} \frac{1}{(2j-1)^2} \frac{1}{\sqrt{(2m)^2 - (2j-1)^2}} \,\mathrm{d}j \right) \\ &= 8\pi^2 \left(\frac{1}{\sqrt{(2m)^2 - 1}} + \frac{\sqrt{(2m)^2 - 1}}{2(2m)^2} \right) \\ &\leq 8\pi^2 \left(\frac{2}{\sqrt{3}\, 2m} + \frac{1}{2(2m)} \right) = 4\pi^2 \left(\frac{4}{\sqrt{3}} + 1 \right) \frac{1}{2m}. \end{aligned}$$

For the lower bound of the singular values, we also use Lemma 3.2. For even n, we extract the summand $j = 0$ and obtain

$$\begin{aligned}\sigma_n^2 &= \frac{16\pi^3}{2n+1}\left(\frac{2\pi^2}{3}\left|\widetilde{P}_n^0(0)\right|^2 + 2\sum_{j=1}^{n}\frac{1}{j^2}\left|\widetilde{P}_n^j(0)\right|^2\right)\\ &\geq \frac{32\pi^5}{3(2n+1)}\left|\widetilde{P}_n^0(0)\right|^2 \geq \frac{16\pi^3}{3(n+1)}.\end{aligned}$$

For odd n, we extract the summand $j = 1$ and obtain

$$\sigma_n^2 \geq \frac{32\pi^3}{2n+1}\left|\widetilde{P}_n^1(0)\right|^2 \geq \frac{16\pi}{\sqrt{(n+1)^2-1}} \geq \frac{16\pi}{n+1}.$$

■

The singular values σ_n decay with rate $n^{-1/2}$. This is the same asymptotic decay rate as of the eigenvalues of the Funk–Radon transform, cf. [39].

4 Special Cases

The recovery of a function f from the arc integrals $\mathscr{A}f$ is overdetermined considered we have full data. In the following subsections, we are going to examine some special cases, where we can reconstruct f from integrals only along certain arcs.

4.1 Arcs Starting in a Fixed Point

As a simple example, we fix one endpoint of the arcs. Without loss of generality, we assume that this endpoint is the north pole. The arc connecting the north pole $\boldsymbol{e}^3$ and an arbitrary other point $\boldsymbol{\xi}(\varphi, \vartheta) \in \mathbb{S}^2$ is given by

$$\gamma(\boldsymbol{e}^3, \boldsymbol{\xi}(\varphi, \vartheta)) = \{\boldsymbol{\eta}(\varphi, \varrho) \in \mathbb{S}^2 : \varrho \in [0, \vartheta]\}.$$

Since, with $Q = Q\left(\frac{\vartheta}{2}, \frac{\pi}{2}, \frac{3\pi}{2} - \varphi\right) \in \mathrm{SO}(3)$, we have $Q\boldsymbol{e}^3 = \boldsymbol{e}_{\frac{\vartheta}{2}}$ and $Q\boldsymbol{\xi} = \boldsymbol{e}_{-\frac{\vartheta}{2}}$. The restriction $\mathscr{B}\colon C(\mathbb{S}^2) \to C(\mathbb{S}^2)$ of the operator $\mathscr{A}$ to these arcs satisfies

$$\mathscr{B}f(\boldsymbol{\xi}(\varphi, \vartheta)) = \mathscr{A}f\left(Q\left(\frac{\vartheta}{2}, \frac{\pi}{2}, \frac{3\pi}{2} - \varphi\right), \frac{\vartheta}{2}\right) = \int_0^{\vartheta} f(\boldsymbol{\eta}(\varphi, \varrho))\,\mathrm{d}\varrho.$$

If f is additionally differentiable, it can be recovered from $\mathscr{B}f$ by

$$f(\boldsymbol{\xi}(\varphi,\vartheta)) = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathscr{B}f(\boldsymbol{\xi}(\varphi,\vartheta)).$$

The following more general result for injectivity is due to [2, Theorem 4.4.1]. Its proof uses a similar idea combined with an extension by density.

Proposition 4.1 *Let S be an open subset of $\mathbb{S}^2$ and $A, B \subset S$ nonempty sets with $\overline{A \cup B} = \overline{S}$. If $f \in C(\mathbb{S}^2)$ and*

$$\int_{\gamma(\boldsymbol{\xi},\boldsymbol{\zeta})} f(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} = 0 \qquad \textit{for all } \boldsymbol{\xi} \in A,\ \boldsymbol{\zeta} \in B,$$

then $f \equiv 0$ on S.

For $A = \{\boldsymbol{e}^3\}$ and $B = \mathbb{S}^2$, we have the arcs starting in the north pole.

4.2 Recovery of Local Functions

A subset $\Omega \subset \mathbb{S}^2$ is called convex if for any two points $\boldsymbol{\xi}, \boldsymbol{\eta} \in \Omega$ the geodesic arc $\gamma(\boldsymbol{\xi}, \boldsymbol{\eta})$ is contained in Ω. We denote by $\partial\Omega$ the boundary of Ω.

Theorem 4.2 *Let $f \in C(\mathbb{S}^2)$ and Ω be a convex subset of $\mathbb{S}^2$ whose closure $\overline{\Omega}$ is strictly contained in a hemisphere, i.e., there exists a $\boldsymbol{\zeta} \in \mathbb{S}^2$ such that $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle > 0$ for all $\boldsymbol{\xi} \in \overline{\Omega}$. If*

$$\int_{\gamma(\boldsymbol{\xi},\boldsymbol{\eta})} f(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} = 0 \qquad \textit{for all } \boldsymbol{\xi}, \boldsymbol{\eta} \in \partial\Omega, \tag{19}$$

then $f = 0$ on Ω.

Proof Without loss of generality, we assume that $\overline{\Omega}$ is strictly contained in the northern hemisphere, i.e., we have $\xi_3 > 0$ for all $\boldsymbol{\xi} \in \overline{\Omega}$. We define the restriction of f to $\overline{\Omega}$ by

$$f_\Omega(\boldsymbol{\xi}) = \begin{cases} f(\boldsymbol{\xi}), & \boldsymbol{\xi} \in \overline{\Omega} \\ 0, & \boldsymbol{\xi} \in \mathbb{S}^2 \setminus \overline{\Omega} \end{cases}$$

Since $\gamma(\boldsymbol{\xi}, \boldsymbol{\eta}) \subset \overline{\Omega}$ for all $\boldsymbol{\xi}, \boldsymbol{\eta} \in \partial\Omega$, the function f_Ω also satisfies (19).

For $\boldsymbol{\xi} \in \mathbb{S}^2$, denote with $\boldsymbol{\xi}^\perp = \{\boldsymbol{\eta} \in \mathbb{S}^2 : \langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle = 0\}$ the great circle perpendicular to $\boldsymbol{\xi}$. We show that the Funk–Radon transform

$$\mathscr{F}f_\Omega(\boldsymbol{\xi}) = \int_{\boldsymbol{\xi}^\perp \cap \overline{\Omega}} f_\Omega(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} + \int_{\boldsymbol{\xi}^\perp \setminus \overline{\Omega}} f_\Omega(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} \tag{20}$$

vanishes everywhere. The second summand of (20) vanishes because f_Ω is zero outside $\overline{\Omega}$ by definition. If $\boldsymbol{\xi}^\perp \cap \overline{\Omega}$ is not empty, there exist two points $\boldsymbol{\eta}^1, \boldsymbol{\eta}^2 \in \partial\Omega$ such that $\gamma(\boldsymbol{\eta}^1, \boldsymbol{\eta}^2) = \boldsymbol{\xi}^\perp \cap \overline{\Omega}$, which shows that also the first summand of (20) vanishes. Hence, $\mathscr{F} f_\Omega = 0$ on $\mathbb{S}^2$. Since the Funk–Radon transform $\mathscr{F}$ is injective for even functions, we see that f_Ω must be odd. Since f_Ω is supported strictly inside the northern hemisphere, so f_Ω must be the zero function. By the construction, we see that $f(\boldsymbol{\xi})$ vanishes for all $\boldsymbol{\xi} \in \Omega$. ■

An analogue to Theorem 4.2 for Ω being the northern hemisphere and the arcs being half circles is shown in [36].

4.3 Arcs with Fixed Length

In the following, we consider circle arcs with fixed length ψ. To this end, we define the restriction

$$\mathscr{A}_\psi(Q) = \mathscr{A}(Q, \psi).$$

Theorem 4.3 *Let $\psi \in (0, \pi)$ be fixed. The operator $\mathscr{A}_\psi : L^2(\mathbb{S}^2) \to L^2(\mathrm{SO}(3))$ has the singular value decomposition*

$$\left\{ \left(Y_n^k, Z_{n,\psi}^k, \mu_n(\psi) \right) : n \in \mathbb{N}_0,\ k \in \{-n, \dots, n\} \right\},$$

with the singular values

$$\mu_n(\psi) = \sqrt{\sum_{j=-n}^{n} \frac{8\pi^2}{2n+1} \left| \widetilde{P}_n^j(0) \right|^2 s_j(\psi)^2} \tag{21}$$

and the singular functions

$$Z_{n,\psi}^k = \frac{\mathscr{A}_\psi Y_n^k}{\mu_n(\psi)} = \frac{1}{\mu_n(\psi)} \sum_{j=-n}^{n} \widetilde{P}_n^j(0)\, s_j(\psi)\, D_n^{j,k}.$$

In particular, $\mathscr{A}_\psi$ is injective.

Proof Let $\psi \in (0, \pi)$ be fixed, $n \in \mathbb{N}_0$ and $k \in \{-n, \dots, n\}$. We have by (8)

$$\left\langle \mathscr{A}_\psi Y_n^k, \mathscr{A}_\psi Y_{n'}^{k'} \right\rangle_{L^2(\mathrm{SO}(3))}$$
$$= \sum_{j=-n}^{n} \sum_{j'=-n'}^{n'} \int_{\mathrm{SO}(3)} D_n^{j,k}(Q)\, \overline{D_{n'}^{j',k'}(Q)}\, \widetilde{P}_n^j(0)\, \widetilde{P}_{n'}^{j'}(0)\, s_j(\psi)\, s_{j'}(\psi)\, \mathrm{d}Q$$

$$= \sum_{j=-n}^{n} \sum_{j'=-n'}^{n'} \frac{8\pi^2}{2n+1} \delta_{nn'}\, \delta_{kk'}\, \delta_{jj'}\, \widetilde{P}_n^j(0)\, \widetilde{P}_{n'}^{j'}(0)\, s_j(\psi)\, s_{j'}(\psi)$$

$$= \delta_{nn'} \delta_{kk'} \sum_{j=-n}^{n} \frac{8\pi^2}{2n+1} \left|\widetilde{P}_n^j(0)\right|^2 s_j(\psi)^2.$$

For the injectivity, we check that the singular values $\mu_n(\psi)$ do not vanish for each $n \in \mathbb{N}_0$. We have $\widetilde{P}_n^j(0) = 0$ if and only if $n-j$ is odd. Furthermore, the definition of s_j in (9) shows that $s_0(\psi) = 2\psi$ vanishes if and only if $\psi = 0$ and $s_1(\psi) = 2\sin(\psi)$ vanishes if and only if ψ is an integer multiple of π. Hence, the functions $\mathscr{A}Y_n^k$ are also orthogonal in the space $L^2(\mathrm{SO}(3))$. ■

Theorem 4.4 *The singular values $\mu_n(\psi)$ of $\mathscr{A}_\psi$ satisfy for odd $n = 2m-1$*

$$\lim_{m\to\infty} \frac{4m-1}{4} \mu_{2m-1}(\psi)^2 = \begin{cases} 4\pi\psi, & \psi \in [0, \frac{\pi}{2}] \\ 4\pi^2 - 4\pi\psi, & \psi \in [\frac{\pi}{2}, \pi], \end{cases} \tag{22}$$

and for even $n = 2m$

$$\lim_{m\to\infty} \frac{4m+1}{4} \mu_{2m}(\psi)^2 = \begin{cases} 4\pi\psi, & \psi \in [0, \frac{\pi}{2}] \\ 12\pi\psi - 4\pi^2, & \psi \in [\frac{\pi}{2}, \pi]. \end{cases} \tag{23}$$

Proof We first show (22). Let $m \in \mathbb{N}$. We have by (21)

$$\frac{4m-1}{4} \mu_{2m-1}(\psi)^2 = 16\pi^2 \sum_{j=1}^{m} \left|\widetilde{P}_{2m-1}^{2j-1}(0)\right|^2 \frac{\sin^2((2j-1)\psi)}{(2j-1)^2}.$$

We denote by $\nu(\psi) = 4\pi\left(\frac{\pi}{2} - \left|\psi - \frac{\pi}{2}\right|\right)$ the right-hand side of (22). The Fourier cosine series of ν reads by [14, 1.444]

$$16 \sum_{k=1}^{\infty} \frac{\sin((2k-1)\psi)^2}{(2k-1)^2} = 16 \sum_{k=1}^{\infty} \frac{1 - \cos((2k-1)2\psi)}{2(2k-1)^2} = \nu(\psi), \qquad \psi \in [0, \pi].$$

We have

$$\left\| \frac{4m-1}{4} \mu_{2m-1}^2 - \nu \right\|_{C([0,\pi])} = \left\| 16 \sum_{j=1}^{\infty} \frac{\pi^2 \left|\widetilde{P}_{2m-1}^{2j-1}(0)\right|^2 - 1}{(2j-1)^2} \sin^2((2j-1)\psi) \right\|_{C([0,\pi])}$$

$$\leq \sum_{j=1}^{\infty} \frac{16 \left| \pi^2 \left|\widetilde{P}_{2m-1}^{2j-1}(0)\right|^2 - 1 \right|}{(2j-1)^2}. \tag{24}$$

We show that (24) goes to zero for $m \to \infty$, which then implies (22). By (12), we see that $\pi^2\, |\widetilde{P}_{2m-1}^{2j-1}(0)|^2$ converges to 1 for $m \to \infty$. Using the singular values (16) together with their bound (18), we obtain the following summable majorant of (24):

$$\sum_{j=1}^{\infty} \frac{16\pi^2 \left|\widetilde{P}_{2m-1}^{2j-1}(0)\right|^2}{(2j-1)^2} \le \frac{4m-1}{2\pi}\,\sigma_{2m-1}^2 \le 2\pi\,\frac{4m-1}{m}\left(\frac{4}{\sqrt{3}}+1\right).$$

Hence, the sum (24) converges to 0 for $m \to \infty$ by the dominated convergence theorem of Lebesgue.

In the second part, we show (23) for the odd singular values. Let $m \in \mathbb{N}$. We have

$$\frac{4m+1}{4}\,\mu_{2m}(\psi)^2 = 8\pi^2 \left|\widetilde{P}_{2m}^{0}(0)\right|^2 \psi^2 + 4\pi^2 \sum_{k=1}^{m} \left|\widetilde{P}_{2m}^{2k}(0)\right|^2 \frac{\sin^2(2k\psi)}{k^2}. \tag{25}$$

We examine both summands on the right side of (25). The first summand converges due to (12):

$$\lim_{m\to\infty} 8\pi^2 \left|\widetilde{P}_{2m}^{0}(0)\right|^2 \psi^2 = 8\psi^2.$$

We denote the second summand of (25) by

$$\lambda_m(\psi) = 4\pi^2 \sum_{k=1}^{m} \left|\widetilde{P}_{2m}^{2k}(0)\right|^2 \frac{\sin^2(2k\psi)}{k^2}$$

and define λ by the following Fourier cosine series, see [14, 1.443],

$$\lambda(\psi) = \sum_{k=1}^{\infty} \frac{\sin^2(2k\psi)}{k^2} = \sum_{k=1}^{\infty} \frac{1-\cos(4k\psi)}{2k^2} = \begin{cases} -2\psi^2 + \pi\psi, & \psi \in [0, \frac{\pi}{2}) \\ -2\psi^2 + 3\pi\psi - \pi^2, & \psi \in [\frac{\pi}{2}, \pi). \end{cases}$$

We have

$$\begin{aligned} \|\lambda_m - \lambda\|_{C([0,\pi])} &= \left\| \sum_{k=1}^{\infty} \frac{\pi^2 \left|\widetilde{P}_{2m}^{2k}(0)\right|^2 - 1}{k^2} \sin^2(2k\psi) \right\|_{C([0,\pi])} \\ &\le \sum_{k=1}^{\infty} \frac{\left|\pi^2 \left|\widetilde{P}_{2m}^{2k}(0)\right|^2 - 1\right|}{(2j-1)^2}. \end{aligned}$$

As in the first part of the proof, we see with (17) that the last sum goes to 0 for $m \to \infty$, which proves (23). ■

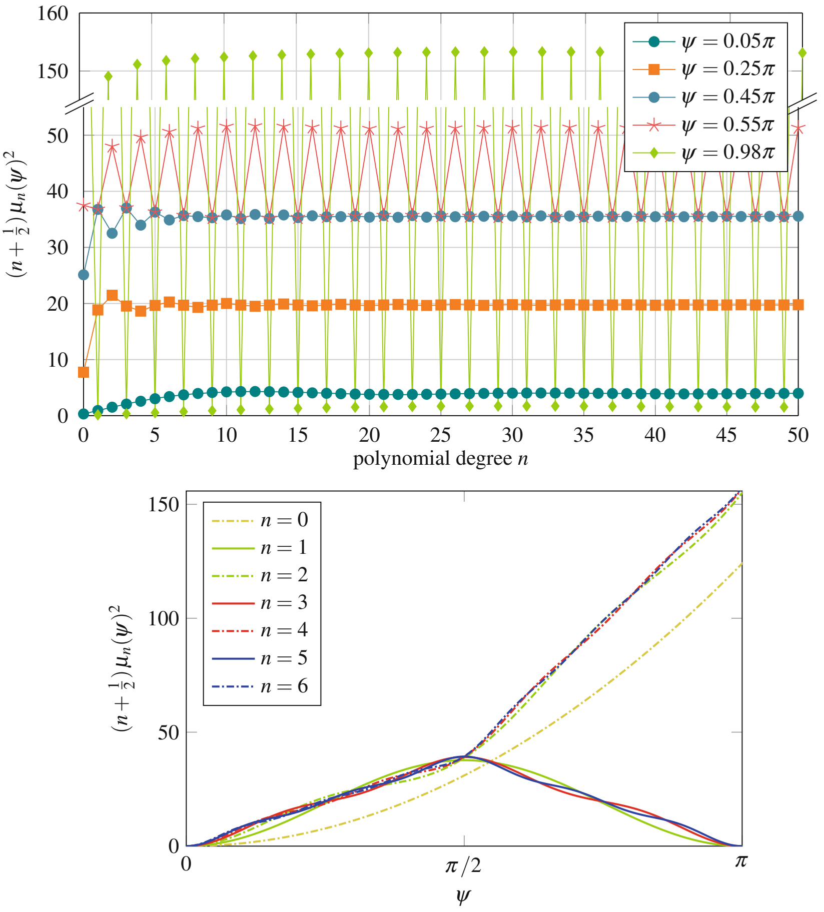


Fig. 2 The (normalized) singular values $(n + \frac{1}{2})\,\mu_n(\psi)^2$. Top: dependency on the degree n. Note the oscillation for $\psi > \frac{\pi}{2}$. Bottom: dependency on the arc-length ψ (dashed lines correspond to even n)

Remark 4.5 Theorem 4.4 shows that the singular values $\mu_n(\psi)$ decay with the same asymptotic rate of $n^{-1/2}$ as σ_n from Theorem 3.3. For $\psi < \frac{\pi}{2}$, the singular values $(n+\frac{1}{2})\,\mu_n(\psi)^2$ for even and odd n converge to the same limit. However, for $\psi > \frac{\pi}{2}$, the singular values for even n become larger than the ones for odd n. This might be explained by the fact that for odd n, the spherical harmonics Y_n^k are odd and integrating them along a circle arc with length 2ψ, which is longer than a half-circle, yields some cancellation. In the limiting case $\psi = \pi$, which is not covered by Theorem 4.3, $\mathscr{A}_\pi$ corresponds to the Funk–Radon transform, which is injective only for even functions and vanishes on odd functions. This behavior is illustrated in Fig. 2. □

Remark 4.6 Since the rotation group SO(3) is three-dimensional, the inversion of the arc transform $\mathscr{A}_\psi$ with fixed length is still overdetermined.

In the case $\psi = \frac{\pi}{2}$, we have the integrals along all half circles. The injectivity of the arc transform for half circles was shown in [18]. The restriction of the arc transform to all half circles that are subsets of either the upper or the lower hemisphere is still injective, see [36]. This is because every function that is supported in the upper (lower) hemisphere can be uniquely reconstructed by its Funk–Radon transform, which then integrates only over the half circles in the upper (lower) hemisphere. □

The singular value decomposition from Theorem 4.3 allows us to reconstruct a function $f \in L^2(\mathbb{S}^2)$ given $g = \mathscr{A}_\psi f$.

Theorem 4.7 *Let $f \in L^2(\mathbb{S}^2)$ and $g = \mathscr{A}_\psi f \in L^2(\mathrm{SO}(3))$. Then f can be reconstructed from the rotational Fourier coefficients $\hat{g}_n^{j,k}$ given in (5) by*

$$f = \sum_{n=0}^{\infty} \sum_{k=-n}^{n} \frac{\sum_{j=-n}^{n} \widetilde{P}_n^j(0)\, s_j(\psi)\, \hat{g}_n^{j,k}}{\sum_{j=-n}^{n} \widetilde{P}_n^j(0)^2\, s_j(\psi)^2}\, Y_n^k. \tag{26}$$

Proof We have by Theorem 4.3 for the spherical Fourier coefficients

$$\begin{aligned}\hat{f}_n^k &= \frac{1}{\mu_n(\psi)} \left\langle g, Z_{n,\psi}^k \right\rangle_{L^2(\mathrm{SO(3)})} \\ &= \frac{1}{\mu_n(\psi)^2} \sum_{j=-n}^{n} \widetilde{P}_n^j(0)\, s_j(\psi) \left\langle g, D_n^{j,k} \right\rangle_{L^2(\mathrm{SO(3)})}.\end{aligned}$$

The assertion follows by (5) and (21). ■

Remark 4.8 A big advantage of using the singular value decomposition for inversion is that it is straightforward to apply Tikhonov-type regularization or the mollifier method [27], which both correspond to a multiplication of the summands in the inversion formula (26) with some filter coefficients c_n, cf. [22]. We obtain

$$f_c = \sum_{n=0}^{\infty} \sum_{k=-n}^{n} c_n \frac{\sum_{j=-n}^{n} \widetilde{P}_n^j(0)\, s_j(\psi)\, \hat{g}_n^{j,k}}{\sum_{j=-n}^{n} \widetilde{P}_n^j(0)^2\, s_j(\psi)^2}\, Y_n^k. \tag{27}$$

Filter coefficients corresponding to Pinsker estimators are optimal for functions in certain Sobolev spaces, cf. [9]. They were applied to the Funk–Radon transform in [22]. □

5 Numerical Tests

We consider the arc transform $\mathscr{A}_\psi$ with fixed length $\psi \in (0, \pi)$ as in Sect. 4.3.

5.1 Forward Algorithm

For given $f \in C(\mathbb{S}^2)$, we want to compute the arc transform $\mathscr{A}_\psi f(Q_m)$ at points $Q_m \in \mathrm{SO}(3)$, $m = 1, \ldots, M$. In order to derive an algorithm, we assume that $f \in \mathscr{P}_N(\mathbb{S}^2)$ is a polynomial. We compute the spherical Fourier coefficients

$$\hat{f}_n^k = \int_{\mathbb{S}^2} f(\boldsymbol{\xi})\, \overline{Y_n^k(\boldsymbol{\xi})}\, \mathrm{d}\boldsymbol{\xi}, \qquad n = 0, \ldots, N,\ k = -n, \ldots, n,$$

with a quadrature rule on $\mathbb{S}^2$ that is exact for polynomials of degree $2N$. The computation of the spherical Fourier coefficients $\hat{f}_n^k$ can be done with the adjoint NFSFT (Nonequispaced Fast Spherical Fourier Transform) algorithm [24] in $\mathscr{O}(N^2 \log^2 N^2 + M)$ steps. Then, by (8),

$$\mathscr{A}_\psi f(Q_m) = \sum_{n=0}^{N} \sum_{j,k=-n}^{n} \hat{f}_n^k\, \widetilde{P}_n^j(0)\, s_j(\psi)\, D_n^{j,k}(Q), \qquad m = 1, \ldots, M, \tag{28}$$

is a discrete rotational Fourier transform of degree N, which can be computed with the NFSOFT (Nonequispaced Fast SO(3) Fourier Transform) algorithm [32] in $\mathscr{O}(M + N^3 \log^2 N)$ steps. Implementations of both NFSFT and NFSOFT are contained in the NFFT library [25].

A simple alternative for the computation of $\mathscr{A}_\psi f$ is the following quadrature with K equidistant nodes

$$\mathscr{A}_\psi f(Q) \approx \frac{2\psi}{K} \sum_{i=1}^{K} f\left(Q^{-1}\left(\boldsymbol{e}_{\rho_i}\right)\right), \qquad \rho_i = \frac{2i - 1 - K}{K}\, \psi. \tag{29}$$

Computing $\mathscr{A}_\psi f(Q_m)$ for $m = 1, \ldots, M$ with the quadrature rule (29) requires $\mathscr{O}(KM)$ operations. Hence, for a high number M of evaluation nodes, the NFSOFT-based algorithm is faster than the quadrature based on (29).

5.2 Inversion

We test the inversion from Theorem 4.7. Let $g = \mathscr{A}_\psi f$. For the computation of the rotational Fourier coefficients $\hat{g}_n^{j,k}$, $n = 0, \ldots, N$, $j, k = -n, \ldots, n$, we use a quadrature formula

$$\hat{g}_n^{j,k} = \int_{\mathrm{SO}(3)} g(Q)\, \overline{D_n^{j,k}(Q)}\, \mathrm{d}Q \approx \sum_{m=1}^{M} w_m\, g(Q_m)\, \overline{D_n^{j,k}(Q_m)} \tag{30}$$

with nodes $Q_m \in \mathrm{SO}(3)$ and weights $w_m > 0, m = 1, \ldots, M$. Again, we assume that $f \in \mathscr{P}_N(\mathbb{S}^2)$, which implies $\hat{g}_n^{j,k} = 0$ for $n > N$. Hence, (30) holds with equality if the quadrature integrates rotational harmonics up to degree $2N$ exactly. There are different ways to obtain such exact quadrature formulas on SO(3). In a tensor product approach, we use Gauss–Legendre quadrature in $\cos\beta$ and a trapezoidal rule in both α and γ. We can also write $\mathrm{SO}(3) \sim \mathbb{S}^1 \times \mathbb{S}^2$ and pair a trapezoidal rule on $\mathbb{S}^1$ in α with a quadrature on $\mathbb{S}^2$ with azimuth γ and polar angle β, see [17]. Furthermore, Gauss-type quadratures on SO(3) that are exact up to machine precision were computed in [16]. In Fig. 3, one can see the circle arcs corresponding to different quadrature rules on SO(3), namely Gauss–Legendre nodes (Fig. 3a), the tensor product of $\mathbb{S}^1 \times \mathbb{S}^2$ (Fig. 3b) and a Gauss-type quadrature on SO(3) (Fig. 3c). We used Gauss-type quadratures on both $\mathbb{S}^2$ and SO(3) from [15]. Note that because of the symmetry of the Gauss-type quadrature on SO(3) we used in Fig. 3c, every arc corresponds to two quadrature nodes on SO(3).

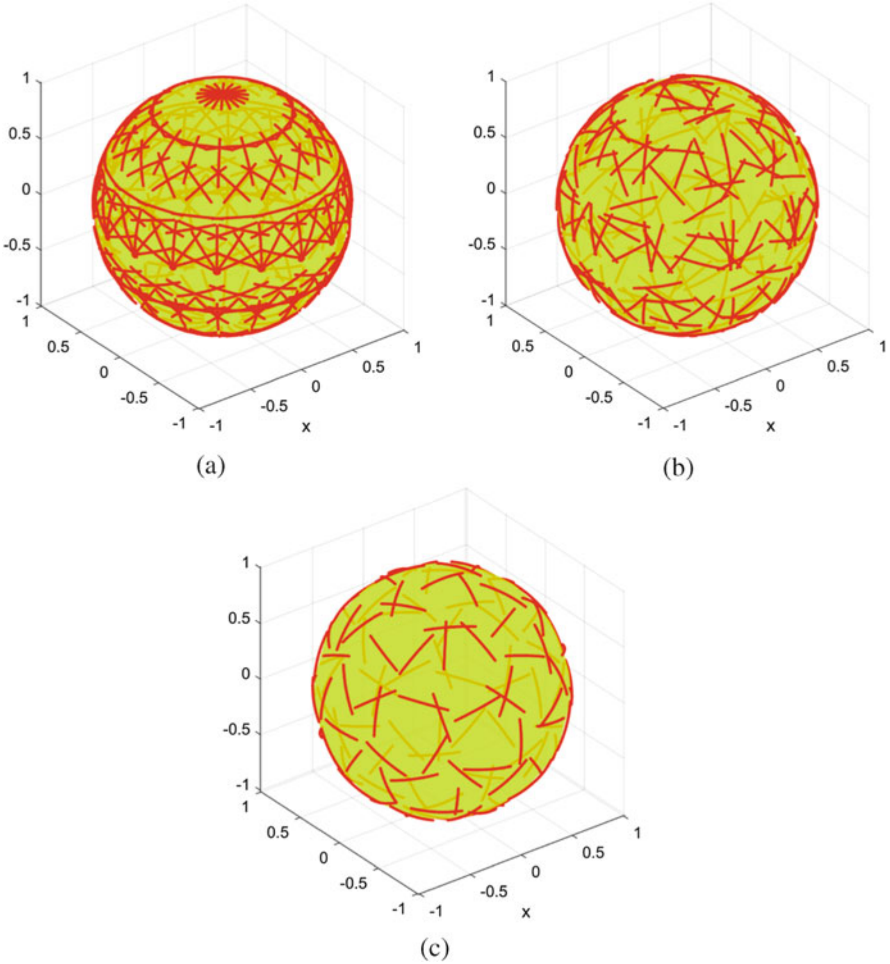
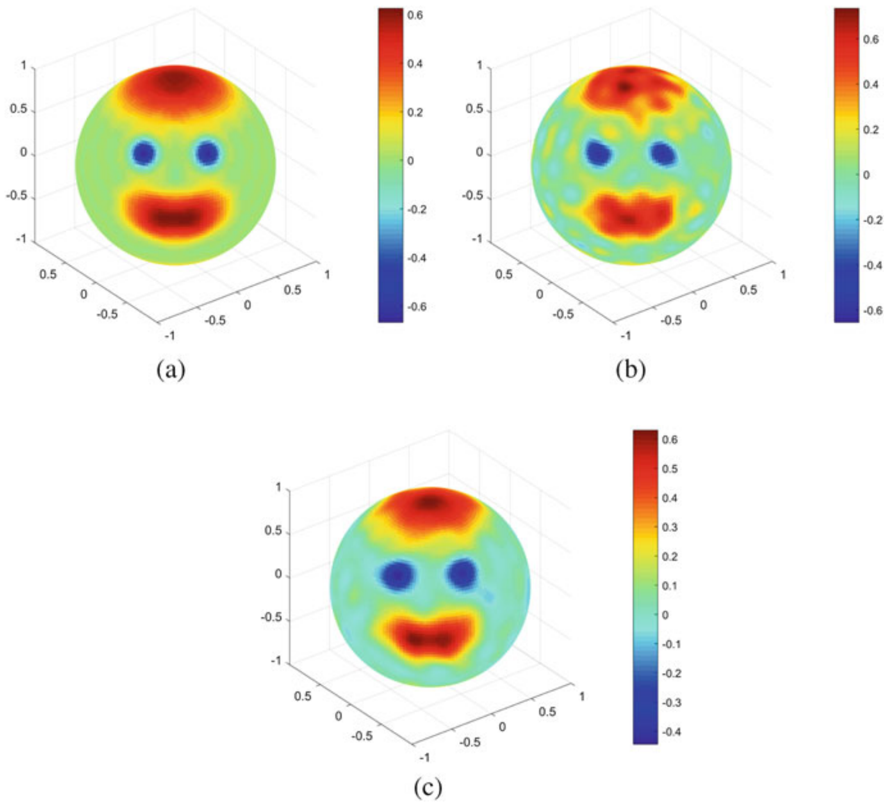


Fig. 3 Circle arcs $\gamma(Q_m, 0.2)$ corresponding to quadrature nodes $Q_m \in \mathrm{SO}(3)$, all quadrature formulas are exact for all rotational harmonics $D_n^{j,k}$ of degree $n \leq 8$. (**a**) Gauss–Legendre with 405 nodes. (**b**) Tensor product $\mathbb{S}^1 \times \mathbb{S}^2$ with 252 nodes. (**c**) Gauss-type with 240 nodes



(a)

(b)

(c)

Fig. 4 Reconstruction of a spherical test function f for degree $N = 22$, $\psi = 0.7$ and a tensor product $\mathbb{S}^1 \times \mathbb{S}^2$ quadrature ($M = 30{,}240$). (**a**) Exact data. (**b**) Noisy data. (**c**) Noisy data & regularization

The reconstruction formula (26) becomes the discrete rotational Fourier transform

$$f = \sum_{n=0}^{N} \sum_{k=-n}^{n} \frac{\sum_{j=-n}^{n} \widetilde{P}_n^j(0)\, s_j(\psi)\, \hat{g}_n^{j,k}}{\sum_{j=-n}^{n} \widetilde{P}_n^j(0)^2\, s_j(\psi)^2}\, Y_n^k.$$

In Fig. 4, we compare the reconstruction results, where we use an artificial test function, the parameter $N = 22$ and the tensor product of a trapezoidal rule on $\mathbb{S}^1$ with a Gauss-type quadrature on $\mathbb{S}^2$ from [15]. The resulting SO(3) quadrature uses $M = 30{,}240$ nodes and is exact for degree 44. We first perform the inversion without any noise in the data. The reconstruction has an RMSE (root mean square error) of 0.0338. Then we add Gaussian white noise with a standard deviation of 0.2 to the data $\mathscr{A}_\psi f(Q_m)$ and achieve an RMSE of 0.2272. Even though we did not perform any regularization, the reconstruction from noisy data still looks considerably well. This might be explained by the fact that the inverse arc transform with fixed opening angle and full SO(3) data is still an overdetermined problem. Applying the regularization (27) truncated to degree $n \leq N$ with filter coefficients from [22] yields a smaller RMSE of 0.1393.

Acknowledgements The authors thank Volker Michel for pointing out the problem of spherical surface wave tomography at the Mecklenburg Workshop on Approximation Methods and Data Analysis 2016 and for fruitful conversations later on. Furthermore, we thank the anonymous reviewer for providing helpful comments and suggestions to improve this article.

References

1. A. Abouelaz, R. Daher, Sur la transformation de Radon de la sphère S^d. Bull. Soc. Math. France **121**(3), 353–382 (1993)
2. A. Amirbekyan, The application of reproducing kernel based spline approximation to seismic surface and body wave tomography: theoretical aspects and numerical results. Dissertation, Technische Universität Kaiserslautern, 2007
3. A. Amirbekyan, V. Michel, F.J. Simons, Parametrizing surface wave tomographic models with harmonic spherical splines. Geophys. J. Int. **174**(2), 617–628 (2008)
4. F.L. Bauer, Remarks on Stirling's formula and on approximations for the double factorial. Math. Intell. **29**(2), 10–14 (2007)
5. R. Daher, Un théorème de support pour une transformation de Radon sur la sphère S^d. C. R. Acad. Sci. Paris **332**(9), 795–798 (2001)
6. F. Dahlen, J. Tromp, *Theoretical Global Seismology* (Princeton University Press, Princeton, 1998)
7. F. Dai, Y. Xu, *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics (Springer, New York, 2013)
8. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*. Mathematics and Its Applications, vol. 375 (Kluwer Academic, Dordrecht, 1996)
9. D. Fournier, L. Gizon, M. Holzke, T. Hohage, Pinsker estimators for local helioseismology: inversion of travel times for mass-conserving flows. Inverse Probl. **32**(10), 105002 (2016)
10. W. Freeden, T. Gervens, M. Schreiner, *Constructive Approximation on the Sphere* (Oxford University Press, Oxford, 1998)
11. P. Funk, Über Flächen mit lauter geschlossenen geodätischen Linien. Math. Ann. **74**(2), 278–300 (1913)
12. S. Gindikin, J. Reeds, L. Shepp, Spherical tomography and spherical integral geometry, in *Tomography, Impedance Imaging, and Integral Geometry*, ed. by E.T. Quinto, M. Cheney, P. Kuchment. Lectures in Applied Mathematics, vol. 30 (American Mathematical Society, South Hadley, MA, 1994), pp. 83–92
13. P. Goodey, W. Weil, Average section functions for star-shaped sets. Adv. Appl. Math. **36**(1), 70–84 (2006)
14. I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products*, 7th edn. (Academic, New York, 2007)
15. M. Gräf, Quadrature rules on manifolds, 2016, http://www.tu-chemnitz.de/~potts/workgroup/graef/quadrature
16. M. Gräf, Efficient algorithms for the computation of optimal quadrature points on Riemannian manifolds. Dissertation, Universitätsverlag Chemnitz, 2013
17. M. Gräf, D. Potts, Sampling sets and quadrature formulae on the rotation group. Numer. Funct. Anal. Optim. **30**, 665–688 (2009)

18. H. Groemer, On a spherical integral transformation and sections of star bodies. Monatsh. Math. **126**(2), 117–124 (1998)
19. D.M. Healy Jr., H. Hendriks, P.T. Kim, Spherical deconvolution. J. Multivariate Anal. **67**, 1–22 (1998)
20. S. Helgason, *Integral Geometry and Radon Transforms* (Springer, Berlin, 2011)
21. R. Hielscher, The Radon transform on the rotation group–inversion and application to texture analysis. Dissertation, Technische Universität Bergakademie Freiberg, 2007
22. R. Hielscher, M. Quellmalz, Optimal mollifiers for spherical deconvolution. Inverse Probl. **31**(8), 085001 (2015)
23. R. Hielscher, M. Quellmalz, Reconstructing a function on the sphere from its means along vertical slices. Inverse Probl. Imaging **10**(3), 711–739 (2016)
24. J. Keiner, D. Potts, Fast evaluation of quadrature formulae on the sphere. Math. Comput. **77**, 397–419 (2008)
25. J. Keiner, S. Kunis, D. Potts, NFFT 3.4, C subroutine library, 2017, http://www.tu-chemnitz.de/~potts/nfft
26. J. Keiner, S. Kunis, D. Potts, Efficient reconstruction of functions on the sphere from scattered data. J. Fourier Anal. Appl. **13**, 435–458 (2007)
27. A.K. Louis, P. Maass, A mollifier method for linear operator equations of the first kind. Inverse Probl. **6**(3), 427–440 (1990)
28. V. Michel, *Lectures on Constructive Approximation: Fourier, Spline, and Wavelet Methods on the Real Line, the Sphere, and the Ball* (Birkhäuser, New York, 2013)
29. G. Nolet, *A Breviary of Seismic Tomography* (Cambridge University Press, Cambridge, 2008)
30. V.P. Palamodov, *Reconstruction from Integral Data*. Monographs and Research Notes in Mathematics (CRC Press, Boca Raton, 2016)
31. V.P. Palamodov, Reconstruction from cone integral transforms. Inverse Probl. **33**(10), 104001 (2017)
32. D. Potts, J. Prestin, A. Vollrath, A fast algorithm for nonequispaced Fourier transforms on the rotation group. Numer. Algorithms **52**, 355–384 (2009)
33. M. Quellmalz, A generalization of the Funk–Radon transform. Inverse Probl. **33**(3), 035016 (2017)
34. B. Rubin, Generalized Minkowski–Funk transforms and small denominators on the sphere. Fract. Calc. Appl. Anal. **3**(2), 177–203 (2000)
35. B. Rubin, Radon transforms and Gegenbauer–Chebyshev integrals, II; examples. Anal. Math. Phys. **7**(4), 349–375 (2017)
36. B. Rubin, On the determination of star bodies from their half-sections. Mathematika **63**(2), 462–468 (2017)
37. Y. Salman, An inversion formula for the spherical transform in S^2 for a special family of circles of integration. Anal. Math. Phys. **6**(1), 43–58 (2016)
38. R. Schneider, Functions on a sphere with vanishing integrals over certain subspheres. J. Math. Anal. Appl. **26**, 381–384 (1969)
39. R.S. Strichartz, L^p estimates for Radon transforms in Euclidean and non–Euclidean spaces. Duke Math. J. **48**(4), 699–727 (1981)
40. J. Trampert, J.H. Woodhouse, Global phase velocity maps of Love and Rayleigh waves between 40 and 150 seconds. Geophys. J. Int. **122**(2), 675–690 (1995)
41. D. Varshalovich, A. Moskalev, V. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific Publishing, Singapore, 1988)
42. E. Wigner, *Gruppentheorie und ihre Anwendung auf die Quantenmechanik der Atomspektren*. Die Wissenschaft, vol. 85 (Friedr. Vieweg & Sohn, Braunschweig, 1931)
43. J.H. Woodhouse, A.M. Dziewonski, Mapping the upper mantle: three-dimensional modeling of earth structure by inversion of seismic waveforms. J. Geophys. Res. Solid Earth **89**(B7), 5953–5986 (1984)
44. G. Zangerl, O. Scherzer, Exact reconstruction in photoacoustic tomography with circular integrating detectors II: spherical geometry. Math. Methods Appl. Sci. **33**(15), 1771–1782 (2010)

Numerical Studies of Recovery Chances for a Simplified EIT Problem

Christopher Hofmann, Bernd Hofmann, and Roman Unger

Abstract This study investigates a simplified discretized EIT model with eight electrodes distributed equally spaced at the boundary of a disc covered with a small number of material 'stripes' of varying conductivity. The goal of this paper is to evaluate the chances of identifying the conductivity values of each stripe from rotating measurements of potential differences. This setting comes from an engineering background, where the used EIT model is exploited for the detection of conductivities in carbon nanotubes (CNT) and carbon nanofibers (CNF). Connections between electrical conductivity and mechanical strain have been of major interest within the engineering community and has motivated the investigation of such a 'stripe' structure. Up to five conductivity values can be recovered from noisy 8×8 data matrices in a stable manner by a least squares approach. Hence, this is a version of regularization by discretization and additional tools for stabilizing the recovery seem to be superfluous. To our astonishment, no local minima of the squared misfit functional were observed, which seems to indicate uniqueness of the recovery if the number of stripes is quite small.

1 Introduction

Electrical impedance tomography (EIT) is an imaging technology that aims to reconstruct the internal electric conductivity of a given object through electrostatic measurements obtained on its boundary. Previously, this class of inverse problems has been studied with a focus on applications in medical imaging and geology. The problem was first posed in a mathematical way by Calderón in [3]. Conductivity distributions appearing in medical applications can be considered as piecewise constant functions under many circumstances. Various body tissues have conductivities which differ sometimes substantially. Therefore the conductivity can be assumed to

C. Hofmann (✉) · B. Hofmann · R. Unger
Faculty of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany
e-mail: christopher.hofmann@mathematik.tu-chemnitz.de;
bernd.hofmann@mathematik.tu-chemnitz.de; roman.unger@mathematik.tu-chemnitz.de

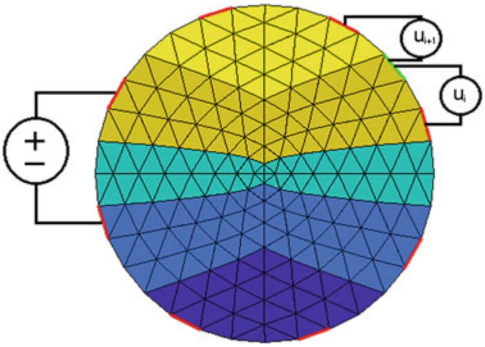
B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_8

have jumps at organ borders. One might also be interested in identifying the size and position of an object, whose conductivity is considerably different from the surrounding tissue, e.g. an organ within a thorax. Numerous results with focus on such applications have been published in recent years (see, e.g., [5, 8, 11, 18]).

In contrast, the studies presented in this paper were motivated by an engineering background. Precisely, the technological goal for the used EIT model is the detection of damages in carbon nanotubes (CNT) and carbon nanofibers (CNF). Connections between electrical conductivity and mechanical strain have been of major interest for engineers in recent years (see, e.g., [4, 17, 21, 31]). In this context, numerous results have been published, preferably with focus on the detection of inclusions or objects within the structure, in this case the carbon nanotube (see, e.g., [10, 12–16, 30]). To achieve satisfying assertions, these methods partly rely on a priori information on the specimen. Usually a known background conductivity and a substantially different conductivity of the inclusion are supposed. In some cases, results had been presented without disclosing the underlying recovery method and algorithm used and moreover the mathematical model is not documented in detail.

The objective of the following model and of corresponding simulations based on it is to evaluate chances and limitations for the recovery of the mechanical strain inside the CNT, which is caused by bending the specimen. For this purpose, this study investigates a simplified discretized EIT model with eight electrodes distributed equally spaced at the boundary $\partial\Omega$ of a disk Ω modelled in two dimensions and covered with a small number n of material 'stripes' of varying conductivity, see a schematic shape in Fig. 1. Each of the 'stripes' is assumed to possess a constant conductivity σ_i $(i = 1, 2, \ldots, n)$, but no assumption on inclusions or background conductivity is made. Results of case studies are presented, in particular, for $n = 2$ and $n = 5$. As is well-known, the EIT-recovery of a full locally distributed conductivity function $\sigma(x)$, $x \in \Omega$, represents a severely ill-posed nonlinear inverse problem, and for example Tikhonov regularization can be helpful for finding stable approximate solutions. But in this study we have a situation of 'regularization by discretization' due to the small number n of unknowns occurring here and thus additional tools for stabilizing the recovery process seem to be superfluous.

Fig. 1 Specimen with electrodes and finite element grid



2 The General EIT Model

For a general two-dimensional conducting object Ω with smooth boundary $\partial\Omega$ and conductivity function $\sigma(x),\ x \in \Omega$, the usual elliptic partial differential equation

$$\nabla \cdot (\sigma(x)\nabla u(x)) = 0 \tag{1}$$

applies, where the state variable $u(x),\ x \in \Omega$, denotes the electric potential and the σ-weighted outer normal derivative $\sigma\partial_\nu u\,|_{\partial\Omega}$ can be interpreted as current. For practical applications it is desirable to apply current and to measure voltages in the sense of potential differences rather than vice versa. We follow this route and consider the current-to-voltage map

$$\Lambda_\sigma : L^2_\diamond(\partial\Omega) \to L^2_\diamond(\partial\Omega), \qquad g|_{\partial\Omega} \mapsto u^g|_{\partial\Omega}\,, \tag{2}$$

where u^g denotes the weak solution of (1) with Neumann boundary values

$$\sigma\partial_\nu u\,|_{\partial\Omega} = g\,|_{\partial\Omega}\ .$$

In this context, we introduce the subspaces

$$L^\infty_+(\Omega) := \{\sigma \in L^\infty(\Omega) : \inf_{x\in\Omega} \sigma(x) > 0\}$$

and

$$L^2_\diamond(\partial\Omega) := \{g \in L^2(\partial\Omega) : \int_{\partial\Omega} g ds = 0\}\,.$$

For fixed $\sigma \in L^\infty_+$ the operator Λ_σ is a compact and self-adjoint linear operator mapping in $L^2_\diamond(\partial\Omega)$ (cf., e.g., [9]). The forward operator of this model situation is then given by

$$F : L^\infty_+(\Omega) \to \mathscr{L}(L^2_\diamond(\partial\Omega)), \qquad \sigma \mapsto \Lambda_\sigma. \tag{3}$$

Consequently, the inverse problem is to retrieve the function $\sigma(x),\ x \in \Omega$, from data of the current-to-voltage map Λ_σ. Various results on the uniqueness of this inverse problem for full and partial data have been published, and we refer to [29] and moreover also to [2, 19, 20, 26].

3 A Simplified and Discretized Specific EIT Model

In practice it is obviously impossible to obtain measurements on the whole boundary $\partial\Omega$ of Ω. Therefore the choice of electrode model is crucial in any numerical study. Widely used electrode models include the Gap Model, Shunt Model and Complete Electrode Model, and we refer for details to the monograph [28] and the handbook article [1]. Electrode models have been extensively studied following a more practical approach in [22, 27] and more recently in [6, 7]. Our investigations below will use the Shunt Model, and its discretized version is outlined in the following. For numerical simulations concerning practical applications the problem has to be discretized with K electrodes ϵ_k, $\bigcup_{k=1}^{K} \epsilon_k \subset \partial\Omega$, on which potential measurements are taken and current is injected. In this context, I_k and U_k denote the associated values of current and voltage, respectively, on the k-th electrode. We further assume steady state $\sum_{k=1}^{K} I_k = 0$ (in- and outgoing currents add up to zero). As the solution of (1) is not unique, it is assumed that the potentials add up to zero as well. With $\mathbb{R}^K_\diamond = \{x \in \mathbb{R}^K \,:\, \sum_{k=1}^{K} x_k = 0\}$ the mapping

$$R_\sigma : \quad (I_k)_{k=1}^{K} \in \mathbb{R}^K_\diamond \quad \mapsto \quad (U_k)_{k=1}^{K} \in \mathbb{R}^K_\diamond$$

is then the basis for required sets of measurements.

As the Shunt Model is used in the following case studies, we assume that no current flows outside the electrodes, i.e. $\sigma\partial_\nu u|_{\partial\Omega\setminus\bigcup_{k=1}^{K}\epsilon_k} = 0$, and that the current on electrode ϵ_k is equally distributed with overall current $I_k = \int_{\epsilon_k} \sigma\partial_\nu u|_{\partial\Omega} ds$. Therefore we have, $\sigma\partial_\nu u|_{\epsilon_k} = \frac{I_k}{|\epsilon_k|}$ with arclength $|\epsilon_k|$ of the electrode ϵ_k. It is further assumed that the potential on every electrode is constant, i.e. $u\,|_{\epsilon_k} = const$. Under these conditions the underlying elliptic boundary value problem is discretized using a FEM code (see for details Sect. 4.1).

With the two-dimensional conducting object Ω in disc form in mind, we concretize the model as follows: We assume that the conductivity is isotropic and we discretize the geometry by using a triangular mesh with 32 boundary edges and $K = 8$ electrodes ϵ_i $(i = 1, .., 8)$ for taking voltage measurements. Neumann boundary conditions are then set on two neighbouring (although not adjacent) electrodes as $\sigma\partial_\nu u_{|\epsilon_i} = 1$ and $\sigma\partial_\nu u_{|\epsilon_{i+1}} = -1$. Moreover, we assume steady state and, in order to overcome non-uniqueness, $u(x) = 0$ for one arbitrary chosen boundary edge which is not an electrode. In its discretized form the forward operator (3) is a mapping

$$\underline{\sigma} = (\sigma_1, \ldots, \sigma_n)^T \in \mathbb{R}^n \quad \mapsto \quad F(\underline{\sigma}) \in \mathbb{R}^{8\times 8},$$

where n denotes the number of 'stripes' inside the disc Ω. Note that the shape (geometry) of the stripes is apparently assumed to be known. We are only searching for the conductivity values σ_i $(i = 1, \ldots, n)$, which are constant on each of the stripes. Moreover, the matrix $F(\underline{\sigma})$ characterizes the noise-free image. To receive the 8×8-matrix $F(\underline{\sigma})$ the electrodes are rotated and the associated elliptic problem is solved in a repeated manner until the starting position is reached. We note that the forward operator $F : \mathbb{R}^n \to \mathbb{R}^{8\times 8}$ is nonlinear such that we have a nonlinear inverse problem under consideration even in this simplified form.

Let us assume that $\underline{\sigma}^* \in \mathbb{R}^n_+$ is the 'true conductivity vector' to be identified. For sufficiently small numbers n it makes sense to compute approximate solutions by a least squares approach if the data are noisy. Hence, we search for approximate solutions

$$\underline{\sigma}^\delta_{LS} = \underset{\underline{\sigma}\in Q}{\arg\min} \, \| F(\underline{\sigma}) - F^\delta(\underline{\sigma}^*) \|_F \,. \tag{4}$$

In this case, $Q \subset \mathbb{R}^n_+$ is the set of admissible solution vectors, for example obtained by imposing box constraints, $\| \cdot \|_F$ designates the Frobenius norm, and the matrix $F^\delta(\underline{\sigma}^*)$ indicates the noisy data associated with some noise level $\delta > 0$.

For the subsequent case studies we carry out simulations, where the exact matrix $F(\underline{\sigma}^*)$ is perturbed in an additive way

$$F^\delta(\underline{\sigma}^*) = F(\underline{\sigma}^*) + \mathscr{E}$$

by means of a matrix $\mathscr{E} = (\varepsilon_{ij}) \in \mathbb{R}^{8\times 8}$ containing Gaussian random i.i.d. entries $\epsilon_{ij} \sim \mathscr{N}(0, d^2)$. For a prescribed averaged relative data error $\delta > 0$ defined by the expectation value

$$\mathbf{E}\left[\frac{\| F^\delta(\underline{\sigma}^*) - F(\underline{\sigma}^*) \|_F^2}{\| F(\underline{\sigma}^*) \|_F^2} \right] = \delta^2, \tag{5}$$

we have to use $d = \frac{\delta}{8} \| F(\underline{\sigma}^*) \|_F$ as standard deviation of the entries in $\mathscr{E}$ for the numerical experiments, since $\mathbf{E}\big[\| F^\delta(\underline{\sigma}^*) - F(\underline{\sigma}^*) \|_F^2 \big] = \mathbf{E}\big[\sum_{i,j=1}^{8} \epsilon_{ij}^2 \big] = 64 d^2$.

4 Numerical Case Studies

4.1 Remarks on Used Finite Element Implementation

To execute the numerical experiments in this case study, a fast finite element solver for the forward operator (3) in its discretized form were needed. Specifically, we have applied an updated 2D Kernel **SPC-PM2Ad** version of an already existing finite element code **SPC**, which has originally been developed in the context of the DFG-funded Collaborative Research Center *SFB 393: Parallel Numerical Simulation for Physics and Mechanics of Continua*. For detailed descriptions of the structure and features of the FEM code we refer to [23–25]. The finite element code is written in FORTRAN and can solve Eq. (1) for the required boundary conditions

Table 1 List of stable error-to-noise ratios

Mean value noise level δ	Random noise level $\frac{\| F(\underline{\sigma}^*)-F^\delta(\underline{\sigma}^*)\|_F}{\| F(\underline{\sigma}^*)\|_F}$	Reconstruction error $\frac{\|\underline{\sigma}_{opt}-\underline{\sigma}^*\|_2}{\|\underline{\sigma}^*\|_2}$	Mean error-to-noise ratio $\frac{\|\underline{\sigma}_{opt}-\underline{\sigma}^*\|_2}{\|\underline{\sigma}^*\|_2}/\delta$	Random error-to-noise ratio $\frac{\|\underline{\sigma}_{opt}-\underline{\sigma}^*\|_2}{\|\underline{\sigma}^*\|_2}/\frac{\| F(\underline{\sigma}^*)-F^\delta(\underline{\sigma}^*)\|_F}{\| F(\underline{\sigma}^*)\|_F}$
0.0100	0.0181	0.0029	0.2925	0.1876
0.0250	0.0361	0.0045	0.1789	0.1403
0.0500	0.0662	0.0076	0.1516	0.1236
0.0550	0.0722	0.0082	0.1488	0.1222
0.1000	0.1199	0.0142	0.1421	0.1181
0.1500	0.1791	0.0213	0.1421	0.1181
0.2000	0.2372	0.0285	0.1425	0.1184
0.2500	0.2935	0.0360	0.1441	0.1198

exploiting appropriate error estimations and adaptive mesh refinement with high accuracy in very short computing time.

For series computation it has been called from a C++ OpenMPI implementation which runs parallel on a distributed memory multicore cluster. Parallelization and the already fast computing times of the FEM code have been essential for preparing Sect. 4.3. For the presented case with five unknowns, $8 \times 31^5 \approx 229$ million of finite element simulations were necessary to calculate values for the forward operator (3) on the whole grid.

In order to present the results of Table 1, the **SPC-PM2Ad** Kernel has been wrapped inside a MATLAB-minimizer based on a specific version of the Levenberg-Marquardt algorithm for solving the nonlinear least squares problem (4).

4.2 The Case of Two Unknown Conductivities

We start our numerical case studies with the investigation of a disc Ω covered by a 'stripe' structure (see Fig. 2) of $n = 2$ materials with different conductivities σ_1 and σ_2. For the set of admissible pairs of values we use the rectangle

$$Q = \{(\sigma_1, \sigma_2) \in [10, 75] \times [5, 46]\},$$

and for applying the discretized forward operator we calculate the corresponding matrices $F(\underline{\sigma}) \in \mathbb{R}^{8\times 8}$ for grids with 51×51 support points.

As an illustrative example we plot the discrepancy norm $\| F(\underline{\sigma}) - F(\underline{\sigma}^*)\|_F$ depending on $\underline{\sigma} = (\sigma_1, \sigma_2)^T$ for $\underline{\sigma}^* = (37.7\ ,\ 7.9)^T$ in Fig. 3 and the corresponding level sets in Fig. 4. One easily sees here and for numerous other examples of two-dimensional points $\underline{\sigma}^*$ that the level sets are of elliptical shape. If, as in our example, $\sigma_2 \ll \sigma_1$, the ellipses are elongated parallel to the axis σ_1-axis. Then the smaller parameter (here σ_2) with lower conductivity can be recovered in a more precise

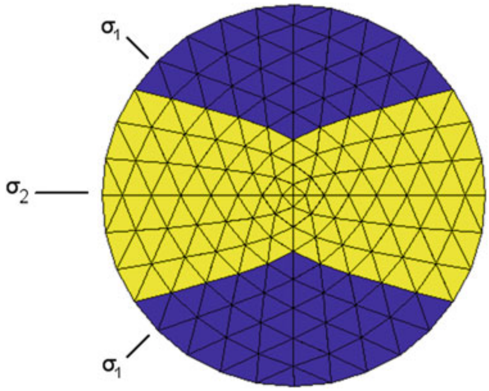


Fig. 2 Material 'stripes' with two unknown conductivities

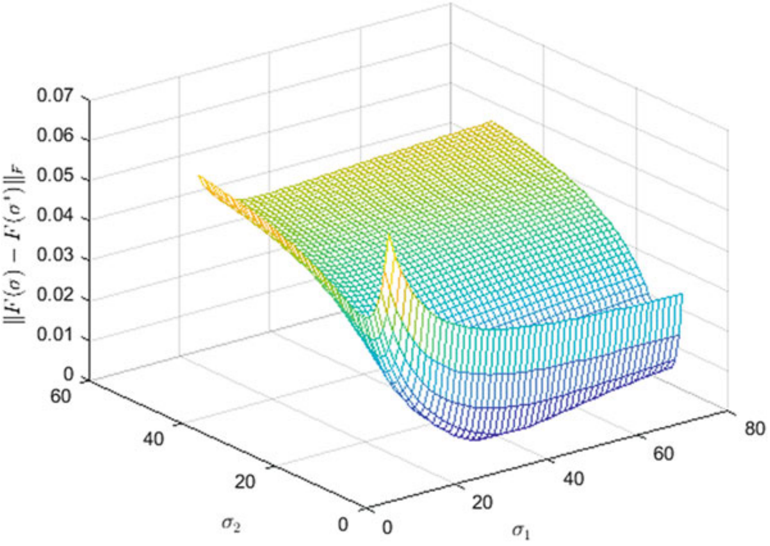


Fig. 3 Perspective drawing of discrepancy norm $\| F(\underline{\sigma}) - F(\underline{\sigma}^*)\|_F$ depending on $\underline{\sigma} = (\sigma_1, \sigma_2)^T$

manner than the parameter with higher conductivity. This observation remains true if the data are noisy. If we have $\sigma_1 \approx \sigma_2$, then the level sets tend to be concentric circles.

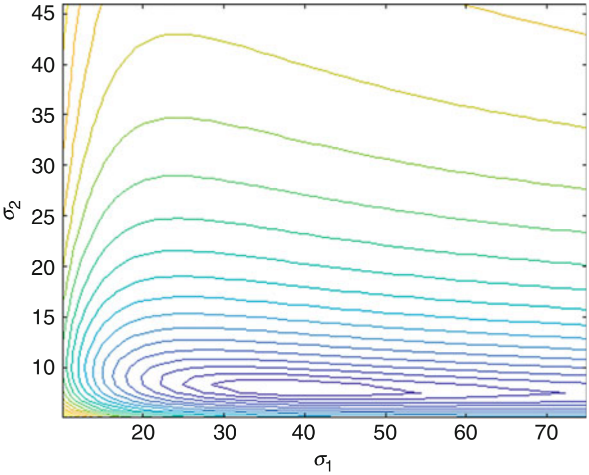


Fig. 4 Level sets $L_c = \{\underline{\sigma} = (\sigma_1, \sigma_2)^T : \| F(\underline{\sigma}) - F(\underline{\sigma}^*)\|_F = c\}$

4.3 The Case of Five Unknown Conductivities

In a more detailed second numerical experiment, we consider ‘stripes’ on the disc Ω with $n = 5$ different materials, where the conductivities σ_1 to σ_5 are arranged from the bottom to the top. Since the finite element calculations tend to be more costly and time consuming, the matrices $F(\underline{\sigma})$ have been calculated for every σ_i $(i = 1, 2, 3, 4, 5)$ approximately in the interval $[1, 50]$ with only 31 support points in every component.

The numerical case study shows that very different conductivity distributions may lead to nearly the same image of the forward operator. An example is presented by Fig. 5.

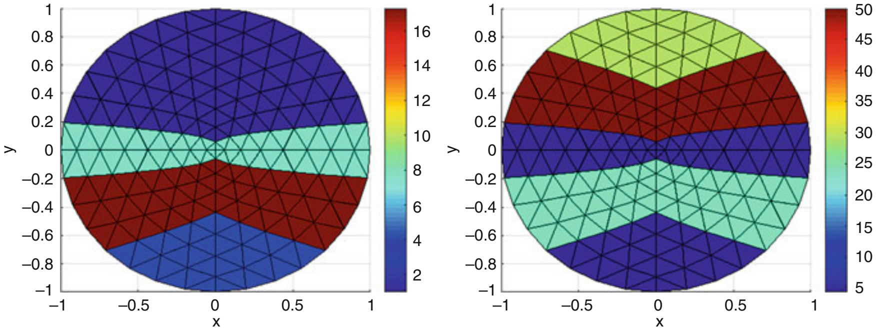


Fig. 5 Two very different conductivity distributions with small discrepancy norm, left picture: $\underline{\sigma}^{(1)} = (4.26, 17.33, 7.65, 0.99, 1.00)^T$, right picture: $\underline{\sigma}^{(2)} = (4.27, 23.87, 4.34, 50.00, 28.99)^T$, $\| F(\underline{\sigma}^{(1)}) - F(\underline{\sigma}^{(2)})\|_F^2 = 0.000099$

On the other hand, Fig. 6 delivers plots of the function

$$f(\lambda) := \| F(\underline{\sigma}^* + \lambda(\underline{\sigma}^{(3)} - \underline{\sigma}^*)) - F^\delta(\underline{\sigma}^*)\|_F, \qquad \lambda \in [-5, 1],$$

characterizing a straight line through the points $\underline{\sigma}^* = (7.53, 22.23, 14.28, 4.26, 4.99)^T$ and $\underline{\sigma}^{(3)} = (7.53, 45.09, 12.63, 4.26, 4.99)^T$, where in both points the components σ_1, σ_4 and σ_5 coincide.

More insight into such two-dimensional cross sections of the five-dimensional space give Figs. 7 and 8. Figure 7 shows the level sets of $\| F(\underline{\sigma}) - F(\underline{\sigma}^*)\|_F$, for $\underline{\sigma}^* = (7.53, 22.23, 14.28, 4.26, 4.99)^T$ with respect to the second and to the third

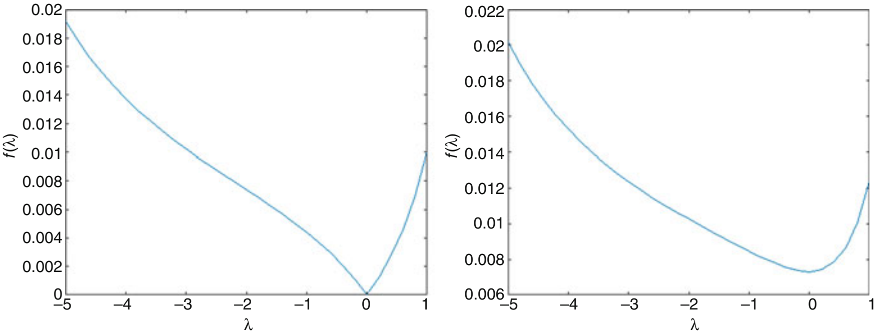


Fig. 6 Graph of $f(\lambda) := \| F(\underline{\sigma}^* + \lambda(\underline{\sigma}^{(3)} - \underline{\sigma}^*)) - F^\delta(\underline{\sigma}^*)\|_F$ for $\lambda \in [-5, 1]$ without noise ($\delta = 0$, left picture) and with 5% noise ($\delta = 0.05$, right picture)

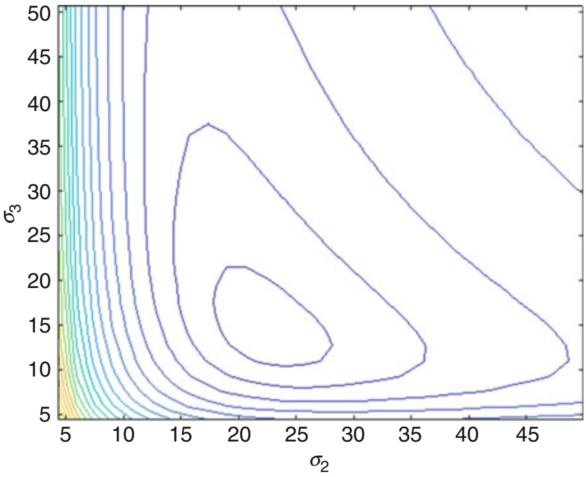


Fig. 7 Level sets $L_c = \{(\sigma_2, \sigma_3) : \| F(\underline{\sigma}) - F^\delta(\underline{\sigma}^*)\|_F = c\}$ without noise ($\delta = 0$) and for fixed σ_1^*, σ_4^* and σ_5^*

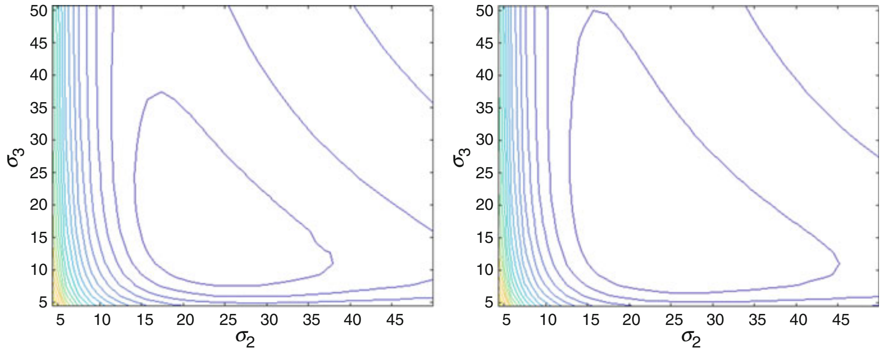


Fig. 8 Level sets L_c as in Fig. 7, but with noise $\delta = 0.05$ (left picture) and $\delta = 0.10$ (right picture)

coordinate. The first, fourth and fifth coordinate are again fixed for this numerical experiment.

If noise is added, i.e. $\delta > 0$, then Fig. 8 shows that near-to-elliptic areas characterized by the sublevel sets $\bigcup_{\tau \leq c} L_\tau$ grow with δ. Hence, the chances for recovering the conductivity distribution with a high level of accuracy decrease with increasing noise level. However, this seems to be the only form of uncertainty if the number n is quite small.

As in the numerical examples used for Fig. 6, where $f(\lambda)$ is strictly decreasing for $\lambda < 0$ and strictly increasing for $\lambda > 0$, and for Figs. 7 and 8, where concentric level sets occur, we did not at all observe inside of boxes $\underline{\sigma} \in Q \subset \mathbb{R}^5$ local minima of the functional $\| F(\underline{\sigma}) - F^\delta(\underline{\sigma}^*) \|_F$ in the five-dimensional case study, even if the noise level $\delta > 0$ is rather large. Overall, numerical evidence obtained from these case studies suggests that the least squares approach (4) has indeed a global minimum, i.e. the minimizer $\underline{\sigma}_{LS}^\delta$ is uniquely determined and no local minima seem to disturb the optimization process when a Levenberg-Marquardt algorithm is applied to find the least squares solution numerically.

Since the Jacobian has to be calculated in every step of the iteration process, which in turn requires multiple calculations of forward operator matrices, we used precalculated values for the discretized F in connection with some kind of multilinear interpolation between the support points. In the following table, $\underline{\sigma}_{opt}$ denotes the optimal solution determined by the algorithm, where the exact conductivity distribution is assumed to be $\underline{\sigma}^* = (37.7, 7.9, 10.7, 18.2, 5.6)^T$ and $\underline{\sigma}_{start} = (9, 32, 7, 1, 37)^T$ has been used as starting vector for the Levenberg-Marquardt iteration. Taking into account the fact that the noise level δ expresses the relative data error in expectation value sense (cf. formula (5)) and that $\frac{\| F(\underline{\sigma}^*) - F^\delta(\underline{\sigma}^*) \|_F}{\| F(\underline{\sigma}^*) \|_F}$ is the random counterpart for one specific realization of the noise matrix $\mathscr{E}$, we can compare the fourth and fifth column of Table 1. The relative reconstruction error in the third column, which uses the Euclidean vector norm $\| \cdot \|_2$, proves the astonishing

Table 2 List of condition numbers of Jacobian $J_h(\underline{\sigma}^*)$ ($h = 0.01$) for varying $\underline{\sigma}^*$

Condition number	σ_1^*	σ_2^*	σ_3^*	σ_4^*	σ_5^*
6.74	5	5	5	5	5
6.76	50	50	50	50	50
14.73	3	4	5	6	7
14.87	30	40	50	60	70
12.55	300	400	500	600	700
4.07	1	30	1	30	1
4.42	10	300	10	300	10
3.99	100	3000	100	3000	100

stability of the recovery process with $n = 5$ unknowns, and we refer in particular to the almost constant quotients in the fifth column even if the noise is up to 25 %.

We complete our investigations by a study of the condition numbers for varying $\underline{\sigma}^*$ of the approximated Jacobian $J_h(\underline{\sigma}^*) \in \mathbb{R}^{64\times 5}$ to $F(\underline{\sigma}^*)$ calculated by finite differences with increment value $h = 0.01$. Let

$$s_1(\underline{\sigma}^*) \geq s_2(\underline{\sigma}^*) \geq s_3(\underline{\sigma}^*) \geq s_4(\underline{\sigma}^*) \geq s_5(\underline{\sigma}^*) > 0$$

denote the singular values of $J_h(\underline{\sigma}^*)$, and

$$\kappa := \frac{s_1(\underline{\sigma}^*)}{s_5(\underline{\sigma}^*)}$$

the corresponding condition number. Some selection of $\underline{\sigma}^*$-situations with associated condition numbers is presented in Table 2.

All results of the table indicate well-conditioning, regardless of whether the five conductivity values σ_i^* $(i = 1, \dots, 5)$ are very different or equal, monotonically increasing or sinusoidal alternating. A proportional growth of all five values does not essentially change the condition numbers.

Acknowledgements The research was supported by Deutsche Forschungsgemeinschaft by DFG grant HO 1454/10-1.

References

1. A. Adler, R. Gaburro, W. Lionheart, Electrical impedance tomography, In *Handbook of Mathematical Methods in Imaging*, ed by O. Scherzer, 2nd edn. (Springer Science+Business Media LCC, New York, 2011), pp. 601–653
2. K. Astala, L. Pivrinta, Calderóns inverse conductivity problem in the plane. Ann. Math. **163**(1), 265–299 (2006)
3. A.P. Calderón, On an inverse boundary value problem. *Seminar on Numerical Analysis and Its Application to Continuum Physics, SOC* (1980), pp. 65–73

4. P. Cardoso et al., Temperature dependence of the electrical conductivity of vapor grown carbon nanofiber/epoxy composites with different filler dispersion levels. Phys. Lett. A **376**(45), 32903294 (2012)
5. M. Hanke, M. Brühl, Recent progress in electrical impedance tomography. Inverse Prob. **19**(6), S65–S90 (2003). Special section on imaging
6. M. Hanke, B. Harrach, N. Hyvönen, Justification of point electrode models in electrical impedance tomography. Math. Models Methods Appl. Sci. **21**(06), 1395–1413 (2011)
7. B. Harrach, Interpolation of missing electrode data in electrical impedance tomography. Inverse Prob. **31**(11), 115008, 20 (2015)
8. B. Harrach, J.K. Seo, Detecting inclusions in electrical impedance tomography without reference measurements. SIAM J. Appl. Math. **69**(6), 1662–1681 (2009)
9. B. Harrach, M. Ullrich, Monotony based imaging in EIT. J. AIP Conf. Proc. **1218**, 1975–1978 (2010)
10. B. Harrach, M. Ullrich, Monotonicity-based shape reconstruction in electrical impedance tomography. SIAM J. Math. Anal. **45**(6), 3382–3403 (2013)
11. B. Harrach, M. Ullrich, Monotony based inclusion detection in EIT for realistic electrode models. J. Phys. Conf. Series **434**(1), 012076 (2013)
12. T.-C. Hou, K.J. Loh, J.P. Lynch, Electrical impedance tomography of carbon nanotube composite materials. Proc. SPIE **6529**(652926), 10 (2007)
13. T.-C. Hou, K.J. Loh, J.P. Lynch, Spatial conductivity mapping of carbon nanotube composite thin films by electrical impedance tomography for sensing applications. Nanotechnology **18**(31), 315501 (2007)
14. B. Jin, P. Maass, An analysis of electrical impedance tomography with applications to Tikhonov regularization. ESAIM Control Optim. Calc. Var. **18**(4), 1027–1048 (2012)
15. B. Jin, T. Khan, P. Maass, A reconstruction algorithm for electrical impedance tomography based on sparsity regularization. Int. J. Numer. Methods Eng. **89**(3), 337–353 (2012)
16. O. Kanoun, U. Tröltzsch, H.-R. Tränkler, Benefits of evolutionary strategy in modeling of impedance spectra. Electrochim. Acta **51**, 1453–1461 (2006)
17. O. Kanoun, C. Müller, A. Benchirouf, A. Sanli, N.T. Dinh, A. Al-Hamry, L. Bu, C. Gerlach, A. Bouhamed, Flexible carbon nanotube films for high performance strain sensors. Sensors **14**, 10042–10071 (2014)
18. K. Knudsen, M. Lassas, J. Mueller, S. Siltanen, Reconstructions of piecewise constant conductivities by the d-bar method for electrical impedance tomography. J. Phys. Conf. Series **124**(1), 012029 (2008)
19. R. Kohn, M. Vogelius, Determining conductivity by boundary measurements. Commun. Pure Appl. Math. **37**, 289–198 (1984)
20. R. Kohn, M. Vogelius, Determining conductivity by boundary measurements ii. Commun. Pure Appl. Math. **37**, 643–667 (1985)
21. S. Kumar, T. Rath, R.N. Mahaling, C.S. Reddy, C.K. Das, K.N. Pandey, R.B. Srivastava, S.B. Yadaw, Study on mechanical, morphological and electrical properties of carbon nanofiber/polyetherimide composites. Mater. Sci. Eng. B **141**, 61–70 (2007)
22. C. Kuo-Sheng, D. Isaacson, J.C. Newell, D.G. Gisser, Electrode models for electric current computed tomography. IEEE Trans. Biomed. Eng. **36**(9), 918–924 (1989)
23. A. Meyer, Programmer's Manual for Adaptive Finite Element Code SPC-PM 2Ad. Preprint SFB393 01-18 TU Chemnitz (2001)
24. A. Meyer, Programmbeschreibung SPC-PM3-AdH-XX Teil 1. Preprint CSC/14-01 TU Chemnitz (2014)
25. A. Meyer, Programmbeschreibung SPC-PM3-AdH-XX Teil 2. Preprint CSC/14-02 TU Chemnitz (2014)
26. A.I. Nachman, Global uniqueness for a two-dimensional inverse boundary value problem. Ann. Math. **143**(1), 71–96 (1996)
27. K. Paulson, B. William, M. Pidcock, Electrode modelling in electrical impedance tomography. SIAM J. Appl. Math. **52**(4), 1012–1022 (1992)

28. S. Siltanen, J. Mueller, *Linear and Nonlinear Inverse Problems with Practical Applications: Electrical Impedance tomography*, (SIAM Philadelphia, 2012) chap. 15, pp. 159–184
29. J. Sylvester, G. Uhlmann, A global uniqueness theorem for an inverse boundary value problem. Ann. Math. **125**(1), 153–169 (1987)
30. T.N. Tallman, S. Gungor, K.W. Wang, C.E. Bakis, Damage detection and conductivity evolution in carbon nanofiber epoxy via electrical impedance tomography. Smart Mater. Struct. **23**(4), 045034, 9 (2014)
31. U. Tröltzsch, O. Kanoun, Generalization of transmission line models for deriving the impedance of diffusion and porous media. Electrochim. Acta **75**, 347–356 (2012)

Bayesian Updating in the Determination of Forces in Euler-Bernoulli Beams

Alexandre Kawano and Abdelmalek Zine

Abstract The beam is among the most important structural elements, and it can fail by different causes. In many cases it is important to access the loading acting on them. The determination of loading on beams is important, for example, for model calibration purposes and or to estimate remaining fatigue life. In this article we first prove that identification of the loading is theoretically possible from the observation of the displacement of small portion of it for an arbitrary small interval of time and then propose a method to infer the spatial distribution of forces acting upon a beam from the measurement of the displacement of one of its points. The Bayesian method is used to combine measurements taken from different points at different times. This method enables an effective way of reducing the practical amount of time for obtaining meaningful loading estimates.

1 Introduction

The beam is among the most important structural elements, and it can fail by different causes. In many cases it is important to access the loading acting on them by indirect methods.

In order to show the importance of the subject, we mention one important example involving the loading acting over risers, which are long tubes used to transport fluids between the sea bottom and the oil platform that is at the sea level. The loading over risers have different origins and is the theme of current research. Needless to say, an accident involving a riser would cause huge environmental damage. Therefore, design codes must be strict when it comes to safety. However,

A. Kawano (✉)
Escola Politécnica, University of Sao Paulo, Sao Paulo, Brazil
e-mail: akawano@usp.br

A. Zine
Ecole Centrale de Lyon, Institut Camille Jordan, Université de Lyon, Ecully, France
e-mail: abdelmalek.zine@ec.fr

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_9

since the history of the use of risers by the oil industry is relatively short, as is the whole history of offshore oil extraction, data for code calibration is still scarce. Constant monitoring of the tubes is the key to avoid catastrophic failures. It is therefore very important to monitor the loading imposed to a riser, and in-situ monitoring strategies are being proposed [10].

The load determination over beams is viewed as an inverse problem related to vibration. Damage detection in beams has been studied, among others, by Barcilon [1], Mclaughlin [8], Morassi [9], and Nicaise and Zair [12]. Here we employ a new method based on almost periodic distributions [7].

Here we are interested in identifying forces acting in a beam. The point of view we take of it is that is an inverse problem [4]. First we prove that the data at our disposal is sufficient for the unique recovery of the loading, and then show a method based on Bayesian updating scheme.

The central equation in this inverse problem is

$$\begin{cases} \rho \dfrac{\partial^2 w}{\partial t^2} + \nu \dfrac{\partial w}{\partial t} + EI \dfrac{\partial^4 w}{\partial x^4} - T \dfrac{\partial^2 w}{\partial x^2} = g(t) f(x), & \text{in }]0, T_0[\times]0, L[\\ w(0,x) = \frac{\partial w}{\partial t}(0,x) = 0, & \forall x \in]0, L[\\ w(t,x) = \frac{\partial^2 w}{\partial x^2}(t,x) = 0, & \forall t \in]0, T_0[, \ \forall x \in \{0, L\}, \end{cases} \tag{1}$$

where $g \in \mathscr{C}^1([0, T_0])$ is a given function with $g(0) \neq 0$ and w is the displacement. The physical parameters are: E is the Young Modulus, I is the moment of inertia of the cross section, ρ is the material's linear density, ν is the damping coefficient and T is the tension force along the beam.

We are interested in determining $f \in \mathrm{H}^{-1}(]0, L[)$. The data available in this inverse problem is the set

$$\Gamma_{T_0, x_0} = \{ w(t, x_0) \ : \ t \in]0, T_0[\}, \tag{2}$$

where $]0, T_0[$ is an arbitrary non empty open set and

$$x_0 \in \{ x \in]0, L[: \ \sin(n\pi x / L) \neq 0, \ \forall n \in \mathbb{N} \}.$$

We must alert the reader to the fact that more information is used besides the one given explicitly by (2). It is important to note that in (1) the boundary conditions are also known. It does not pose any practical problem, for sensors can be put at the extremes of the portion of the beam being analyzed to measure their relative displacements and the bending moments acting there. Since the problem is linear, in the formulation we are allowed to put null boundary conditions.

Due to the presence of damping, measurements taken long after the start of the process, the initial conditions become irrelevant.

2 The Direct Problem

We can obtain a formal solution to Problem (1) by a Galerkin method. Consider the eigenproblem, where $S \in \mathrm{H}_0^1(]0,L[)$:

$$\begin{cases} \dfrac{EI}{\rho}\dfrac{\partial^4 S}{\partial x^4} - \dfrac{T}{\rho}\dfrac{\partial^2 S}{\partial x^2} = \beta S, \quad \text{in }]0,L[, \\ S(0) = S(L) = 0, \\ S''(0) = S''(L) = 0. \end{cases} \tag{3}$$

The eigenvectors are

$$S_n(x) = C_n \sin(\frac{n\pi x}{L}), \tag{4}$$

with the corresponding eigenvalues

$$\beta_n = \frac{EI}{\rho}\left(\frac{n\pi}{L}\right)^4 + \frac{T}{\rho}\left(\frac{n\pi}{L}\right)^2.$$

The constant $C_n = \sqrt{2/L}$ is chosen so that $||S_n||_{\mathrm{L}^2(]0,L[)} = 1$.

Following a standard method, we conclude that $\{S_n : n \in \mathbb{N}\}$ is orthogonal and dense in $\mathrm{H}_0^1(]0,L[)$, as well as orthonormal in $\mathrm{L}^2(]0,L[)$. It follows easily that $\| S_n \|_{\mathrm{H}_0^1(]0,L[)} = \frac{n\pi}{L} \| S_n \|_{\mathrm{L}^2(]0,L[)} = \mathcal{O}(n)$. The set $\left(\frac{S_n}{n}\right)_{n\in\mathbb{N}}$ forms an orthonormal Hilbert basis of $\mathrm{H}_0^1(]0,L[)$, and any function in $\mathrm{H}_0^1(]0,L[)$ can be expressed as $\sum_{n=1}^{+\infty} \frac{A_n}{n} S_n$, with $(A_n)_{n\in\mathbb{N}} \in \ell^2$. From duality pairing, we see that Any distribution $h \in \mathrm{H}^{-1}(]0,L[)$ can be represented uniquely as

$$h = \sum_{n=1}^{+\infty} \tilde{A}_n n S_n,$$

with $(\tilde{A}_n)_{n\in\mathbb{N}} \in \ell^2$. Now we use $\{S_n : n \in \mathbb{N}\}$ to represent the spatial distribution $f \in \mathrm{H}^{-1}(]0,L[)$ as

$$f = \sum_{n=1}^{\infty} A_n S_n, \tag{5}$$

with $(A_n/n)_{n\in\mathbb{N}} \in \ell^2$.

By the method of separation of variables, we assume that the dynamic response of $w(t,x)$ can be represented by:

$$w(t,x) = \sum_{n=1}^{\infty} G_n(t) S_n(x). \tag{6}$$

From (5) and (6) into (1), we obtain:

$$\sum_{n=1}^{\infty}\left(G_n''(t)+\frac{\nu}{\rho}G_n'(t)+\beta_n G_n(t)-A_n\frac{g(t)}{\rho}\right)S_n(x)=0. \tag{7}$$

From the orthogonality of $(S_n)_{n\in\mathbb{N}}$, we have

$$G_n''(t)+\frac{\nu}{\rho}G_n'(t)+\beta_n G_n(t)-A_n\frac{g(t)}{\rho}=0, \tag{8}$$

of which solution is

$$G_n(t)=A_n\int_0^t\frac{g(\tau)}{\omega_n}\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}(t-\tau)}\sin((t-\tau)\omega_n)\,\mathrm{d}\tau, \tag{9}$$

where $\omega_n=\sqrt{\beta_n-\left(\frac{\nu}{2\rho}\right)^2}$.

From (6), we have

$$w(t,x)=\sum_{n=1}^{\infty}\frac{A_n}{\omega_n}\left[\int_0^t g(\tau)\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}(t-\tau)}\sin(\omega_n(t-\tau))\,\mathrm{d}\tau\right]S_n(x). \tag{10}$$

If $\omega_n=0$, then $\sin(\tau\omega_n)/\omega_n$ should be replaced by τ in the formulas above.

Proposition 1 *For any $t\in[0,T_0]$, the traces of ω at the boundaries are well defined, and that $w\in\mathscr{C}([0,T_0],\mathrm{H}_0^1(]0,L[))\cap\mathscr{C}^1([0,T_0],\mathrm{H}_0^1(]0,L[))$.*

Proof From the fact that $g\in\mathscr{C}^1([0,T_0])$, by an integration by parts we see that there is a $C_{T_0}>0$ independent of ω_n and of g such that

$$\int_0^t g(t-\tau)\sin(\omega_n\tau)\,\mathrm{d}\tau<\frac{C_{T_0}}{\omega_n}\parallel g\parallel_{\mathscr{C}^1[0,T_0]}, \tag{11}$$

$$\frac{\partial}{\partial t}\int_0^t g(t-\tau)\sin(\omega_n\tau)\,\mathrm{d}\tau<C_{T_0}\parallel g\parallel_{\mathscr{C}^1[0,T_0]}. \tag{12}$$

From (10) and (11) we conclude that for any $t\in[0,T_0]$, $w(t)$ is an element of $\mathrm{H}_0^1(]0,L[)$, and therefore the traces of ω at the boundaries are well defined. Also from (10) and (11), we obtain that $w\in\mathscr{C}([0,T_0],\mathrm{H}_0^1(]0,L[))$.

To see that $w\in\mathscr{C}^1([0,T_0],\mathrm{H}_0^1(]0,L[))$, just take an arbitrary $\phi\in\mathrm{H}^{-1}(]0,L[)$ and consider the function $t\mapsto\langle\frac{\partial w}{\partial t}(t,\cdot)\,,\,\phi\rangle$. Using (12) and the fact that $(\langle S_n\,,\,\phi\rangle)_{n\in\mathbb{N}}\in\ell^2$, we conclude that $w\in\mathscr{C}^1([0,T_0],\mathrm{H}_0^1(]0,L[))$. □

By applying the Laplace transform in the t variable, we can readily see that $w\in\mathscr{C}^1([0,T_0],\mathrm{H}_0^1(]0,L[))$ that satisfies (1) is unique.

From a simple application of the Theorem of Dominated Convergence, we obtain that the solution (10) can be written as

$$w(t,x) = \int_0^t g(t-\tau)\left[\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}\tau}\sum_{n=1}^{\infty}\frac{A_n}{\omega_n}\sin(\omega_n\tau)S_n(x)\right]\mathrm{d}\tau, \tag{13}$$

for $t \in [0, T_0]$ and $x \in [0, L]$.

3 The Inverse Problem

We are going to use a result (Theorem 1 below) found in [12], of which proof can be found in [11] (see also [3]).

For $T_0 > 0$, $g \in \mathscr{C}^1([0, T_0])$, with $g(0) \neq 0$, define the operator $K : \mathrm{L}^2(0, T_0) \to \mathrm{L}^2(0, T_0)$ by

$$(K\psi)(t) = \int_0^t g(t-s)\psi(s)\,\mathrm{d}s, \quad \forall t \in]0, T_0[.$$

Define the space $\mathscr{G} \subset \mathrm{L}^2(0, T_0)$ by

$$\mathscr{G} = \left\{\eta \in \mathrm{L}^2(0, T_0) \,:\, (g, \eta)_{\mathrm{L}^2(0,T_0)} = 0\right\},$$

and projection operator $P : \mathrm{L}^2(0, T_0) \to \mathscr{G}$.

Theorem 1 *The operator* $PK : \mathrm{L}^2(0, T_0) \to \mathscr{G}$ *can be extended to a bounded operator from* $\mathrm{H}^{-1}(0, T_0)$ *into* $\mathrm{L}^2(0, T_0)$ *that satisfies*

$$C^{-1} \parallel PK\psi \parallel_{\mathrm{L}^2(0,T_0)} \leq \parallel \psi \parallel_{\mathrm{H}^{-1}(0,T_0)} \leq C \parallel PK\psi \parallel_{\mathrm{L}^2(0,T_0)},$$

for some constant $C > 0$.

Now apply Theorem 1 to (13), using the data $\Gamma_{T_0,x_0} = \{0\}$, to conclude that

$$\sum_{n=1}^{\infty}\frac{A_n}{\omega_n}\sin(\omega_n\tau)S_n(x_0) = 0, \quad \forall \tau \in [0, T_0[. \tag{14}$$

Now we invoke a theorem for the uniqueness of the sequence of coefficients in an almost periodic distribution posed in a form like (14).

The following result is from [7]. We recall that a sequence $(\lambda_n)_{n\in\mathbb{N}} \subset \mathbb{C}$ is uniformly discrete if there is an $\epsilon > 0$ such that $p \neq q \Rightarrow \left|\lambda_p - \lambda_q\right| > \epsilon$, and s' is the space of slowly growing sequences. That is, if $(a_n)_{n\in\mathbb{N}} \in s'$, then $\exists q \in \mathbb{Z}_+$ such that $(n^{-q}a_n)_{n\in\mathbb{N}} \in \ell^1$.

Theorem 2 *Given $\Lambda = (\lambda_n)_{n\in\mathbb{N}}$, uniformly discrete, is such that $\exists n_0 \in \mathbb{N}$, $\exists C \in \mathbb{R}_+$ such that $n > n_0 \Rightarrow |\lambda_n| > Cn^\alpha$, if $\alpha > 1$ and $(a_n)_{n\in\mathbb{N}} \in s'$, then if there is a $\tau > 0$ such that $\sum_{n\in\mathbb{N}} a_n \, \mathrm{e}^{\imath \lambda t} = 0, \forall t \in [-\tau, \tau]$, then $(a_n)_{n\in\mathbb{N}} = \{0\}$.*

Now apply Theorem 2 to (14) to conclude that $A_n = 0$. We conclude that the data $\Gamma_{T_0,x_0} = \{0\}$ is enough to determine uniquely $(A_n)_{n\in\mathbb{N}}$, and consequently the distribution $f \in \mathrm{H}^{-1}(]0, L[)$ in (1).

To summarize, we have just proved the following uniqueness result.

Theorem 3 *In problem (1), with $g \in \mathscr{C}^1([0, T_0])$, $T_0 > 0$, $g(0) \neq 0$, the data $\Gamma_{T_0,x_0} = \{w(t, x_0) : t \in]0, T_0[\}$ is enough to determine uniquely $f \in \mathrm{H}^{-1}(]0, L[)$ in (1).*

Then rearranging the terms, we would end up with a sum

$$\sum_{n=1}^{\infty} \frac{A''_n}{\omega''_n} \sin(\omega''_n \tau) S''_n(x) = 0.$$

If this is true $\forall t \in]0, T_0[$, then the only possibility is $A''_n = 0$, $\forall n \in \mathbb{N}$. That is, there exists at maximum only one representation of the form

$$\sum_{n=1}^{\infty} \frac{A_n}{\omega_n} \sin(\omega_n \tau) S_n(x),$$

used in (14), and therefore, the set $\{\omega_n : n \in \mathbb{N}\}$ is unique.

3.1 Recovery Procedure

In this section, we propose a method for the recovery of $f \in \mathrm{H}^{-1}(]0, L[)$.

Suppose that

$$g_M(t) = \sum_{k=0}^{M} k_m \cos(\tilde{\omega}_m t),$$

where $\tilde{\omega}_m = \pi m t / T_0$, is obtained by the truncation of the Fourier series of the even extension of $g \in \mathscr{C}^1[0, T_0]$ to the interval $[-T_0, T_0]$, for $M \in \mathbb{N}$. It is known that $g_M \to g$ absolutely and uniformly. Due to (10) and (11), if w_M is the solution that satisfies (1) when g is replaced by g_M, then $w_M \to w$ uniformly in $[0, T_0] \times [0, L]$. Of course, if we have an a priori estimate for the unknown coefficients

$$\| (A_n)_{n\in\mathbb{N}} \|_{\ell^2} < \epsilon_0,$$

which is obtained easily from an upper bound of the L^2 norm of g in the interval $[0, T_0]$, then, given any $\epsilon > 0$, it is always possible to choose a $M \in \mathbb{N}$ such that, for each $x \in \Omega$, $\| w(\cdot, x) - w_N(\cdot, x) \|_{\mathrm{L}^2(]0,T_0[)} < \epsilon$.

We can regard the solution of (1) generating the data for an inverse problem with g_M instead of the original g.

Solving (1) for g replaced by g_M, we arrive at the following solution:

$$
\begin{aligned}
w_M(t,x) = &\sum_{n=1}^{\infty}\sum_{m=1}^{M} \frac{A_n S_n(x)}{\omega_n(\nu^2 + 4\rho^2(\tilde{\omega}_m - \omega_n)^2)(\nu^2 + 4\rho^2(\tilde{\omega}_m + \omega_n)^2)} \\
&\Big[\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}t}[-2\nu\rho(\nu^2 + 4\rho^2(\tilde{\omega}_m^2 + \omega_n^2))\sin(\omega_n t) \\
&-4\omega_n\rho^2(\nu^2 + 4\rho^2(\omega_n^2 - \tilde{\omega}_m^2))\cos(\omega_n t)] \\
&+4\omega_n\rho^2[(\nu^2 + 4\rho^2(\omega_n^2 - \tilde{\omega}_m^2))\cos(\tilde{\omega}_m t) + 4\tilde{\omega}_m\nu\rho\sin(\tilde{\omega}_m t)]\Big].
\end{aligned}
\tag{15}
$$

Because we are dealing with absolutely convergent series, we can rewrite (15) as

$$
\begin{aligned}
w_M(t,x) = &\sum_{n=1}^{\infty} A_n \sum_{m=1}^{M} \frac{S_n(x)}{\omega_n(\nu^2 + 4\rho^2(\tilde{\omega}_m - \omega_n)^2)(\nu^2 + 4\rho^2(\tilde{\omega}_m + \omega_n)^2)} \\
&\Big[\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}t}[-2\nu\rho(\nu^2 + 4\rho^2(\tilde{\omega}_m^2 + \omega_n^2))\sin(\omega_n t) \\
&-4\omega_n\rho^2(\nu^2 + 4\rho^2(\omega_n^2 - \tilde{\omega}_m^2))\cos(\omega_n t)]\Big] \\
&+\sum_{m=1}^{M}\sum_{n=1}^{\infty} \frac{A_n S_n(x)}{\omega_n(\nu^2 + 4\rho^2(\tilde{\omega}_m - \omega_n)^2)(\nu^2 + 4\rho^2(\tilde{\omega}_m + \omega_n)^2)} \\
&4\omega_n\rho^2[(\nu^2 + 4\rho^2(\omega_n^2 - \tilde{\omega}_m^2))\cos(\tilde{\omega}_m t) + 4\tilde{\omega}_m\nu\rho\sin(\tilde{\omega}_m t)].
\end{aligned}
\tag{16}
$$

Observe that is always possible to choose $T_0 > 0$ so that

$$\left\{\tilde{\omega}_m = \frac{\pi m}{T_0} \,:\, m = 1, \ldots, M\right\} \cap \{\omega_n \,:\, n \in \mathbb{N}\} = \emptyset.$$

In this case, the set $\tilde{\Lambda} = \{\tilde{\omega}_m \,:\, m = 1, \ldots, M\} \cup \{\omega_n \,:\, n \in \mathbb{N}\}$ is still uniformly discrete.

Manipulating (16), we get

$$
\begin{aligned}
w_M(t,x) = &\sum_{n=1}^{\infty}\left[\left(\sum_{m=1}^{M} B_{m,n}^{(1)}\right)\sin(\omega_n t) + \left(\sum_{m=1}^{M} B_{m,n}^{(2)}\right)\cos(\omega_n t)\right] \\
&+\sum_{m=1}^{M}\left[\left(\sum_{n=1}^{\infty} C_{m,n}^{(1)}\right)\sin(\tilde{\omega}_m t) + \left(\sum_{n=1}^{\infty} C_{m,n}^{(2)}\right)\cos(\tilde{\omega}_m t)\right],
\end{aligned}
\tag{17}
$$

where

$$\begin{aligned}
B^{(1)}_{m,n} =& \frac{-2\nu\rho A_n S_n(x)\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}t}(\nu^2+4\rho^2(\tilde{\omega}_m^2+\omega_n^2))}{\omega_n(\nu^2+4\rho^2(\tilde{\omega}_m-\omega_n)^2)(\nu^2+4\rho^2(\tilde{\omega}_m+\omega_n)^2)},\\
B^{(2)}_{m,n} =& \frac{-4\rho^2 A_n S_n(x)\omega_n\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}t}(\nu^2+4\rho^2(\omega_n^2-\tilde{\omega}_m^2))}{\omega_n(\nu^2+4\rho^2(\tilde{\omega}_m-\omega_n)^2)(\nu^2+4\rho^2(\tilde{\omega}_m+\omega_n)^2)},\\
C^{(1)}_{m,n} =& \frac{16 A_n S_n(x)\omega_n\tilde{\omega}_m\nu\rho^3}{\omega_n(\nu^2+4\rho^2(\tilde{\omega}_m-\omega_n)^2)(\nu^2+4\rho^2(\tilde{\omega}_m+\omega_n)^2)},\\
C^{(2)}_{m,n} =& \frac{4 A_n S_n(x)\omega_n\rho^2(\nu^2+4\rho^2(\omega_n^2-\tilde{\omega}_m^2))}{\omega_n(\nu^2+4\rho^2(\tilde{\omega}_m-\omega_n)^2)(\nu^2+4\rho^2(\tilde{\omega}_m+\omega_n)^2)}
\end{aligned}$$

Writing the trigonometric functions as sums of exponentials, after some manipulation we obtain

$$w_M(t,x) = \sum_{n=1}^{+\infty}\tilde{B}_n\mathrm{e}^{\imath\,\lambda_{n,1}t} + \sum_{m=1}^{M}\tilde{C}_m\mathrm{e}^{\imath\,\lambda_{m,2}t}, \tag{18}$$

where $\lambda_{n,1} = (-1)^n\omega_{\lceil\frac{n}{2}\rceil}$ and $\lambda_{m,2} = (-1)^m\tilde{\omega}_{\lceil\frac{m}{2}\rceil}$. $\lceil n/2\rceil$ denotes the least integer x such that $n/2 \le x$. Besides,

$$\begin{aligned}
\tilde{B}_n &= (-1)^n\frac{\sum_{m=1}^{M}B^{(1)}_{m,\lceil n/2\rceil}}{2\imath} + \frac{\sum_{m=1}^{M}B^{(2)}_{m,\lceil n/2\rceil}}{2},\\
\tilde{C}_m &= (-1)^m\frac{\sum_{n=1}^{+\infty}C^{(1)}_{\lceil m/2\rceil,n}}{2\imath} + \frac{\sum_{n=1}^{+\infty}C^{(2)}_{\lceil m/2\rceil,n}}{2}.
\end{aligned} \tag{19}$$

Define $\Lambda_1 = (\lambda_{n,1})_{n\in\mathbb{N}} \cup (\lambda_{m,2})_{m\in\mathbb{Z}_+}$. Now reorder Λ_1 increasingly with respect to the absolute value of its elements to obtain from Λ_1 the new ordered sequence $\Lambda = (\lambda_n)_{n\in\mathbb{N}}$. Clearly, (18) can be put in the form

$$w_M(t,x) = \sum_{n=1}^{+\infty}\alpha_n\mathrm{e}^{\imath\,\lambda_n t}, \tag{20}$$

where the majority of coefficients α_n belongs to the infinite set $(\tilde{B}_n)_{n\in\mathbb{N}}$ with a finite number of them coming from $(\tilde{C}_m)_{m\in\{1,\ldots,2M\}}$.

In order to use the method to be described below, we approximate each α_n which depends on t, to a constant mean value in time. This amounts to use the average

$$\frac{1}{T_0}\int_0^{T_0}\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}t}\,\mathrm{d}t = \frac{1}{2}\frac{\rho}{\nu}(1-\mathrm{e}^{-\frac{T_0}{2}\frac{\nu}{\rho}})$$

in place of the function $\mathrm{e}^{-\frac{1}{2}\frac{\nu}{\rho}t}$.

Following [7], we use a family of functions

$$\phi_{1,m,\tau}(\xi) = \frac{[\sin((\xi - \lambda_m)\tau)]^2}{(\xi - \lambda_m)^2\tau^2}, \quad \forall m \in \mathbb{N},\ \forall \tau > 0.$$

Observe that their Fourier transform are compactly supported:

$$\widehat{\phi_{1,m,\tau}}(t) = (H_\tau * H_\tau)(t)e^{-it\lambda_m}, \quad H_\tau(t) = \frac{1}{2\tau}\chi_{]-\tau,\tau[}(t), \tag{21}$$

Define now $V(m) = \langle w(\cdot, x), \widehat{\phi_{1,m,\tau}}\rangle$ and $P_\tau(m,n) = \phi_{1,m,\tau}(\lambda_n)$. Consider the operator $T : \ell^2 \to s'$, given $(\alpha_n)_{n\in\mathbb{N}} \to (V(m))_{m\in\mathbb{N}}$. Formally, applying T can be interpreted as performing a product with a matrix of infinite order,

$$\begin{bmatrix} V(1) \\ V(2) \\ \vdots \end{bmatrix} = \begin{bmatrix} P_\tau(1,1) & P_\tau(1,2) & P_\tau(1,3) & \dots \\ P_\tau(2,1) & P_\tau(2,2) & P_\tau(2,3) & \dots \\ \vdots & \vdots & \vdots & \dots \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \end{bmatrix}. \tag{22}$$

Perform now a truncation of the system in (22) to obtain

$$\begin{bmatrix} V(1) \\ V(2) \\ \vdots \\ V(N) \end{bmatrix} = \underbrace{\begin{bmatrix} P_\tau(1,1) & P_\tau(1,2) & \dots & P_\tau(1,N) \\ P_\tau(2,1) & P_\tau(2,2) & \dots & P_\tau(2,N) \\ \vdots & \vdots & \vdots & \vdots \\ P_\tau(N,1) & P_\tau(N,2) & \dots & P_\tau(N,N) \end{bmatrix}}_{\mathrm{T_N}} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}. \tag{23}$$

By solving the linear system (23), we obtain the first elements of the desired sequence $(\alpha_n)_{n\in\mathbb{N}}$. All formal operations above, including convergence and stability considerations, are proved in [7].

As far as stability is concerned, suppose that the information we have at our disposal is $w_e(t, x_0) \doteq w(t, x_0) + e(t)$, where $e(t)$ is the measurement error, which can be bounded a priori by

$$||e||_{L^2(]0,T_0[)} \leq \epsilon_{\text{error}}.$$

When the measurements contain errors, then instead of $V(m)$, we have at our disposal $\tilde{V}(m) = \langle w_e(t, x_0), \widehat{\phi_{1,m,\tau}}\rangle$.

The solution of the linear system (23) gives

$$(\tilde{\alpha}_n)_{n=1}^N = \mathrm{T_N}^{-1}((V(m))_{m=1}^N) + \mathrm{T_N}^{-1}((V_e(m))_{m=1}^N),$$

where

$$V_e(m) = \langle e, \widehat{\phi_{1,m,\tau}} \rangle.$$

In [7] it is proved that ${\mathrm{T_N}}^{-1}((V(m))_{m=1}^N) \overset{N \to +\infty}{\longrightarrow} ((\alpha_n)_{n \in \mathbb{N}})$ and that there exists $C > 0$ such that $||\mathrm{T}_\mathrm{N}^{-1}((V_e(m))_{m=1}^N)||$ given by

$$||{\mathrm{T_N}}^{-1}(V_e(m)))_{m=1}^N|||_{\ell^2} \le C\epsilon_{\text{error}} \sqrt{\frac{N+1}{2\tau}}.$$

In practice, the error ϵ_{error} incorporates not only the measurement errors but also those arising from numerical computation. Because of this, as usual, precision is lower than the one expected if only measurement error were present.

3.2 Bayesian Updating

Now we apply the Bayesian updating method. It is well known that the Tikhonov regularization may be regarded as a special case of Bayesian updating process (see for instance, [6]). Furthermore, the method makes it possible to incorporate previous experimental results and even subjective expert opinion into the analysis.

We suppose in this section that vector $[V] = [V(1) \ \ldots \ V(N)]^t$ in (23) is corrupted by noise. That is, $[V]$ is replaced by

$$[\tilde{V}] = [V] + [\mathscr{E}],$$

where $\mathscr{E}$ is random variable. We further suppose that $\mathscr{E}$ is normally distributed with zero mean, and that it possesses a covariance matrix $\sigma_\epsilon^2[I]$.

The likelihood function for $[\alpha]$, given the observation $[\tilde{V}]$ is

$$L([\alpha] \mid [\tilde{V}]) = \frac{1}{(\sigma_\epsilon \sqrt{2\pi})^N} \exp\left[\sigma_\epsilon^{-2N}([\tilde{V}] - [\mathrm{T_N}][\alpha])^t([\tilde{V}] - [\mathrm{T_N}][\alpha])\right] \tag{24}$$

If we attach to $[\alpha]$ a probability density functions in the Bayesian sense, we can assign to $[\alpha]$ a prior and a posterior probability density functions, f_{prior} and f_{post} respectively, that quantify the knowledge about $[\alpha]$. The prior density f_{prior} can incorporate previous analysis and also subjective opinions [5, 6].

By the Bayes rule, we have

$$f_{post}([\alpha]) = \frac{L([\alpha] \mid [\tilde{V}]) f_{prior}([\alpha])}{K([\tilde{V}])},$$

where $K([\tilde{V}])$ is a normalizing constant defined so that

$$\int f_{post}([\alpha]) = 1.$$

To ease all computations, we use a Bayesian conjugate pair. Since the likelihood function L assumes the form of a Gaussian distribution, we assume that the prior also has this form in the computation of the posterior distribution.

In fact, it is known (see for instance, [2]) that taking $[\alpha_{prior}] \sim N([\alpha_0], \sigma_\epsilon^2[N_0])$, that is, $[\alpha_{prior}]$ is normally distributed, then $[\alpha_{post}]$ is also normally distributed, with posterior distribution mean is given

$$[\alpha_1] = \frac{n_0\mu_0 + n\bar{x}}{n_0 + n}, \tag{25}$$

where $\bar{x}$ is the average of the $[\alpha]$ that solves (23), and the pair (α_0, n_0) corresponds to the prior. Note that n_0 can be interpreted as the weight attributed to the initial guess.

When all data gathered from measurements are incorporated into the analysis, the best estimator of $[\alpha]$ is the mean (25).

Now it is important to realize that after the posterior distribution is obtained, it can be reused as a new prior for a new application of the Bayes rule, when new data is acquired. This is one of the advantages of the Bayesian method, since it can incorporate previous experimental evidences in a easy way. Also observe that the first prior employed in the beginning of the process incorporates subjective opinion regarding the parameters, about their joint distribution, mean and dispersion.

In the next section, the Bayesian method is illustrated by an example in which in the first step, the displacement of a point x_0 is used as the data to obtain the posterior distribution. Then this posterior is reused as a prior for the next step, but the observation point is taken in another location x_2. This process is continued iteratively for a finite number of steps.

4 Numerical Experiments

To illustrate the theory above, we show some numerical experiments. Consider a beam that models a span of $L = 100m$ of a riser under traction with the following parameters (all values are in the metric system, SI): $EI = 2.7 \times 10^7$, $\rho = 1.3 \times 10^1$, $T = 5.0 \times 10^5$, $\nu = 0.1$.

The excitation force used to simulate the dynamics of the system is

$$h(t,x) = \cos(\tilde{\omega}_1 t) f(x),$$

with $\tilde{\omega}_1 = 6$, $g(x) = \sum_{j=1}^{15} A_n \sqrt{2/L} \sin(\pi j x/L)$, $(A_n)_{n=1,\dots 15} = (3.64186, -2.94632, -0.463688, 2.03586, -0.900316, -0.143288, -0.198724, 0.388815, 0.404651, -0.900316, 0.331078, 0.25921, -0.107005, -0.061409, -0.300105)$.

Function g is an approximation for an unit uniformly distributed load spanning from $x = 0.5L$ and $0.8L$. Of course, since the problem is linear, it suffices to multiply the excitation by a constant factor to obtain more realistic displacement values.

The function $w(t, x_0)$ (in this case, $w = w_M$ with $M = 1$) generated with the data above is shown in Fig. 1. A random noise uniformly distributed over $[-\epsilon, \epsilon]$, $\epsilon = 0.01$ was added to $w(t, x_0)$. This disturbed data is used for the recovery of the first five elements of (A_n).

Observe that the synthetic data is obtained using f expressed as a Fourier series with 15 terms. However, we are going to recover only its first seven terms. Since each sin or cos function originates two exponential terms, after some manipulation and reordering, in (20) will end up with 16 terms (14 for the spatial Fourier series and 2 for the time term $\cos(\tilde{\omega}_1 t)$).

Solving the 16×16 linear system (23), we get $\tilde{\alpha}_n$, $n = 1, \dots, 16$. From $\tilde{\alpha}_n$ we obtain A_n from (19).

The measurements are done in intervals of time of $T_0 = 20$ and $T_0 = 40$ s at point $x_0 = \frac{5}{11}L$. The results from the recovery process are shown in Figs. 2 and 3. Note that since we are recovering only the first seven terms of a summation of 15, the best result possible is represented by the solid line in these figures marked as "Target f".

Now we combine information gathered from several measurements by using the Bayesian updating scheme. The interval of time used in the recovery of f was considerably less: $T_0 = 4$ s only, but as it is shown in Fig. 4, even with just one measure point $x_0 = \frac{1}{11}L$, the result is better. This is due to the fact that in the Bayesian scheme there is built-in also a Tikhonov regularization.

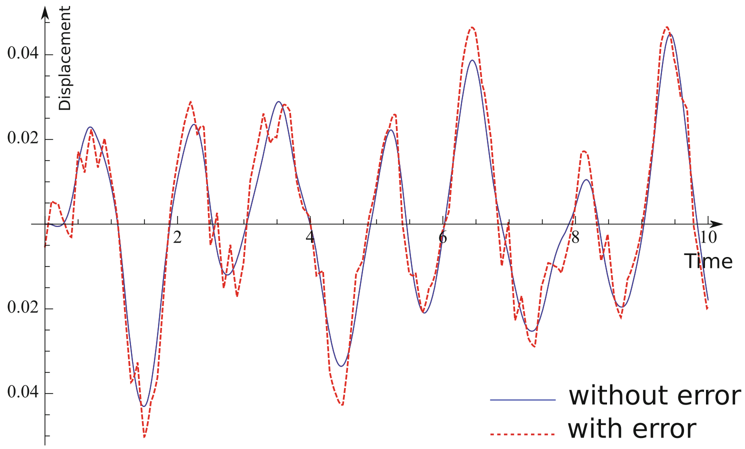


Fig. 1 Displacement $w(t, x_0)$ without and with error

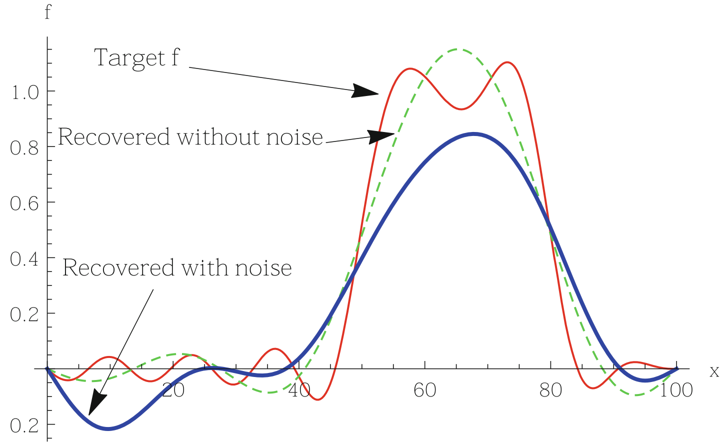


Fig. 2 Recovered f with $T_0 = 20\,\mathrm{s}$

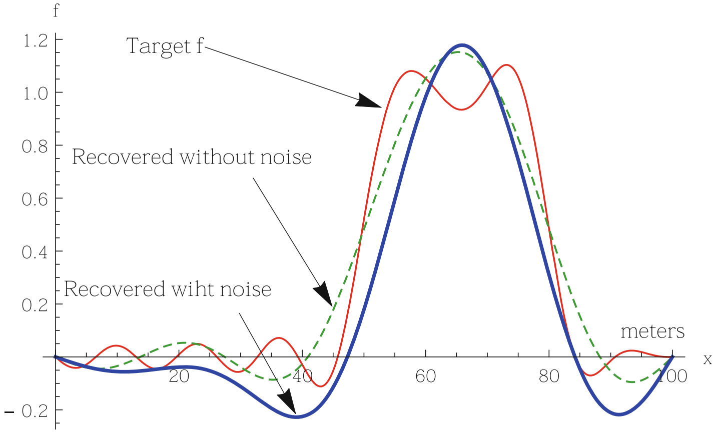


Fig. 3 Recovered f with $T_0 = 40\,\mathrm{s}$

In Fig. 5 it is shown the result when two and three measurement points at $\frac{1}{11}L$ and $\frac{2}{11}L$ are added to $x_0 = \frac{5}{11}L$. The time interval is still $T_0 = 4\,\mathrm{s}$. Observe that the results obtained for two and three observation points coincide. It becomes clear that the Bayesian updating scheme can combine measurements taken at different points, and that the quality of the result is far superior if just one measurement is taken with no regularization besides truncation.

Now show several numerical experiments similar to the experiment above. We consider the same beam under the same loading, but we have chosen points x_0 of the sequence $\{\frac{100}{101}, \frac{200}{101}, \frac{300}{101} \ldots\}$, in increasing order.

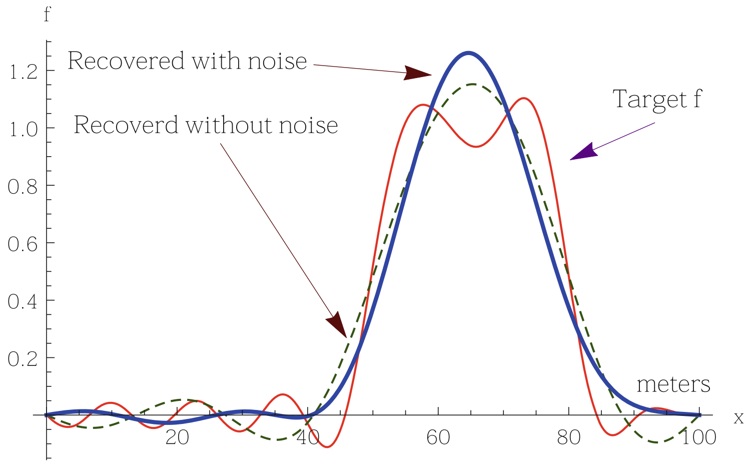


Fig. 4 Recovered f with $T_0 = 4$ s; one observation point with Bayesian updating

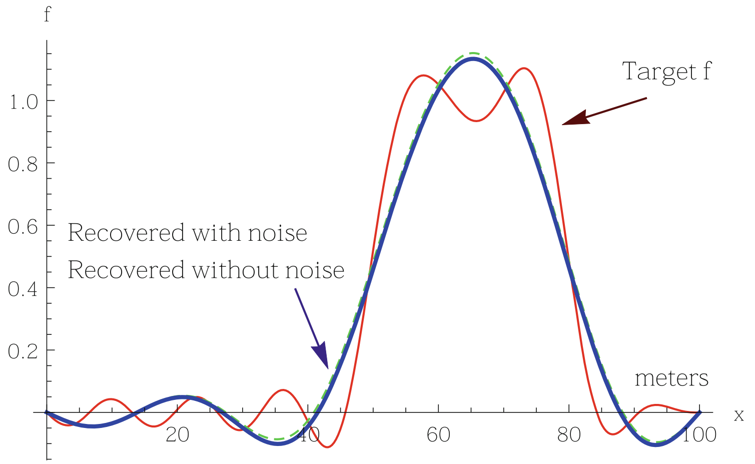


Fig. 5 Recovered f with $T_0 = 4$ s; two and three observation points with Bayesian updating

The data is organized in tables. We quantify the L^1 norm of the error between the target function f and the function estimated by the Bayesian method as

$$\tilde{\mathscr{E}} = \|f - f_{estimated}\|_1 .$$

We define also the mean L^1 error by $\mathscr{E} = \tilde{\mathscr{E}}/L$. In the tables below, n is the number of measurements and T_0 is the period of observation.

To exemplify how the convergence changes with the change of the measurement error, we show a table, with two distinct ϵ, that denote the standard deviation of the error.

	$\epsilon = 0.05$		$\epsilon = 0.2$	
T_0	n	$\mathscr{E}$	n	$\mathscr{E}$
10.0	1	19.868	1	21.0924
10.0	5	6.40178	5	4.89268
10.0	10	5.921161	10	4.44392
5.0	1	19.5052	1	18.9911
5.0	5	6.55553	5	6.93032
5.0	10	4.6693	10	6.89589
2.5	1	19.7712	1	20.8417
2.5	5	7.72417	5	7.3979
2.5	10	6.89102	10	7.1898
2.0	1	22.8154	1	25.2022
2.0	5	12.3089	5	12.4535
2.0	10	11.682	10	12.116

We made measurements until $n = 10$, but, as we can see, the mean error $\mathscr{E}$ increases as the time decreases due to the error generated added to the numerical instability. The condition number of the matrix T_N, when $T_0 = 1.5$, is 15,636, and its determinant is 1.11×10^{-7}.

5 Conclusion

In this paper we proved that the spatial distribution of the loading acting on an elastic beam can be uniquely determined by knowing the displacement of a point over an arbitrarily small interval of time. However, although it is an relevant mathematical fact, from the applications point of view, it is more important to verify if such observation can be used to really identify loads. To answer this question, we performed some numerical experiments.

If we consider just one observation point, we see, comparing Figs. 2 and 3, that the time observation span is important. This is a result that we may call intuitive. What is not intuitive is the efficiency of the Bayesian updating scheme. Comparing Figs. 2 and 4, we see that the Bayesian scheme applied to an observation of 4 s is better that one obtained with 20 s without that scheme. As remarked earlier, this is explained by the Tikhonov regularization that is built-in in the Bayesian method. With two observation points using data gathered for only 4 s we obtained a result that, from a practical point of view, reached the full recovery of the loading.

The numerical experiments backup the theoretical uniqueness result. Furthermore, inferring from them we may say that the Bayesian updating scheme renders the recovery of the loading practical.

Acknowledgements The authors thank the partial support of FAPESP Proc. 2015/50461-0.

References

1. V. Barcilon, Inverse problem for the vibrating beam in the free–clamped configuration. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Sci. (1934–1990) **304**(1483), 211–251 (1982)
2. J.M. Bernardo, A.F.M. Smith, *Bayesian Theory* (Wiley, New York, 2000)
3. G. Bruckner, M. Yamamoto, On the determination of point sources by boundary observations: uniqueness, stability and reconstruction. Preprint WIAS **252**, 1–15 (1996)
4. V. Isakov, *Inverse Problems for Partial Differential Equations*, 2nd edn. (Springer, Berlin, 2006)
5. R. Jeffrey, *Subjective Probability* (Cambridge University Press, Cambridge, 2014)
6. J. Kaipio, E. Somersalo, *Statistical and Computational Inverse Problems* (Springer, Berlin, 2005)
7. A. Kawano, A. Zine, Uniqueness and nonuniqueness results for a certain class of almost periodic distributions. SIAM J. Math. Anal. **43**(1), 135–152 (2011)
8. J.R. Mclaughlin, An inverse eigenvalue problem of order four. SIAM J. Math. Anal. **7**(5), 646–661 (1976)
9. A. Morassi, Damage detection and generalized Fourier coefficients. J. Sound Vib. **302**(1–2), 229–259 (2007)
10. H. Mukundan, Y. Modarres-Sadeghi, J.M. Dahl, F.S. Hover, M.S. Triantafyllou, Monitoring VIV fatigue damage on marine risers. J. Fluids Struct. **25**(4), 617–628 (2009)
11. S. Nicaise, O. Zair, Identifiability, stability and reconstruction results of point sources by boundary measurements in heterogeneous trees. Revista Matemática Complutense **16**(1), 151–178 (2003)
12. S. Nicaise, O. Zair, Determination of point sources in vibrating beams by boundary measurements: identifiability, stability, and reconstruction results. Electron. J. Differ. Equ. **2004**(20), 1–17 (2004)

On Nonstationary Iterated Tikhonov Methods for Ill-Posed Equations in Banach Spaces

M. P. Machado, F. Margotti, and Antonio Leitão

Abstract In this article we propose a novel *nonstationary iterated Tikhonov* (nIT) type method for obtaining stable approximate solutions to ill-posed operator equations modeled by linear operators acting between Banach spaces. We propose a novel a posteriori strategy for choosing the sequence of regularization parameters (or, equivalently, the Lagrange multipliers) for the nIT iteration, aiming to obtain a fast decay of the residual.

Numerical experiments are presented for a 1D convolution problem (smooth Tikhonov functional and Banach parameter-space), and for a 2D deblurring problem (nonsmooth Tikhonov functional and Hilbert parameter-space).

1 Introduction

In this article we propose and (numerically) investigate a new *nonstationary Iterated Tikhonov* (nIT) type method [6, 9] for obtaining stable approximations of linear ill-posed problems modeled by operators mapping between Banach spaces.

The novelty of our approach consists in adopting an a posteriori strategy for the choice of the Lagrange multipliers, which aims to achieve a predefined decay of the residual in each iteration. This strategy differs from the classical choice for the Lagrange multipliers in [9, 10], which propose an a priori strategy, and leads to an unknown decay rate of the residual.

The *inverse problem* we are interested in consists of determining an unknown quantity $x \in X$ from the set of data $y \in Y$, where X, Y are Banach spaces. In practical situations, one does not know the data exactly; instead, only approximate

M. P. Machado
IMPA, Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, Brazil
e-mail: majentcha@gmail.com

F. Margotti · A. Leitão (✉)
Department of Mathematics, Federal University of St. Catarina, P.O. Box 476,
88040-900 Florianópolis, Brazil
e-mail: fabiomarg@gmail.com; acgleitao@gmail.com

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_10

measured data $y^\delta \in Y$ are available with

$$\|y^\delta - y\|_Y \leq \delta, \tag{1}$$

where $\delta > 0$ is the (known) noise level. The available data y^δ are obtained by indirect measurements of the parameter x, this process being described by the ill-posed operator equation

$$Ax = y^\delta, \tag{2}$$

where $A : X \to Y$ is a bounded linear operator, whose inverse $A^{-1} : R(A) \to X$ either does not exist, or is not continuous. For a comprehensive study of this type of problems, we refer the reader to the text book [13] and to the references therein.

Iterated Tikhonov type methods are typically used for linear inverse problems. Applications of this method for linear operator equations in Hilbert spaces can be found in [9] (see also [4] for the nonlinear case). In the Hilbert space setting, both a priori and a posteriori strategies for choosing the Lagrange multipliers have been extensively analyzed [6].

The research on the Banach space setting is still ongoing. Some preliminary results can be found in [10] for linear operator equations, and in [11] for nonlinear systems. In both references above, the authors consider a priori strategies for choosing the Lagrange multipliers.

The approach presented here is devoted to the Banach space setting, and consists in adopting an a posteriori strategy for the choice of the Lagrange multipliers. The penalty terms used in our Tikhonov functionals are the same as in [11] and consist of Bregman distances induced by (uniformly) convex functionals (e.g., the sum of the L^2-norm with the TV-seminorm).

This chapter is outlined as follows: In Sect. 2 a revision of relevant background material is presented. In Sect. 3 we introduce our nIT method. In Sect. 4 possible implementations of our method are discussed; the evaluation of the Lagrange multipliers is addressed, as well as the issue of minimizing the Tikhonov functionals. Section 5 is devoted to numerical experiments, while in Sect. 6 we present some final remarks and conclusions.

2 Background Material

For details about the material discussed in this section, we refer the reader to the textbooks [3] and [13].

Unless the contrary is explicitly stated, we always consider X a *real* Banach space. The *effective domain* of the convex functional $f : X \to \overline{\mathbb{R}} := (-\infty, \infty]$ is defined as

$$\mathrm{Dom}\,(f) := \{x \in X : f(x) < \infty\}.$$

The set $\mathrm{Dom}\,(f)$ is always convex and we call f *proper* provided $\mathrm{Dom}\,(f)$ is not empty. The functional f is called *uniformly convex* if there exists a continuous and strictly increasing function $\varphi\colon \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with the property $\varphi\,(t) = 0$ implies $t = 0$, such that

$$f\,(\lambda x + (1-\lambda)\,y) + \lambda\,(1-\lambda)\,\varphi\,(\|x-y\|) \le \lambda f\,(x) + (1-\lambda) f\,(y)\,, \qquad (3)$$

for all $\lambda \in (0,1)$ and $x, y \in X$. Of course f uniformly convex implies f strictly convex, which in turn implies f convex. The functional f is *lower semi-continuous* (in short l.s.c.) if for any sequence $(x_k)_{k\in\mathbb{N}} \subset X$ satisfying $x_k \to x$, it holds

$$f\,(x) \le \lim_{k\to\infty}\inf f\,(x_k)\,.$$

It is called *weakly lower semi-continuous* (w.l.s.c.) if above property holds true with $x_k \to x$ replaced by $x_k \rightharpoonup x$. Obviously every w.l.s.c functional is l.s.c. Further, any Banach space norm is w.l.s.c.

The *sub-differential* of a functional $f\colon X \to \overline{\mathbb{R}}$ is the point-to-set mapping $\partial f\colon X \to 2^{X^*}$ defined by

$$\partial f\,(x) := \{\xi \in X^* :\ f\,(x) + \langle \xi, y - x\rangle \le f\,(y) \qquad \text{for all} \ \ y \in X\}\,.$$

Any element in the set $\partial f\,(x)$ is called a *sub-gradient* of f at x. The effective domain of ∂f is the set

$$\mathrm{Dom}\,(\partial f) := \{x \in X :\ \ \partial f\,(x) \neq \emptyset\}\,.$$

It is clear that the inclusion $\mathrm{Dom}\,(\partial f) \subset \mathrm{Dom}\,(f)$ holds whenever f is proper.

Sub-differentiable and convex l.s.c. functionals are strongly connected to each other. In fact, a sub-differentiable functional f is convex and l.s.c. in any open convex set of $\mathrm{Dom}\,(f)$. On the other hand, a proper, convex and l.s.c. functional is always sub-differentiable on its effective domain.

The definition of sub-differential readily yields

$$0 \in \partial f\,(x) \Longleftrightarrow f\,(x) \le f\,(y) \qquad \text{for all} \ \ y \in X.$$

If $f, g\colon X \to \overline{\mathbb{R}}$ are convex functionals and there is a point $x \in \mathrm{Dom}\,(f) \cap \mathrm{Dom}\,(g)$ where f is continuous, then

$$\partial\,(f + g)\,(x) = \partial f\,(x) + \partial g\,(x) \qquad \text{for all} \ \ x \in X. \qquad (4)$$

Moreover, if Y is also a real Banach space, $h\colon Y \to \overline{\mathbb{R}}$ is convex, $b \in Y$, $A\colon X \to Y$ is a bounded linear operator and h is continuous at some point of the range of A, then

$$\partial\,(h\,(\cdot - b))\,(y) = (\partial h)\,(y - b) \qquad \text{and} \qquad \partial\,(h \circ A)\,(x) = A^*\,(\partial h\,(Ax))\,,$$

for all $x \in X$ and $y \in Y$, where $A^*: Y^* \to X^*$ is the Banach-adjoint of A. As a consequence,

$$\partial \left(h \left(A \cdot - b \right) \right) (x) = A^* \left(\partial h \right) (Ax - b) \qquad \text{for all } \ x \in X. \tag{5}$$

If a convex functional $f: X \to \overline{\mathbb{R}}$ is Gâteaux-differentiable at $x \in X$, then f has a unique sub-gradient at x, namely, the Gâteaux-derivative itself: $\partial f(x) = \{\nabla f(x)\}$.

The sub-differential of the convex functional

$$f(x) = \frac{1}{p} \|x\|^p, \quad p > 1, \tag{6}$$

is called the *duality mapping* and is denoted by J_p. It can be shown that for all $x \in X$,

$$J_p(x) = \left\{ x^* \in X^* : \langle x^*, x \rangle = \|x^*\| \, \|x\| \quad \text{and} \quad \|x^*\| = \|x\|^{p-1} \right\}.$$

Thus, the duality mapping has the inner-product-like properties:

$$\langle x^*, y \rangle \leq \|x\|^{p-1} \|y\| \text{ and } \langle x^*, x \rangle = \|x\|^p,$$

for all $x^* \in J_p(x)$. In a Hilbert spaces X, by using the Riesz Representation Theorem, one can prove that $J_2(x) = x$ for all $x \in X$. Further, only in Hilbert spaces J_2 is a linear map.

Banach spaces are classified according with their geometrical characteristics. Many concepts concerning these characteristics are usually defined using the *modulus of convexity* and the *modulus of smoothness*, but most of these definitions can be equivalently stated observing the properties of the functional f defined in (6).[1] This functional is convex and sub-differentiable in any Banach space X. If (6) is Gâteaux-differentiable in the whole space X, this Banach space is called *smooth*. In this case, $J_p(x) = \partial f(x) = \{\nabla f(x)\}$ and therefore, the duality mapping $J_p: X \to X^*$ is single-valued. If the functional f in (6) is Fréchet-differentiable in X, this space is called *locally uniformly smooth* and it is called *uniformly smooth* provided f is uniformly Fréchet-differentiable in bounded sets. As a result, the duality mapping is continuous (resp. uniformly continuous in bounded sets) in locally uniformly smooth (resp. uniformly smooth) spaces. It is immediate that uniform smoothness implies local uniform smoothness, which in turn implies smoothness. Further, none reciprocal is true. Similarly, a Banach space X is called *strictly convex* whenever (6) is a strictly convex functional. Moreover, X is called *uniformly convex* if the functional f in (6) is uniformly convex. It is clear that uniform convexity implies strict convexity. It is well-known that both uniformly smooth and uniformly convex Banach spaces are reflexive.

[1] Normally, the differentiability and convexity properties of this functional are independent of the particular choice of $p > 1$.

Assume f is proper. Then choosing elements $x, y \in X$ with $y \in \mathrm{Dom}\,(\partial f)$, we define the *Bregman distance* between x and y in the direction of $\xi \in \partial f\,(y)$ as

$$D_\xi f\,(x,y) := f\,(x) - f\,(y) - \langle \xi, x - y \rangle\,.$$

Obviously $D_\xi f\,(y,y) = 0$, and since $\xi \in \partial f\,(y)$, it additionally holds $D_\xi f\,(x,y) \geq 0$. Moreover, it is straightforward proving the *Three Points Identity*:

$$D_{\xi_1} f\,(x_2,x_1) - D_{\xi_1} f\,(x_3,x_1) = D_{\xi_3} f\,(x_2,x_3) + \langle \xi_3 - \xi_1, x_2 - x_3 \rangle\,,$$

for all $x_2 \in X$, $x_1, x_3 \in \mathrm{Dom}\,(\partial f)$, $\xi_1 \in \partial f\,(x_1)$ and $\xi_3 \in \partial f\,(x_3)$. Further, the functional $D_\xi f\,(\cdot, y)$ is strictly convex whenever f is strictly convex, and in this case, $D_\xi f\,(x,y) = 0$ iff $x = y$.

When f is the functional defined in (6) and X is a smooth Banach space, the Bregman distance has the special notation $\Delta_p\,(x,y)$, i.e.,

$$\Delta_p\,(x,y) := \frac{1}{p}\,\|x\|^p - \frac{1}{p}\,\|y\|^p - \left\langle J_p\,(y)\,, x - y \right\rangle.$$

Since J_2 is the identity operator in Hilbert spaces, a simple application of the polarization identity shows that $\Delta_2\,(x,y) = \frac{1}{2}\,\|x-y\|^2$ in these spaces.

If $f : X \to \overline{\mathbb{R}}$ is uniformly convex, then for all $y \in X$, $x \in \mathrm{Dom}\,(\partial f)$, $\xi \in \partial f\,(x)$ and $\lambda \in (0,1)$,

$$\begin{aligned} f\,(\lambda x + (1-\lambda)\,y) &\geq f\,(x) + \langle \xi, (\lambda x + (1-\lambda)\,y) - x \rangle \\ &= f\,(x) + (1-\lambda)\,\langle \xi, y - x \rangle\,, \end{aligned}$$

which in view of (3) implies

$$\langle \xi, y - x \rangle + \lambda \varphi\,(\|x-y\|) \leq f\,(y) - f\,(x)\,.$$

Now, letting $\lambda \to 1^-$, we obtain $\varphi\,(\|x-y\|) \leq D_\xi f\,(y,x)$. Analogously, the inequality

$$\varphi\,(\|x-y\|) \leq D_\xi f\,(x,y) \tag{7}$$

holds true for all $x \in X$, $y \in \mathrm{Dom}\,(\partial f)$ and $\xi \in \partial f\,(y)$, whenever f is uniformly convex. In particular, in a smooth and uniformly convex Banach space X, the above inequality reads $\varphi\,(\|x-y\|) \leq \Delta_p\,(x,y)$.

It is well-known that for $1 < p < \infty$, the Lebesgue space $L^p\,(\Omega)$, the Sobolev space $W^{n,p}\,(\Omega)$ and the space of p-summable sequences $\ell^p\,(\mathbb{R})$ are uniformly smooth and uniformly convex Banach spaces.

3 The Iterative Method

In this section we introduce the nonstationary iterated Tikhonov method to solve (2). The method we propose here is in the spirit of the method in [11], with the distinguish feature of using an endogenous strategy for the choice of the Lagrange multipliers λ_k^δ.

Specifically, fixing $r > 0$ and a uniformly convex penalty term f, the iterative method defines sequences (x_k^δ) in X and (ξ_k^δ) in X^* iteratively by

$$\begin{aligned} x_k^\delta &:= \arg\min_{x\in X} \tfrac{\lambda_k^\delta}{r} \left\| Ax - y^\delta \right\|^r + D_{\xi_{k-1}^\delta} f\left(x, x_{k-1}^\delta\right) \\ \xi_k^\delta &:= \xi_{k-1}^\delta - \lambda_k^\delta A^* J_r(Ax_k^\delta - y^\delta), \end{aligned}$$

where the multiplier λ_k^δ will be determined using only information about A, δ, y^δ and x_{k-1}^δ.

Our strategy for selecting the Lagrange multipliers is inspired in the recent work [1], where it was proposed an endogenous strategy for the choice of the Lagrange multiplier in the iterative method for solving (2), when X and Y are Hilbert spaces. This method is based on successive orthogonal projection methods onto a family of shrinking, separating convex sets. Specifically, the iterative method in [1] obtains the new iterate projecting the current one onto a levelset of the residual function, whose level belongs to a range defined by the current residual and by the noise level. Further, the admissible Lagrange multipliers (in each iteration) shall belong to a non-degenerate interval.

With the view to extend this framework to Banach space setting we are forced to work with Bregman distance and *Bregman projections*. This is due to the well-known fact that in Banach spaces the *metric projection* onto a convex and closed set C, defined as $P_C(x) = \mathit{arg\,min}_{z\in C}\|z - x\|^2$, loses the decreasing distance property of the orthogonal projection in Hilbert spaces. In order to recover this property, one should minimize in Banach spaces the Bregman distance, instead of the norm-induced distance.

In what follows we assume the following conditions:

(A.1) There exist an element $x^\star \in X$ such that $Ax^\star = y$, where $y \in R(A)$ is the exact data.

(A.2) f is a l.s.c. function.

(A.3) f is a uniformly convex function.

(A.4) X and Y are reflexive Banach spaces and Y is smooth.

We define Ω_μ^r, the μ-levelset of the residual functional $\|Ax - y^\delta\|$, as

$$\Omega_\mu^r := \left\{ x \in X \, : \, \frac{1}{r}\|Ax - y^\delta\|^r \le \frac{1}{r}\mu^r \right\} .$$

We observe that since A is a continuous linear operator it follows that Ω_μ^r is closed and convex.

Now, given $\hat{x} \in \mathrm{Dom}(\partial f)$ and $\xi \in \partial f(\hat{x})$, we can define the *Bregman projection* of $\hat{x}$ onto Ω_μ^r, as a solution of the minimization problem

$$\begin{cases} \min & D_\xi f(x, \hat{x}) \\ \text{s.t.} & \frac{1}{r}\|Ax - y^\delta\|^r \leq \frac{1}{r}\mu^r. \end{cases} \tag{8}$$

It is clear that a solution of the above problem depends on the sub-gradient ξ. Furthermore, since $D_\xi f(\cdot, \hat{x})$ is strictly convex, which follows from the uniformly convexity of f, problem (8) has at most one solution.

The fact that the projection is well defined when $\mu > \delta$, and in this case we can set $P^f_{\Omega_\mu^r}(\hat{x}) := \arg\min_{x \in \Omega_\mu^r} D_\xi f(x, \hat{x})$, is a consequence of the following lemma.

Lemma 1 *If $\mu > \delta$ then problem (8) has a solution.*

Proof Hypothesis (A.1), together with Eq. (1) and the assumption that $\mu > \delta$, imply that the feasible set of problem (8), i.e. the set Ω_μ^r, is nonempty.

By conditions (A.2) and (A.3) we have that $D_\xi f(\cdot, \hat{x})$ is proper, convex and l.s.c. Furthermore, relation (7) implies that $D_\xi f(\cdot, \hat{x})$ is a coercive function. Hence, the lemma follows using that X is a reflexive space and applying [2, Corollary 3.23]. □

It is easy to see that if $0 \leq \mu' \leq \mu$ then $\Omega_{\mu'}^r \subseteq \Omega_\mu^r$, and $A^{-1}(y) \subset \Omega_\mu^r$ for all $\mu \geq \delta$. Furthermore, with the available information of the solution set of (2), Ω_δ^r is the set of best possible approximate solution for this inverse problem. However, since problem (8) may be ill-posed when $\mu = \delta$, our best choice is to generate x_k^δ from $x_{k-1}^\delta \notin \Omega_\delta^r$ as a solution of problem (8), with $\hat{x} = x_{k-1}^\delta$ and $\mu = \mu_k$ such that we guarantee a reduction of the residual norm while preventing ill-posedness of (8).

For this purpose, we now analyze the minimization problem (8) by means of Lagrange multipliers. The Lagrangian function associated with problem (8) is

$$\mathscr{L}(x, \lambda) = \frac{\lambda}{r}(\|Ax - y^\delta\|^r - \mu^r) + D_\xi f(x, \hat{x}).$$

We observe that for each $\lambda > 0$ the function $\mathscr{L}(\cdot, \lambda) : X \to \overline{\mathbb{R}}$ is l.s.c. and convex. For any $\lambda > 0$ define the following functions

$$\pi(\hat{x}, \lambda) = \arg\min_{x \in X} \mathscr{L}(x, \lambda), \qquad G_{\hat{x}}(\lambda) = \|A\pi(\hat{x}, \lambda) - y^\delta\|^r. \tag{9}$$

The next lemma gives a classical Lagrange multiplier result for problem (8), which will be useful for formulating the nIT method.

Lemma 2 *Suppose that* $\|A\hat{x}-y^\delta\| > \mu > \delta$*, then the following assertions are equivalent*

1. *x is a solution of (8);*
2. *there exists $\lambda^* > 0$ such that $x = \pi(\hat{x}, \lambda^*)$ and $G_{\hat{x}}(\lambda^*) = \mu^r$.*

Proof By (1), hypothesis (A.1) and the assumption $\mu > \delta$, we have that $x^\star \in X$ is such that

$$\|Ax^\star - y^\delta\|^r < \mu^r.$$

Inequality above implies the Slater condition for problem (8). Thus, using that A is continuous and $D_\xi f(\cdot, \hat{x})$ is l.s.c., we have that x is a solution of (8) if and only if there exists $\lambda \in \mathbb{R}$ such that the point (x, λ) satisfies the *Karush-Kuhn-Tucker* (KKT) conditions for this minimization problem, see [12].

The KKT conditions [12] for (8) are

$$\lambda \geq 0, \qquad G_{\hat{x}}(\lambda) \leq \mu^r, \qquad \lambda(G_{\hat{x}}(\lambda) - \mu^r) = 0, \qquad 0 \in \partial_x \mathscr{L}(x, \lambda).$$

If we suppose that $\lambda = 0$ in relations above, then the definition of the Lagrangian function, together with the strictly convexity of $D_\xi f(\cdot, \hat{x})$, implies that $\hat{x}$ is the unique minimizer of $\mathscr{L}(\cdot, 0)$. Since $\|A\hat{x} - y^\delta\| > \mu$ we conclude that the pair $(\hat{x}, 0)$ does not satisfy the KKT conditions. Hence, we have $\lambda > 0$ and $G_{\hat{x}}(\lambda) - \mu^r = 0$. We conclude the lemma using the definition of $\pi(\hat{x}, \lambda)$. □

We are now ready to formulate the nIT method for solving (2).

Properties (4) and (5), together with the definition of the duality mapping, imply that the point $x_k^\delta \in X$ minimizes the *Tikhonov functional*

$$T_\lambda^\delta(x) := \frac{\lambda_k^\delta}{r}\left\|Ax - y^\delta\right\|^r + D_{\xi_{k-1}^\delta} f\left(x, x_{k-1}^\delta\right),$$

if and only if

$$0 \in \lambda_k^\delta A^* J_r\left(Ax_k^\delta - y^\delta\right) + \partial f\left(x_k^\delta\right) - \xi_{k-1}^\delta. \tag{10}$$

Hence, since Y is a smooth Banach space, we have that the duality mapping J_r is single valued and

$$\xi_{k-1}^\delta - \lambda_k^\delta A^* J_r\left(Ax_k^\delta - y^\delta\right) \in \partial f\left(x_k^\delta\right).$$

Therefore, ξ_k^δ in step 3.2 of Algorithm 1 is well defined and it is a sub-gradient of f at x_k^δ.

Algorithm 1 The iterative method

```
[1] choose an initial guess $x_0 \in X$ and $\xi_0 \in \partial f(x_0)$;
[2] choose $\eta \in (0,1)$, $\tau > 1$ and set $k := 0$;
[3] while $\big(\|Ax_k^\delta - y^\delta\| > \tau\delta\big)$ do
    [3.1] $k := k+1$;
    [3.2] compute $\lambda_k^\delta, x_k^\delta$ such that $x_k^\delta = \arg\min_{x\in X} \frac{\lambda_k^\delta}{r}\|Ax - y^\delta\|^r + D_{\xi_{k-1}^\delta} f(x, x_{k-1}^\delta)$,
          and $\delta^r < G_{x_{k-1}^\delta}(\lambda_k^\delta) \le \big(\eta\delta + (1-\eta)\|Ax_{k-1}^\delta - y^\delta\|\big)^r$.
          Set $\xi_k^\delta = \xi_{k-1}^\delta - \lambda_k^\delta A^* J_r(Ax_k^\delta - y^\delta)$.
```

4 Algorithms and Numerical Implementation

4.1 Determining the Lagrange Multipliers

As before, we consider the function $G_{\hat{x}}(\lambda) = \left\|Ax_\lambda - y^\delta\right\|^r$, where $x_\lambda = \pi(\lambda, \hat{x})$ represents the minimizer of the Tikhonov functional

$$T_\lambda(x) = \frac{\lambda}{r}\left\|Ax - y^\delta\right\|^r + D_\xi f(x, \hat{x}). \tag{11}$$

In order to determine the Lagrange multiplier in the iteration k, we need to calculate $\lambda_k > 0$ such that $G_{x_{k-1}}(\lambda_k) \in [a_k, b_k]$, where

$$a_k := \delta^r \quad \text{and} \quad b_k := (\eta\delta + (1-\eta)\|Ax_{k-1} - y^\delta\|)^r,$$

with $0 < \eta < 1$ pre-defined.

For doing that, we have employed three different methods: the well-known secant and Newton methods and a third strategy, called *adaptive method*, which we explain now: fix $\sigma_1, \sigma_2 \in (0,1)$, $c_1 > 1$ and start with $\lambda_0^\delta > 0$. In the k-th iteration, $k \ge 1$, we define $\lambda_k^\delta = c_k\lambda_{k-1}^\delta$, where

$$c_k = \begin{cases} c_{k-1}\sigma_1, & \text{if } G_{x_{k-2}}(\lambda_{k-1}^\delta) < a_{k-1} \\ c_{k-1}/\sigma_2, & \text{if } G_{x_{k-2}}(\lambda_{k-1}^\delta) > b_{k-1} \\ c_{k-1}, & \text{otherwise} \end{cases}, \text{ for } k \ge 2.$$

The idea behind the adaptive method is observing the behavior of the residual in last iterations and trying to determine how much the Lagrange multiplier should be increased in the next iteration. For example, the residual $G_{x_{k-2}}(\lambda_{k-1}^\delta) = \|Ax_{k-1} - y^\delta\|^r$ lying on the left of the target interval $[a_{k-1}, b_{k-1}]$, means that λ_{k-1}^δ was too large. We thus multiply the number c_{k-1} by a number $\sigma_1 \in (0,1)$ in order to reduce the speed of growing of the Lagrange multipliers λ_k^δ, trying to hit the target in the next iteration.

Although the Newton method is efficient, in the sense that it normally finds a good approximation for the Lagrange multiplier in very few steps, it has the drawback of demanding the differentiability of the Tikhonov functional, and therefore it cannot be applied in all situations.

Because it does not require the evaluation of derivatives, the secant method can be used even for a nonsmooth Tikhonov functional. A disadvantage of this method is the high computational effort required to perform it.

Among these three possibilities, the adaptive strategy is the cheapest one, since it only demands one minimization of the Tikhonov functional per iteration. Further, this simple strategy does not request the derivative of this functional, which makes it fit in a large range of applications.

Notice that this third strategy may generate a λ_k^δ such that $G_{x_{k-1}}(\lambda_k^\delta) \notin [a_k, b_k]$ in some iterative steps. This is the reason for correcting the factors c_k in each iteration. In our numerical experiments, the condition $G_{x_{k-1}}(\lambda_k^\delta) \in [a_k, b_k]$ was satisfied in almost all steps (see the slope of the green curve on Fig. 3; bottom picture).

4.2 Minimization of the Tikhonov Functional

In our numerical experiments, we are interested in solving the inverse problem (2), where the linear and bounded operator $A : L^p(\Omega) \to L^2(\Omega)$, $1 < p < \infty$, the noisy data y^δ and the noise level $\delta > 0$ are known.

In order to apply the iterative method (Algorithm 1), a minimizer of the Tikhonov functional (11) needs to be calculated on each iteration. Minimizing this functional can be itself a very challenging task. We have used two algorithms for achieving this goal in our numerical experiments: (1) the Newton method was used for minimizing this functional in the case $p \neq 2$ and with a smooth function f, which induces the Bregman distance in the penalization term. (2) The so called ADMM method was employed in order to minimize the Tikhonov functional for the case $p = 2$ (Hilbert space) and a nonsmooth functional f. In the following, we explain the details.

First we consider the Newton method. Define the Bregman distance induced by the norm-functional $f(g) := \frac{1}{p}\|g\|_{L^p}^p$, $1 < p < \infty$, which leads to the smooth penalization term $D_\xi f(g,h) = \Delta_p(g,h)$, see Sect. 2. The resultant Tikhonov functional is

$$T_\lambda(g) = \frac{\lambda}{2}\left\|Ag - y^\delta\right\|^2 + \Delta_p(g, g_{k-1}),$$

where g_{k-1} is the current iterate.[2] In this case, the optimality condition (10) reads:

$$F(\overline{g}) = \lambda A^* y^\delta + J_p(g_{k-1}), \tag{12}$$

where $\overline{g} \in L^p(\Omega)$ is the minimizer of the above Tikhonov functional and $F(g) := \lambda A^* A g + J_p(g)$.

[2] Here (2) is replaced by $Ag = y^\delta$.

In order to apply the Newton method to the nonlinear equation (12), one needs to evaluate the derivative of F, which (if it exists) is given by $F'(g) = \lambda A^*A + J_p'(g)$. Next, we prove that J_p is at least Gâteaux-differentiable in $L^p(\Omega)$, if $p \geq 2$. Further, we present an explicit expression for $J_p'(g)$, which will be used later in our numerical experiments.

The key for finding a formula for $J_p'(g)$ is observing the differentiability of the function $\gamma : \mathbb{R} \to \mathbb{R}$, $x \mapsto \frac{1}{p}|x|^p$. This function is differentiable in $\mathbb{R}$ whenever $p > 1$, and in this case,

$$\gamma'(x) = |x|^{p-1}\operatorname{sign}(x), \qquad \text{where } \operatorname{sign}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases}. \tag{13}$$

Furthermore, γ is twice differentiable in $\mathbb{R}$ if $p \geq 2$, with derivative given by

$$\gamma''(x) = (p-1)|x|^{p-2}. \tag{14}$$

This formula still holds true for $1 < p < 2$, but only in $\mathbb{R}\setminus\{0\}$. In this case, $\gamma''(0)$ does not exist and $\gamma''(x)$ grows to infinity as x approaches to zero.

Since $J_p(g) = \left(\frac{1}{p}\|g\|_{L^p}^p\right)'$ can be identified with (see [3])

$$J_p(g) = |g|^{p-1}\operatorname{sign}(g), \tag{15}$$

which looks very similar to γ' in (13), the bounded linear operator $J_p'(g) : L^p(\Omega) \to L^{p^*}(\Omega)$ is similar to γ'' in (14). Indeed, for any fixed $g \in L^p(\Omega)$, with $p \geq 2$, we have

$$\left\langle J_p'(g), h\right\rangle = \left\langle (p-1)|g|^{p-2}, h\right\rangle, \tag{16}$$

for every $h \in L^p(\Omega)$, where the linear operator $(p-1)|g|^{p-2}$ is understood pointwise: $h \mapsto (p-1)|g(\cdot)|^{p-2}h(\cdot)$. This ensures that J_p is Gâteaux-differentiable in $L^p(\Omega)$ and its derivative J_p' can be identified with $(p-1)|\cdot|^{p-2}$.

In the discretized setting, $J_p'(g)$ is a diagonal matrix whose i-th element on its diagonal is $(p-1)|g(x_i)|^{p-2}$, with x_i being the i-th point of the chosen mesh.

In our numerical simulations, we consider the situation where the sought solution is sparse and, therefore, the case $p \approx 1$ is of our interest. We stress the fact that Eq. (14) holds true even for $1 < p < 2$ whenever $x \neq 0$. Using this fact, one can prove that (16) holds true for these values of p, for instance, if g does not change signal in Ω (i.e., $g > 0$ or $g < 0$ in Ω) and the direction h is a bounded function in this set. However, these strong hypotheses are very difficult to check, and even if they are satisfied, we still expect having stability problems for inverting the matrix $F'(g)$ in (12) if the function g has a small value in some point of the mesh, because the function in (14) satisfies $\gamma''(x) \to \infty$ as $x \to 0$. In order to avoid this kind of problem in our numerical experiments, we have replaced the i-th element on the diagonal of the matrix $J_p'(g)$ by $\max\left\{(p-1)\,|g(x_i)|^{p-2}, 10^6\right\}$.

The second method that we used in our experiments was the well-known *Alternating Direction Method of Multipliers* (ADMM), which has been implemented to minimize the Tikhonov functional associated with the inverse problem $Ax = y^\delta$, where $X = Y = \mathbb{R}^n$, $A : \mathbb{R}^n \to \mathbb{R}^n$, and $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a nonsmooth function.

ADMM is an optimization scheme for solving linearly constrained programming problems with decomposable structure [5], which goes back to the works of Glowinski and Marrocco [8], and of Gabay and Mercier [7]. Specifically, this algorithm solves problems in the form:

$$\min_{(x,z)} \{\varphi(x) + \phi(z) \;:\; Mx + Bz = d\}, \tag{17}$$

where $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\phi : \mathbb{R}^m \to \overline{\mathbb{R}}$ are convex proper l.s.c. functions, $M : \mathbb{R}^n \to \mathbb{R}^l$ and $B : \mathbb{R}^m \to \mathbb{R}^l$ are linear operators, and $d \in \mathbb{R}^l$.

ADMM solves the coupled problem (17) performing a sequences of steps that decouple functions φ and ϕ, making it possible to exploit the individual structure of these functions. It can be interpreted in terms of alternating minimization, with respect to x and z, of the augmented Lagrangian function associated with problem (17). Indeed, ADMM consists of the iterations

$$\begin{aligned} x_{k+1} &= \arg\min_x \mathscr{L}_\rho(x, z_k, u_k) \\ z_{k+1} &= \arg\min_z \mathscr{L}_\rho(x_{k+1}, z, u_k) \\ u_{k+1} &= u_k + \rho(Mx_{k+1} + Bz_{k+1} - d), \end{aligned}$$

where $\rho > 0$ and $\mathscr{L}_\rho$ is the augmented Lagrangian function

$$\mathscr{L}_\rho(x, z, u) := \varphi(x) + \phi(z) + \langle u, Mx + Bz - d\rangle + \frac{\rho}{2}\|Mx + Bz - d\|_2^2 .$$

The convergence results for ADMM guarantee, under suitable assumptions, that the sequences (x_k), (z_k) and (u_k), generated by the method, are such that $Mx_k + Bz_k - d \to 0$, $\varphi(x_k) + \phi(z_k) \to s^\star$ and $u_k \to u^\star$, where $s^\star$ is the optimal value of problem (17) and $u^\star$ is a solution of the dual problem associated with (17).

For minimizing the Tikhonov functional using ADMM we introduce an additional decision variable z such that problem

$$\min_{x \in X} T^\delta_{\lambda^\delta_k}(x)$$

is rewritten into the form of (17). The specific choice of the functions φ, ϕ and the operators M and B is problem dependent. For a concrete example, please see Sect. 5.2. This allows us to exploit the special form of the functional $T^\delta_{\lambda^\delta_k}$ and pose the problem in a more suitable manner to solve it numerically.

In our numerical simulations we stopped ADMM when $\|Mx_k\| + Bz_k - d$ was less than a prefixed tolerance.

5 Numerical Experiments

5.1 Deconvolution

The first application considered here is the deconvolution problem modeled by the linear integral operator

$$A\,x := \int_0^1 K(s,t)\,x(t)\,dt \;=\; y(s)\,,$$

where the kernel K is the continuous function defined by

$$K(s,t) \;=\; \begin{cases} 49s(1-t)\,, & s \le t \\ 49t(1-s)\,, & s > t \end{cases}\,.$$

This benchmark problem is considered in [10]. There, it is observed that $A : L^p[0,1] \to C[0,1]$ is continuous and bounded for $1 \le p \le \infty$. Thus $A : L^p[0,1] \to L^r[0,1]$ is compact, for $1 \le r < \infty$.

In our experiment, A is replaced by the discrete operator A_d, where the above integral is computed using a quadrature formula (trapezoidal rule) over an uniform partition of the interval $[0,1]$ with 400 nodes.

The exact solution of the discrete problem is the vector $x^\star \in \mathbb{R}^{400}$ with $x^\star(48) = 2$, $x^\star(200) = 1.5$, $x^\star(270) = 1.75$ and $x^\star(i) = 0$, elsewhere.

We compute $y = A_d x^\star$, the exact data, and add random Gaussian noise to $y \in \mathbb{R}^{400}$ to get the noisy data y^δ satisfying $\| y - y^\delta \|_Y \le \delta$.

We follow [10] in the experimental setting and choose $\delta = 0.0005$, $\tau = 1.001$ (discrepancy principle), and $Y = L^2$. For the parameter space, two distinct choices are considered, namely $X = L^{1.001}$ and $X = L^2$.

Numerical results are presented in Fig. 1.[3] The following methods are implemented:

- (Blue) L^2-penalization, Geometric sequence;
- (Green) L^2-penalization, Secant method;
- (Red) $L^{1.001}$-penalization, Geometric sequence;
- (Pink) $L^{1.001}$-penalization, Secant method;
- (Black) $L^{1.001}$-penalization, Newton method.

The six pictures in Fig. 1 represent:

[Top] Iteration error in L^2-norm (left)[4]; residual in L^2-norm (right);
[Center] Number of linear systems/step (left); Lagrange multipliers (right);

[3] For simplicity, all legends in this figure refers to the space L^1; however, we used $p = 1.001$ in the computations.

[4] For the purpose of comparison, the iteration error is plotted in the in L^2-norm for both choices of the parameter space $X = L^2$ and $X = L^{1.001}$.

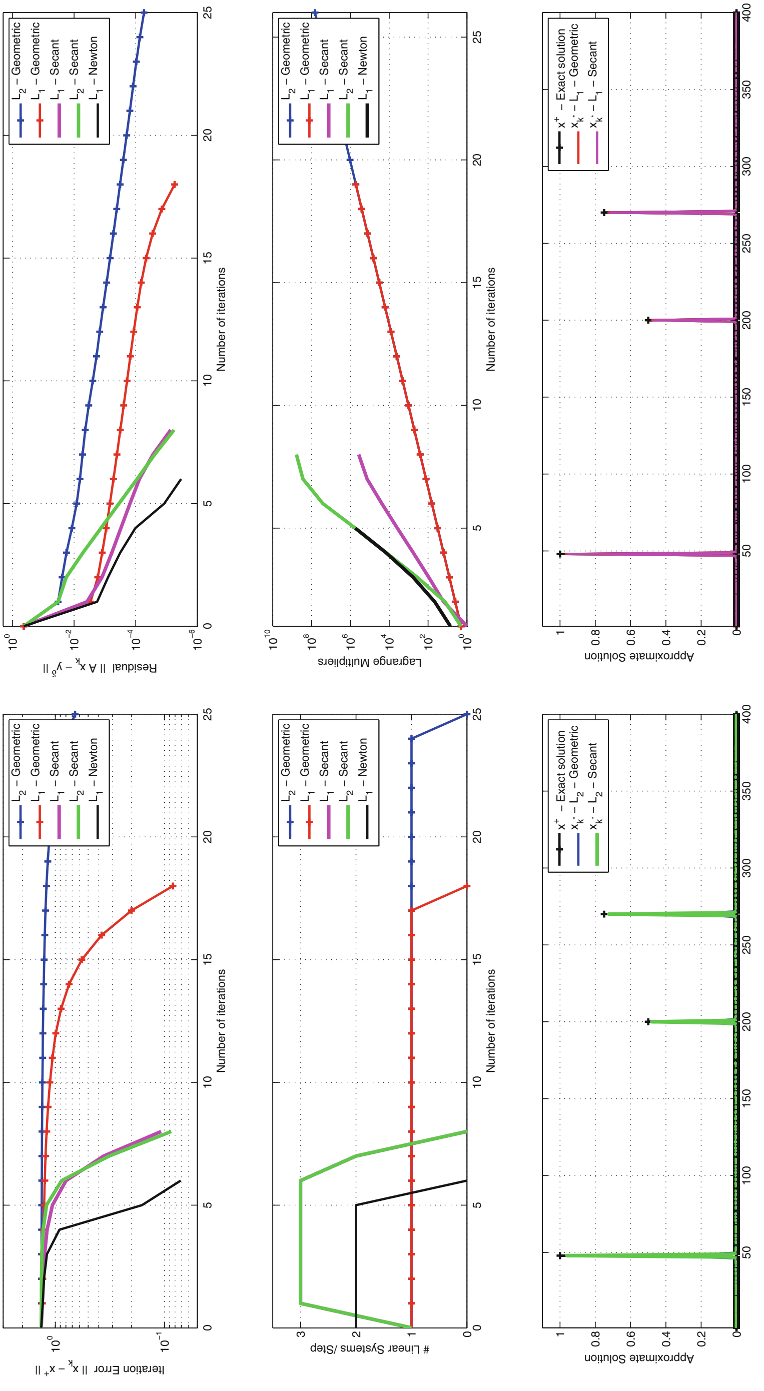


Fig. 1 Deconvolution problem: numerical experiments

[Bottom] Exact solution and reconstructions with L^2-penalization (left); exact solution and reconstructions with $L^{1.001}$-penalization (right).

5.2 Image Deblurring

The second application of the nIT method that we consider is the image deblurring problem. This is a finite dimensional problem with spaces $X = \mathbb{R}^n \times \mathbb{R}^n$ and $Y = \mathbb{R}^n \times \mathbb{R}^n$. The vector $x \in X$ represents the pixel values of the original image to be restored, and $y \in Y$ contains the pixel values of the observed blurred image. In practice, only noisy blurred data $y^\delta \in Y$ satisfying (1) is available. The linear transformation A represents some blurring operator.

For our numerical simulations we consider the situation where the blur of the image is modeled by a space invariant point spread function (PSF). We use the 256×256 *Cameraman* test image, and y^δ is obtained adding artificial noise to the exact data $Ax = y$ (here A is the convolution operator corresponding to the PSF).

For this problem we implemented the nIT method with two different penalization terms, namely $f(x) = \|x\|_2^2$ (L^2 penalization) and $f(x) = \frac{\mu}{2}\|x\|_2^2 + TV(x)$ ($L^2 + TV$ penalization). Here $\mu > 0$ is a regularization parameter and $TV(x) = \|\nabla x\|_1$ is the *total variation* norm of x, where $\nabla : \mathbb{R}^n \times \mathbb{R}^n \to (\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^n \times \mathbb{R}^n)$ is the *discrete gradient* operator.

We minimize the Tikhonov functional associated with the $L^2 + TV$ penalization term using the ADMM described in Sect. 4. Specifically, if $f(x) = \frac{\mu}{2}\|x\|_2^2 + \|\nabla x\|_1$, then on each iteration we need to solve

$$\min_{x\in X} \frac{\lambda_k^\delta}{2} \left\|Ax - y^\delta\right\|^2 + \frac{\mu}{2}\|x - x_{k-1}^\delta\|^2 + \|\nabla x\|_1 - \|\nabla x_{k-1}^\delta\|_1 - \left\langle \xi_{k-1}^\delta, x - x_{k-1}^\delta \right\rangle .$$

To use ADMM we sate this problem into the form of problem (17) defining $z = \nabla x$, $\varphi(x) := \frac{\lambda_k^\delta}{2}\|Ax - y^\delta\|^2 + \frac{\mu}{2}\|x - x_{k-1}^\delta\|^2 - \left\langle \xi_{k-1}^\delta, x - x_{k-1}^\delta \right\rangle$, $\phi(z) = \|z\|_1 - \|\nabla x_{k-1}^\delta\|_1$, $M = -\nabla$, $B = I$ and $d = 0$.

In the experiments we choose $\mu = 10^{-4}$, $\delta = 0.00001$ and $\tau = 1.5$. Moreover, we take as initial guesses $x_0 = y^\delta$ and $\xi_0 = \nabla^*(sign(\nabla x_0))$.

Figure 2 shows the recovered images using the two penalization terms, and the different strategies we considered for choosing the Lagrange multipliers.

Figure 3 presents some numerical results. We implemented for this example the following methods:

- (Blue) L^2-penalization, Geometric sequence;
- (Red) $L^2 + TV$-penalization, Geometric sequence;
- (Pink) $L^2 + TV$-penalization, Secant method;
- (Green) $L^2 + TV$-penalization, Adaptive method.

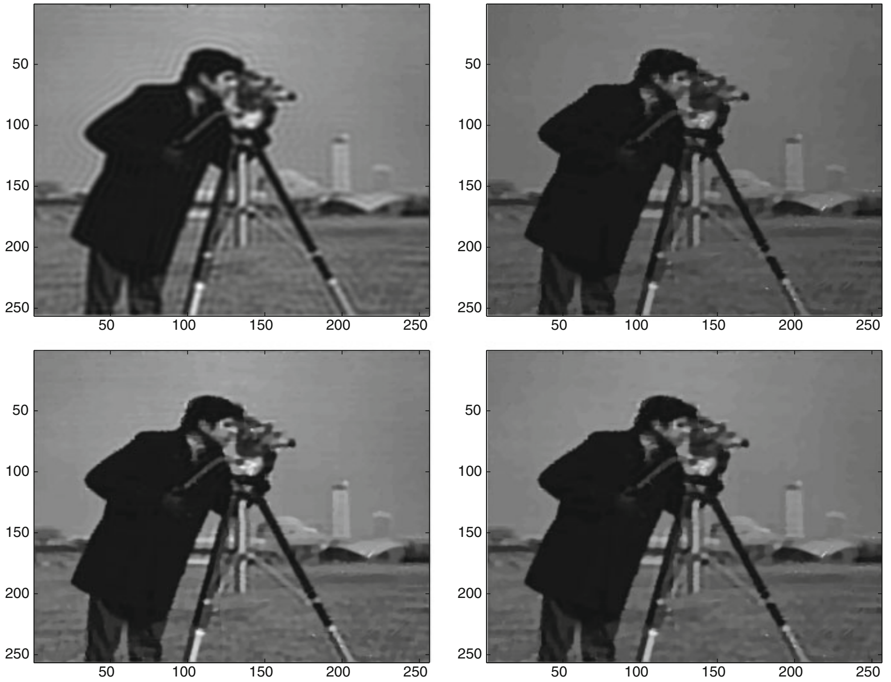


Fig. 2 Image deblurring problem: (top left) Geometric sequence, L^2 penalization; (top right) Geometric sequence, L^2 + TV penalization; (bottom left) Secant method, L^2 + TV penalization; (bottom right) Adaptive method, L^2 + TV penalization

The four pictures in Fig. 3 represent:

[Top] Iteration error $\|x^\star - x_k^\delta\|$;
[Center top] Residual $\|Ax_k^\delta - y^\delta\|$;
[Center bottom] Number of linear systems solved in each step;
[Bottom] Lagrange multiplier λ_k^δ.

6 Conclusions

In this chapter we propose a novel nonstationary iterated Tikhonov (nIT) type method for obtaining stable approximate solutions to ill-posed operator equations modeled by linear operators acting between Banach spaces.

The novelty of our approach consists in defining strategies for choosing a sequence of regularization parameters (Lagrange multipliers) for the nIT method.

The Lagrange multipliers are chosen (a posteriori) in order to enforce a fast decay of the residual functional (see Algorithm 1 and Sect. 4.1). The computation of these multipliers is performed by means of three distinct methods: (1) a secant

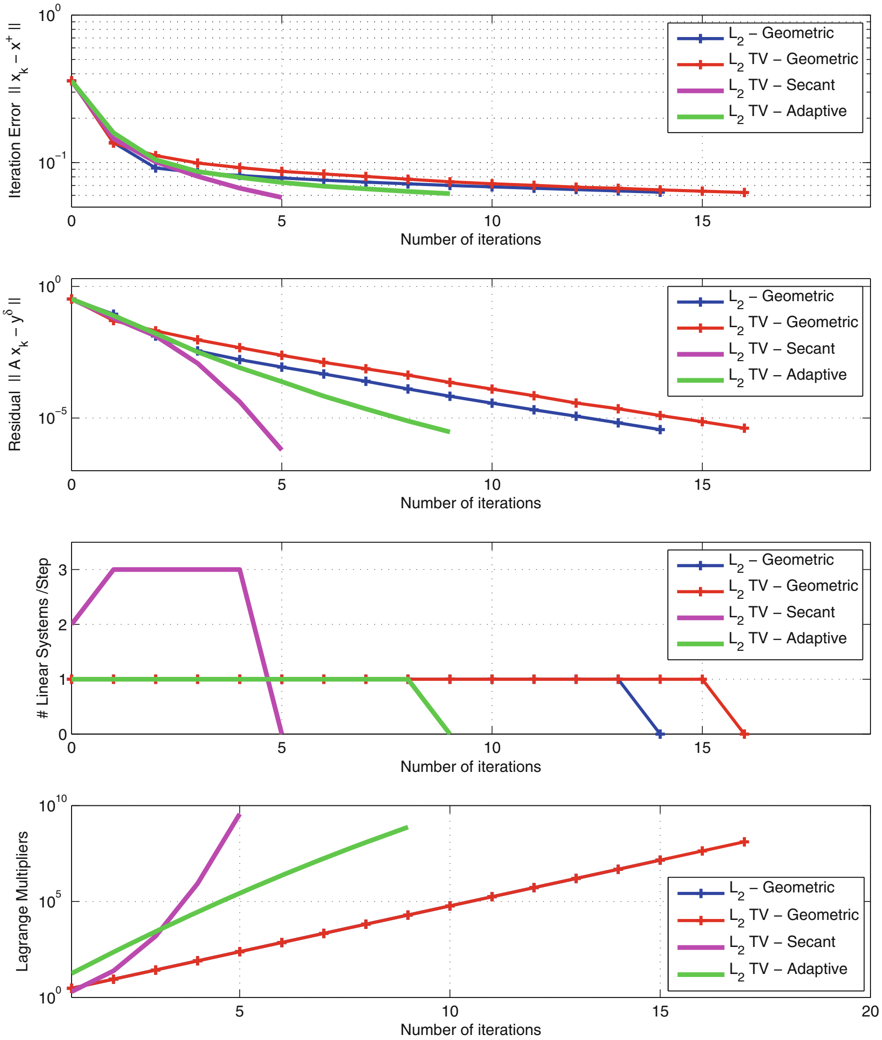


Fig. 3 Image deblurring problem: numerical experiments

type method; (2) a Newton type method; (3) an adaptive method using a geometric sequence with non-constant growth rate, where the rate is updated after each step.

The computation of the iterative step of the nIT method requires the minimization of a Tikhonov type Functional (see Sect. 4.2). This task is solved here using two distinct methods: (1) in the case of smooth penalization and Banach parameter-spaces the optimality condition (related to the Tikhonov functional) leads to a nonlinear equation, which is solved using a Newton type method; (2) in the case of nonsmooth penalization and Hilbert parameter-space, the ADMM method is used for minimizing the Tikhonov functional.

What concerns the Deconvolution problem in Sect. 5.1[5]:

- The secant and the Newton methods produce a sequence of multipliers with faster growth, when compared to the geometric (a priori) choice of multipliers.
- The fact above is observed in both parameter spaces L^2 and $L^{1.001}$.
- The secant and the Newton methods converge within fewer iterations than the geometric choice of multipliers.
- The numerical effort required by the secant type method is similar to the one required by the geometric choice of multipliers.
- The Newton method requires the smallest amount of computational effort.
- As expected, the sparse solution $x^\star$ is better approximated by the methods operating in the $L^{1.001}$ parameter-space.

What concerns the Deblurring problem in Sect. 5.2[6]:

- The secant and the adaptive methods produce a sequence of multipliers with faster growth, when compared to the geometric (a priori) choice of multipliers.
- The secant and the adaptive methods converge within fewer iterations.
- The numerical effort required by the secant type method is similar to the one required by the geometric choice of multipliers.
- The adaptive method requires the smallest amount of computational effort.
- The first reconstructed image (L^2 penalization) differs from the other three reconstructions ($L^2 + TV$ penalization), which produce images with sharper edges and better defined contours.

Acknowledgements F. Margotti is supported by CAPES and IMPA. A. Leitão acknowledges support from CNPq (grant 309767/2013-0), and from the AvH Foundation.

References

1. R. Boiger, A. Leitão, B.F. Svaiter, Endogenous strategies for choosing the Lagrange multipliers in nonstationary iterated Tikhonov method for ill-posed problems (2016, submitted)
2. H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext (Springer, New York, 2011)
3. I. Cioranescu, *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*. Mathematics and Its Applications, vol. 62 (Kluwer Academic, Dordrecht, 1990)
4. A. De Cezaro, J. Baumeister, A. Leitão, Modified iterated Tikhonov methods for solving systems of nonlinear ill-posed equations. Inverse Prob. Imaging **5**(1), 1–17 (2011)
5. J. Eckstein, D.P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. Ser. A **55**(3), 293–318 (1992)
6. H.W. Engl, M Hanke, A Neubauer, *Regularization of Inverse Problems*. Mathematics and Its Applications, vol. 375 (Kluwer Academic, Dordrecht, 1996)

[5] In this situation we have smooth penalization terms and Banach parameter-spaces.

[6] In this situation we have nonsmooth penalization terms and Hilbert parameter-spaces.

7. D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
8. R. Glowinski, A. Marrocco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér. **9**(R-2), 41–76 (1975)
9. M. Hanke, C.W. Groetsch, Nonstationary iterated Tikhonov regularization. J. Optim. Theory Appl. **98**(1), 37–53 (1998)
10. Q. Jin, L. Stals, Nonstationary iterated Tikhonov regularization for ill-posed problems in Banach spaces. Inv. Probl. **28**(3), 104011 (2012)
11. Q. Jin, M. Zhong, Nonstationary iterated Tikhonov regularization in Banach spaces with uniformly convex penalty terms. Numer. Math. **127**(3), 485–513 (2014)
12. R.T. Rockafellar, *Conjugate Duality and Optimization* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974). Lectures given at the Johns Hopkins University, Baltimore, MD, June, 1973. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, vol. 16
13. T. Schuster, B. Kaltenbacher, B. Hofmann, K.S. Kazimierski, *Regularization Methods in Banach Spaces* (de Gruyter, Berlin, 2012)

The Product Midpoint Rule for Abel-Type Integral Equations of the First Kind with Perturbed Data

Robert Plato

Abstract We consider the regularizing properties of the product midpoint rule for the stable solution of Abel-type integral equations of the first kind with perturbed right-hand sides. The impact of continuity and smoothness properties of solutions on the convergence rates is described in detailed manner by using a scale of Hölder spaces. In addition, correcting starting weights are introduced to get rid of undesirable initial conditions. The proof of the inverse stability of the quadrature weights relies on Banach algebra techniques. Finally, numerical results are presented.

1 Introduction

1.1 Preliminary Remarks

In this contribution we consider linear Abel-type integral equations of the following form,

$$(Au)(x) = \frac{1}{\Gamma(\alpha)} \int_0^x (x-y)^{\alpha-1} k(x,y) u(y)\, dy = f(x) \ \text{ for } \ 0 \le x \le a, \tag{1}$$

with $0 < \alpha < 1$ and $a > 0$, and with a sufficiently smooth kernel function $k : \{\,(x,y) \in \mathbb{R}^2 \ \mid \ 0 \le y \le x \le a\,\} \to \mathbb{R}$, and Γ denotes Euler's gamma function. Moreover, the function $f : [0,a] \to \mathbb{R}$ is approximately given, and a function $u : [0,a] \to \mathbb{R}$ satisfying Eq. (1) is to be determined.

R. Plato (✉)
Department of Mathematics, University of Siegen, Walter-Flex-Str. 3, 57068 Siegen, Germany
e-mail: plato@mathematik.uni-siegen.de

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_11

In the following we suppose that the kernel function does not vanish on the diagonal $0 \le x = y \le a$, and without loss of generality we may assume that

$$k(x, x) = 1 \ \text{ for } \ 0 \le x \le a \tag{2}$$

holds.

There exist many quadrature methods for the approximate solution of Eq. (1), see e.g., Brunner/van der Houwen [4], Linz [18], and Hackbusch [13]. One of these methods is the product midpoint rule which is considered in detail, e.g., in Weiss and Anderssen [30] and in Eggermont [8], see also [18, Section 10.4].

In the present text we investigate, for perturbed right-hand sides in Eq. (1), the regularizing properties of the product midpoint rule, and we also consider a modification of this method. Continuity and smoothness is classified in terms of Hölder continuity of the solution and its derivative, respectively. We also give a new proof of the inverse stability of the quadrature weights which relies on Banach algebra techniques and may be of independent interest. Finally, some numerical illustrations are presented.

1.2 The Abel Integral Operator

As a first step we consider in (1) the special situation $k \equiv 1$. For technical reasons we allow arbitrary intervals $[0, b]$ with $0 < b \le a$ instead of the fixed interval $[0, a]$. The resulting integral operator is the Abel integral operator

$$(\mathscr{V}^\alpha \varphi)(x) = \frac{1}{\Gamma(\alpha)} \int_0^x (x-y)^{\alpha-1} \varphi(y)\, dy \ \text{ for } \ 0 \le x \le b, \tag{3}$$

where $\varphi : [0, b] \to \mathbb{R}$ is supposed to be a piecewise continuous function. One of the basic properties of the Abel integral operator is as follows,

$$(\mathscr{V}^\alpha y^q)(x) = \tfrac{\Gamma(q+1)}{\Gamma(q+1+\alpha)} x^{q+\alpha} \ \text{ for } x \ge 0 \qquad (q \ge 0), \tag{4}$$

where y^q is short notation for the mapping $y \mapsto y^q$. In the following, frequently we make use of the following elementary estimate:

$$\sup_{0 \le x \le b} |(\mathscr{V}^\alpha \varphi)(x)| \le \frac{b^\alpha}{\Gamma(\alpha+1)} \sup_{0 \le y \le b} |\varphi(y)|, \tag{5}$$

where $\varphi : [0, b] \to \mathbb{R}$ is a piecewise continuous function. Other basic properties of the Abel integral operator can be found e.g., in Gorenflo and Vessella [12] or Hackbusch [13].

2 The Product Midpoint Rule for Abel Integrals

2.1 The Method

For the numerical approximation of the Abel integral operator (3) we introduce equidistant grid points

$$x_n = nh, \qquad n = \frac{k}{2}, \quad k = 0, 1, \ldots, 2N, \quad \text{with } h = \frac{a}{N}, \tag{6}$$

where N is a positive integer. For a given continuous function $\varphi : [0, x_n] \to \mathbb{R}$ $(n \in \{1, 2, \ldots, N\})$, the product midpoint rule for the numerical approximation of the Abel integral $(\mathscr{V}^\alpha \varphi)(x_n)$ is obtained by replacing the function φ on each subinterval $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, n$, by the constant term $\varphi(x_{j-1/2})$, respectively:

$$\begin{aligned}(\mathscr{V}^\alpha \varphi)(x_n) &\approx \frac{1}{\Gamma(\alpha)} \sum_{j=1}^{n} \Big\{ \int_{x_{j-1}}^{x_j} (x_n - y)^{\alpha-1}\, dy \Big\} \varphi(x_{j-1/2}) &(7)\\ &= \frac{1}{\Gamma(\alpha+1)} \sum_{j=1}^{n} \big\{ (x_n - x_{j-1})^\alpha - (x_n - x_j)^\alpha \big\} \varphi(x_{j-1/2}) \\ &= \frac{h^\alpha}{\Gamma(\alpha+1)} \sum_{j=1}^{n} \big\{ (n-j+1)^\alpha - (n-j)^\alpha \big\} \varphi(x_{j-1/2}) \\ &= h^\alpha \sum_{j=1}^{n} \omega_{n-j} \varphi(x_{j-1/2}) =: (\Omega_h^\alpha \varphi)(x_n), &(8)\end{aligned}$$

where the quadrature weights $\omega_0, \omega_1, \ldots$ are given by

$$\omega_s = \frac{1}{\Gamma(\alpha+1)} \big\{ (s+1)^\alpha - s^\alpha \big\} \ \text{ for } s = 0, 1, \ldots\ . \tag{9}$$

The weights have the asymptotic behavior $\omega_s = \frac{1}{\Gamma(\alpha)} s^{\alpha-1} + \mathscr{O}(s^{\alpha-2})$ as $s \to \infty$.

2.2 The Integration Error: Preparations

In the sequel, we consider the integration error

$$(E_h^\alpha \varphi)(x_n) = (\mathscr{V}^\alpha \varphi)(x_n) - (\Omega_h^\alpha \varphi)(x_n) \tag{10}$$

under different smoothness assumptions on the function φ. As a preparation, for $c < d, L \geq 0, m = 0, 1, \ldots$ and $0 < \beta \leq 1$, we introduce the space $F_L^{m+\beta}[c, d]$ of all functions $\varphi : [c, d] \to \mathbb{R}$ that are continuously differentiable up to order m, and the derivative $\varphi^{(m)}$ of order m is Hölder continuous of order β with Hölder constant $L \geq 0$, i.e.,

$$F_L^{m+\beta}[c,d] = \{\, \varphi \in C^m[c,d] \mid |\varphi^{(m)}(x) - \varphi^{(m)}(y)| \leq L|x-y|^\beta \ \text{ for } x, y \in [c,d] \,\}. \tag{11}$$

The space of Hölder continuous functions of order $m + \beta$ on the interval $[c,d]$ is then given by

$$F^{m+\beta}[c,d] = \{\, \varphi : [c,d] \to \mathbb{R} \mid \varphi \in F_L^{m+\beta}[c,d] \text{ for some constant } L \geq 0 \,\}.$$

Other notations for the latter spaces are quite common, e.g., $C^{m,\beta}[c,d]$, cf. [3, Section 2].

As a preparation, for $n \in \{1,2,\ldots,N\}$ and $\varphi : [0,x_n] \to \mathbb{R}$ we introduce the piecewise constant interpolating spline $q_h\varphi : [0,x_n] \to \mathbb{R}$, i.e.,

$$(q_h\varphi)(y) \equiv \varphi(x_{j-1/2}) \ \text{ for } x_{j-1} \leq y < x_j \qquad (j = 1,2,\ldots,n), \tag{12}$$

and in the latter case $j = n$, this setting is also valid for $y = x_n$. For $\varphi \in F^p[0,x_n]$ with $0 < p \leq 1$, it follows from zero order Taylor expansions at the grid points that

$$\varphi(y) = (q_h\varphi)(y) + \mathscr{O}(h^p), \quad 0 \leq y \leq x_n, \tag{13}$$

uniformly both on $[0,x_n]$ and for $\varphi \in F_L^p[0,x_n]$, with any arbitrary but fixed constant $L \geq 0$, and also uniformly for $n = 1,2,\ldots,N$.

We consider the smooth case $\varphi \in C^1[0,x_n]$, $n \in \{1,2,\ldots,N\}$, next. Let $r_h\varphi : [0,x_n] \to \mathbb{R}$ be given by

$$(r_h\varphi)(y) = \varphi(x_{j-1/2}) + (y - x_{j-1/2})\varphi'(x_{j-1/2}) \ \text{ for } x_{j-1} \leq y < x_j \quad (j = 1,\ldots,n), \tag{14}$$

and in the latter case $j = n$, this definition is extended to the case $y = x_n$. For $\varphi \in F^p[0,x_n]$ with $1 < p \leq 2$, first order Taylor expansions at the grid points yield

$$\varphi(y) = (r_h\varphi)(y) + \mathscr{O}(h^p), \quad 0 \leq y \leq x_n, \tag{15}$$

uniformly in the same manner as for (13).

2.3 The Integration Error

We are now in a position to consider, under different smoothness conditions on the function φ, representations for the integration errors $(E_h^\alpha\varphi)(x_n)$ introduced in (10).

Lemma 1 *Let $n \in \{1, 2, \ldots, N\}$, and moreover let $\varphi : [0, x_n] \to \mathbb{R}$ be a continuous function. We have the following representations for the quadrature error $(E_h^\alpha \varphi)(x_n)$ introduced in (10):*

(a) We have

$$(E_h^\alpha \varphi)(x_n) = (\mathscr{V}^\alpha(\varphi - q_h\varphi))(x_n). \tag{16}$$

(b) For $\varphi \in C^1[0, x_n]$ we have

$$(E_h^\alpha \varphi)(x_n) = h^{\alpha+1} \sum_{j=1}^{n} \tau_{n-j} \varphi'(x_{j-1/2}) + (\mathscr{V}^\alpha(\varphi - r_h\varphi))(x_n), \tag{17}$$

where

$$\tau_s = \frac{1}{\Gamma(\alpha+2)}\{(s+1)^{\alpha+1} - s^{\alpha+1}\} - \frac{1}{2\Gamma(\alpha+1)}\{(s+1)^\alpha + s^\alpha\} \tag{18}$$

$$\text{for } s = 0, 1, \ldots .$$

Proof The error representation (16) is an immediate consequence of the identities (7) and (8). For the verification of the second error representation (17), we use the decomposition

$$(E_h^\alpha \varphi)(x_n) = (\mathscr{V}^\alpha(\varphi - q_h\varphi))(x_n) = (\mathscr{V}^\alpha(r_h\varphi - q_h\varphi))(x_n) + (\mathscr{V}^\alpha(\varphi - r_h\varphi))(x_n),$$

and we have to consider the first term on the right-hand side in more detail. Elementary computations show that

$$\frac{1}{\Gamma(\alpha)} \int_{x_{j-1}}^{x_j} (x_n - y)^{\alpha-1}(y - x_{j-1/2})\, dy = h^{\alpha+1}\tau_{n-j} \quad \text{for } j = 1, 2, \ldots, n. \tag{19}$$

From (19), the second error representation (17) already follows. This completes the proof of the lemma. □

A Taylor expansion of the right-hand side of (18) shows that the coefficients τ_s have the following asymptotic behavior:

$$\tau_s = \frac{1-\alpha}{12\Gamma(\alpha)} s^{\alpha-2} + \mathscr{O}(s^{\alpha-3}) \quad \text{as } s \to \infty. \tag{20}$$

Lemma 1 is needed in the proof of our main theorem. It is stated in explicit form here since it immediately becomes clear from this lemma that, for each $\varphi \in F^p[0, a]$ with $0 < p \le \alpha + 1$, the interpolation error satisfies

$$(E_h^\alpha \varphi)(x_n) = \mathscr{O}(h^p) \quad \text{as } h \to 0$$

uniformly for $n = 0, 1, \ldots, N$. This follows from (13) and (15), and from the absolute summability $\sum_{s=0}^\infty |\tau_s| < \infty$, cf. (20).

3 The Product Midpoint Rule for Abel-Type First-Kind Integral Equations with Perturbations

3.1 Some Preparations

We now return to the Abel-type integral equation (1). For the numerical approximation we consider this equation at grid points $x_n = nh, n = 1, 2, \ldots, N$ with $h = a/N$, cf. (6). The resulting integrals are approximated by the product midpoint rule, respectively, see (8) with $\varphi(y) = k(x_n, y)u(y)$ for $0 \le y \le x_n$.

In what follows, we suppose that the right-hand side of Eq. (1) is only approximately given, with

$$|f_n^\delta - f(x_n)| \le \delta \quad \text{for } n = 1, 2, \ldots, N, \tag{21}$$

where $\delta > 0$ is a known noise level. For this setting, the product midpoint rule for the numerical solution of Eq. (1) looks as follows:

$$h^\alpha \sum_{j=1}^{n} \omega_{n-j}\, k(x_n, x_{j-1/2})\, u_{j-1/2}^\delta = f_n^\delta, \qquad n = 1, 2, \ldots, N. \tag{22}$$

The approximations $u_{n-1/2}^\delta \approx u(x_{n-1/2})$ for $n = 1, 2, \ldots, N$ can be determined recursively by using scheme (22).

For the main error estimates, we impose the following conditions.

Assumption 1

(a) There exists a solution $u : [0, a] \to \mathbb{R}$ to the integral equation (1) which satisfies $u \in F^p[0, a]$, where $c_\alpha := \min\{\alpha, 1 - \alpha\} < p \le 2$.
(b) There holds $k(x, x) = 1$ for each $0 \le x \le a$.
(c) The kernel function k has Lipschitz continuous partial derivatives of second order.
(d) The grid points x_n are given by (6).
(e) The right-hand side of Eq. (1) is approximately given by (21).

3.2 Formal Power Series

As a preparation for the proof of the main stability result of the present paper, cf. Theorem 1, we next consider power series. In what follows, we identify sequences $(b_n)_{n\ge0}$ of complex numbers with their (formal) power series $b(\xi) = \sum_{n=0}^\infty b_n \xi^n$, with $\xi \in \mathbb{C}$. Pointwise multiplication of two power series

$$\Big(\sum_{\ell=0}^{\infty} b_\ell \xi^\ell\Big) \cdot \Big(\sum_{j=0}^{\infty} c_j \xi^j\Big) = \sum_{n=0}^{\infty} d_n \xi^n, \quad \text{with } d_n := \sum_{\ell=0}^{n} b_\ell c_{n-\ell} \quad \text{for } n = 0, 1, \ldots$$

makes the set of power series into a complex commutative algebra with unit element $1 + 0 \cdot \xi + 0 \cdot \xi^2 + \cdots$. For any power series $b(\xi) = \sum_{n=0}^{\infty} b_n \xi^n$ with $b_0 \neq 0$, there exists a power series which inverts the power series $b(\xi)$ with respect to pointwise multiplication, and it is denoted by $1/b(\xi)$ or by $[b(\xi)]^{-1}$. For a comprehensive introduction to formal power series see, e.g., Henrici [15].

In what follows, we consider the inverse

$$[\omega(\xi)]^{-1} = \sum_{n=0}^{\infty} \omega_n^{(-1)} \xi^n \tag{23}$$

of the generating function $\omega(\xi) = \sum_{n=0}^{\infty} \omega_n \, \xi^n$, with ω_n as in (9).

Lemma 2 *The coefficients in (23) have the following properties:*

$$\omega_0^{(-1)} > 0, \qquad \omega_n^{(-1)} < 0 \ \textit{for} \ n = 1, 2, \ldots, \tag{24}$$

$$\omega_0^{(-1)} = \Gamma(\alpha + 1) = \sum_{n=1}^{\infty} |\omega_n^{(-1)}|, \tag{25}$$

$$\omega_n^{(-1)} = \mathscr{O}(n^{-\alpha-1}) \quad \textit{as} \ n \to \infty. \tag{26}$$

Estimate (26) can be found in [8]. Another proof of (26) which uses Banach algebra theory and may be of independent interest is given in Sect. 7 of the present paper. Section 7 also contains proofs of the other statements in Lemma 2.

Lemma 2 is needed in the proof of our main result, cf. Theorem 1 below and Sect. 8. We state the lemma here in explicit form since it is fundamental in the stability estimates.

3.3 The Main Result

We next present the first main result of this paper, cf. the following theorem, where different continuity and smoothness properties of the solution u are considered. For comments on the estimates presented in the theorem, see Remark 1 below.

Theorem 1 *Let the conditions of Assumption 1 be satisfied, and consider the approximations $u_{1/2}^{\delta}, u_{3/2}^{\delta}, \ldots, u_{N-1/2}^{\delta}$ determined by scheme (22). Let* $c_\alpha := \min\{\alpha, 1 - \alpha\}$.

(a) If $c_\alpha < p \le 1 + c_\alpha$, then we have

$$\max_{n=1,2,\ldots,N} |u_{n-1/2}^{\delta} - u(x_{n-1/2})| = \mathscr{O}(h^{p-c_\alpha} + \frac{\delta}{h^\alpha}) \quad \textit{as} \ (h, \delta) \to 0. \tag{27}$$

(b) Let $2 - \alpha < p \le 2$, and in addition let $u(0) = u'(0) = 0$ be satisfied. Then

$$\max_{n=1,2,\ldots,N} |u_{n-1/2}^{\delta} - u(x_{n-1/2})| = \mathscr{O}(h^{p-1+\alpha} + \frac{\delta}{h^\alpha}) \quad \textit{as} \ (h, \delta) \to 0. \tag{28}$$

The proof of Theorem 1 is given in Sect. 8.

Remark 1 We give some comments on Theorem 1. Due to the special form of the term c_α appearing in Theorem 1, it makes sense to distinguish the cases $\alpha \le \frac{1}{2}$ and $\alpha \ge \frac{1}{2}$ which in fact will be done in items (a) and (b).

(a) In the case $0 < \alpha \le \frac{1}{2}$, the following estimate holds:

$$\max_{n=1,\dots,N} |u_n^\delta - u(x_{n-1/2})| = \begin{cases} \mathscr{O}(h^{p-\alpha} + \frac{\delta}{h^\alpha}), \text{ if } \alpha < p \le \alpha + 1, \\ \mathscr{O}(h^{p-1+\alpha} + \frac{\delta}{h^\alpha}), \text{ if } 2 - \alpha < p \le 2, \ u(0) = u'(0) = 0. \end{cases}$$

(b) In the case $\frac{1}{2} \le \alpha < 1$, the following estimate holds:

$$\max_{n=1,\dots,N} |u_n^\delta - u(x_{n-1/2})| = \mathscr{O}(h^{p-1+\alpha} + \frac{\delta}{h^\alpha}), \text{ if } \ 1 - \alpha < p \le 2 - \alpha, \\ \text{or if } \ 2 - \alpha < p \le 2, \ u(0) = u'(0) = 0.$$

For an extension to Volterra integral equations of the first kind with smooth kernels ($\alpha = 1$), cf. Remark 3 below.

(c) The noise-free rates, obtained for $p = 1$ and $p = 2$, basically coincide with those given in the papers by Weiss and Anderssen [30] and by Eggermont [9].

(d) The maximal rate in the noise-free case $\delta = 0$ and without initial conditions is $\mathscr{O}(h)$, and it is obtained for $p = 1 + c_\alpha$. This rate is indeed maximal for sufficiently smooth functions, as can be seen by considering the error at the first grid point $x_{1/2}$, obtained for the function $u(y) = y$, cf. Weiss and Anderssen [30]. Under the additional assumption $u(0) = u'(0) = 0$, the maximal rate is $\mathscr{O}(h^{\alpha+1})$, obtained for $p = 2$.

(e) It is not clear if the rates presented in Theorem 1 are optimal under the respective continuity and smoothness conditions. △

In what follows, for step sizes $h = a/N$ we write, with a slight abuse of notation, $h \sim \delta^\beta$ as $\delta \to 0$, if there exist real constants $c_2 \ge c_1 > 0$ such that $c_1 h \le \delta^\beta \le c_2 h$ holds for $\delta \to 0$. As an immediate consequence of Theorem 1 we obtain the following main result of this paper.

Corollary 1 *Let Assumption 1 be satisfied.*

- *Let $\alpha \le 1/2$ and $\alpha < p \le \alpha + 1$. For $h = h(\delta) \sim \delta^{1/p}$ we have*

$$\max_{n=1,2,\dots,N} |u_{n-1/2}^\delta - u(x_{n-1/2})| = \mathscr{O}(\delta^{1-\alpha/p}) \quad \textit{as } \ \delta \to 0.$$

- *Let one of the following two conditions be satisfied: (a) $\alpha \ge 1/2$, $1 - \alpha < p \le 2 - \alpha$, or (b) $2 - \alpha < p \le 2$, $u(0) = u'(0) = 0$. Then for $h = h(\delta) \sim \delta^{1/(p-1+2\alpha)}$ we have*

$$\max_{n=1,2,\dots,N} |u_{n-1/2}^\delta - u(x_{n-1/2})| = \mathscr{O}\big(\delta^{1-\frac{\alpha}{p-1+2\alpha}}\big) \quad \textit{as } \ \delta \to 0.$$

Note that in the case $\alpha < \frac{1}{2}$, for the class of functions satisfying the initial conditions $u(0) = u'(0) = 0$, there is a gap for $\alpha + 1 < p \le 2 - \alpha$ where no improvement in the rates is obtained, i.e., we have piecewise saturation $\mathscr{O}(\delta^{1/(\alpha+1)})$ for the given range of p. This is due to different techniques used in the proof of Theorem 1.

We conclude this section with some more remarks.

Remark 2

(a) We mention some other quadrature schemes for the approximate solution of Abel-type integral equations of the first kind. The product trapezoidal method is considered, e.g., in Weiss [29], Eggermont [9], and in [22]. Fractional multistep methods are treated in Lubich [19, 20] and in [21]. Backward difference product integration methods are analyzed in Cameron and McKee [6, 7]. Galerkin methods for Abel-type integral equations are considered, e.g., in Eggermont [9] and in Vögeli et al. [28]. Some general references are already given in the beginning of the present paper.

(b) For other special regularization methods for the approximate solution of Volterra integral equations of the first kind with perturbed right-hand sides and with possibly algebraic-type weakly singular kernels, see e.g., Anderssen [2], Bughgeim [5], Gorenflo and Vessella [12], and the references therein.

Remark 3 The results of Theorem 1 and Corollary 1 can be extended to linear Volterra integral equations of the first kind with smooth kernels, that is, for $\alpha = 1$. The resulting method is in fact the classical repeated midpoint rule, and the main error estimate is as follows: if $0 < p \le 2$, then we have

$$\max_{n=1,2,\dots,N} |u^\delta_{n-1/2} - u(x_{n-1/2})| = \mathscr{O}(h^p + \frac{\delta}{h}) \quad \text{as} \quad (h, \delta) \to 0,$$

and initial conditions are not required anymore then. The choice $h = h(\delta) \sim \delta^{1/(p+1)}$ then gives

$$\max_{n=1,2,\dots,N} |u^\delta_{n-1/2} - u(x_{n-1/2})| = \mathscr{O}(\delta^{p/(p+1)}) \quad \text{as} \quad \delta \to 0.$$

The proof follows the lines used in the present paper, with a lot of simplifications then. In particular, the inverse stability results derived in Sect. 7 can be discarded in this case. We leave the details to the reader and indicate the basic ingredients only: we have $\omega_n = 1$ and $\tau_n = 0$ for $n = 0, 1, \dots$ then, and in addition, $\omega_0^{(-1)} = 1, \omega_1^{(-1)} = -1$, and $\omega_n^{(-1)} = 0$ for $n = 2, 3, \dots$ holds. For other results on the regularizing properties of the repeated midpoint rule for solving linear Volterra integral equations of the first kind with smooth kernels, see [23] and Kaltenbacher [16].

4 Modified Starting Weights

For the product midpoint rule (8), applied to a continuous function $\varphi : [0, a] \to \mathbb{R}$, and with grid points as in (6), with $1 \le n \le N$ and $N \ge 2$, we now would like to overcome the conditions $\varphi(0) = \varphi'(0) = 0$. For this purpose we consider the modification

$$(\widetilde{\Omega}_h \varphi)(x_n) := \overbrace{h^\alpha \sum_{j=1}^{n} \omega_{n-j}\, \varphi(x_{j-1/2})}^{=(\Omega_h \varphi)(x_n)} + h^\alpha \sum_{j=1}^{2} w_{nj}\, \varphi(x_{j-1/2}) \tag{29}$$

as approximation to the fractional integral $(\mathscr{V}^\alpha \varphi)(x_n)$ at the considered grid points x_n, respectively. See Lubich [19, 20] and [21] for a similar approach for fractional multistep methods. In (29), w_{n1} and w_{n2} are correction weights for the starting values that are specified in the following. In fact, for each $n = 1, 2, \ldots, N$ the correction weights are chosen such that the modified product midpoint rule (29) is exact at $x_n = nh$ for polynomials of degree ≤ 1, i.e.,

$$(\widetilde{\Omega}_h y^q)(x_n) = (\mathscr{V}^\alpha y^q)(x_n) \quad \text{for } q = 0, 1. \tag{30}$$

4.1 Computation of the Correction Weights

For each $n = 1, 2, \ldots, N$, a reformulation of (30) gives the following linear system of two equations for the correction weights w_{nj}, $j = 1, 2$:

$$h^\alpha (w_{n1} + w_{n2}) = (E_h^\alpha 1)(x_n), \qquad h^{\alpha+1}(\tfrac{1}{2} w_{n1} + \tfrac{3}{2} w_{n2}) = (E_h^\alpha y)(x_n),$$

cf. (10) for the introduction of E_h^α. On the other hand we have

$$(E_h^\alpha 1)(x_n) = 0, \qquad (E_h^\alpha y)(x_n) = h^{\alpha+1} \sum_{s=0}^{n-1} \tau_s.$$

Those identities follow from the representations (16) and (17), respectively. From this we obtain

$$-w_{n1} = w_{n2} = \sum_{s=0}^{n-1} \tau_s. \tag{31}$$

This in particular means that the correction weights are independent of h. We finally note that the asymptotic behavior of the coefficients τ_s, cf. (20), implies

$$w_{nj} = \mathscr{O}(1) \quad \text{as } n \to \infty \qquad \text{for} \quad j = 1, 2. \tag{32}$$

4.2 Integration Error of the Modified Quadrature Method

We now consider, for each $n = 1, 2, \ldots, N$, the error of the modified product midpoint rule,

$$(\widetilde{E}_h^\alpha \varphi)(x_n) = (\mathcal{V}^\alpha \varphi)(x_n) - (\widetilde{\Omega}_h \varphi)(x_n), \tag{33}$$

where $\varphi : [0, a] \to \mathbb{R}$ denotes a continuous function.

Lemma 3 *Let $n \in \{1, 2, \ldots, N\}$, and moreover let $\varphi \in F^p[0, a]$, with $0 \leq p \leq 2$. We have the following representations of the modified quadrature error $(\widetilde{E}_h^\alpha \varphi)(x_n)$ introduced in (33):*

(a) In the case $0 < p \leq 1$ we have $(\widetilde{E}_h^\alpha \varphi)(x_n) = (E_h^\alpha \varphi)(x_n) + \mathcal{O}(h^{p+\alpha})$ as $h \to 0$.
(b) In the case $1 < p \leq 2$ we have, with $\widetilde{\varphi}(y) := \varphi(y) - \varphi(0) - \varphi'(0)y$ for $0 \leq y \leq a$,

$$(\widetilde{E}_h^\alpha \varphi)(x_n) = (E_h^\alpha \widetilde{\varphi})(x_n) + \mathcal{O}(h^{p+\alpha}) \quad \text{as } h \to 0.$$

Both statements hold uniformly for $n = 1, 2, \ldots, N$, and for $\varphi \in F_L^p[0, a]$, with $L \geq 0$ arbitrary but fixed.

Proof

(a) This follows immediately from (29) and (31)–(33):

$$\begin{aligned}(\widetilde{E}_h^\alpha \varphi)(x_n) &= (E_h^\alpha \varphi)(x_n) + h^\alpha w_{n1}\big(\varphi(x_{3/2}) - \varphi(x_{1/2})\big) \\ &= (E_h^\alpha \varphi)(x_n) + \mathcal{O}(h^{p+\alpha}) \quad \text{as } h \to 0.\end{aligned}$$

(b) Using the notation $q(y) := \varphi(0) + \varphi'(0)y$, we have $\varphi = \widetilde{\varphi} + q$, and the linearity of the modified error functional gives

$$\begin{aligned}(\widetilde{E}_h^\alpha \varphi)(x_n) &= (\widetilde{E}_h^\alpha \widetilde{\varphi})(x_n) + \overbrace{(\widetilde{E}_h^\alpha q)(x_n)}^{=0} = (E_h^\alpha \widetilde{\varphi})(x_n) - h^\alpha \sum_{j=1}^{2} w_{nj} \widetilde{\varphi}(x_{j-1/2}) \\ &= (E_h^\alpha \widetilde{\varphi})(x_n) + \mathcal{O}(h^{p+\alpha}),\end{aligned}$$

where $\widetilde{\varphi}(y) = \mathcal{O}(y^p)$ as $y \to 0$ has been used, and the boundedness of the correction weights, cf. (32), is also taken into account. □

4.3 Application to the Abel-Type Integral Equation of the First Kind

In what follows, the modified product midpoint rule (29) is applied to solve the algebraic-type weakly singular Volterra integral equation (1) numerically, with noisy data as in (21). In order to make the starting procedure applicable, in the following we assume that the kernel k can be smoothly extended beyond the triangle $\{ 0 \le y \le x \le a \}$. For simplicity we assume that the kernel is defined on the whole square.

Assumption 2 *The kernel function k has Lipschitz continuous partial derivatives of second order on* $[0, a] \times [0, a]$.

For each $n = 1, 2, \ldots, N$, we consider the modified product midpoint rule (29) with $\varphi(y) = k(x_n, y)u(y)$ for $0 \le y \le a$, $n = 1, 2, \ldots, N$. This results in the following modified scheme:

$$h^\alpha \sum_{j=1}^{n} \omega_{n-j} k(x_n, x_{j-1/2}) \widetilde{u}^\delta_{j-1/2} + h^\alpha \sum_{j=1}^{2} w_{nj} k(x_n, x_{j-1/2}) \widetilde{u}^\delta_{j-1/2} = f^\delta_n \tag{34}$$

for $n = 1, 2, \ldots, N$. This scheme can be realized by first solving a coupled linear system of two equations for the approximations $\widetilde{u}^\delta_{n-1/2} \approx u(x_{n-1/2}),\ n = 1, 2$. The approximations $\widetilde{u}^\delta_{n-1/2} \approx u(x_{n-1/2})$ for $n = 3, 4, \ldots, N$ then can be determined recursively by using scheme (34).

4.4 Uniqueness, Existence and Approximation Properties of the Starting Values

We next consider uniqueness, existence and approximation properties of the two starting values $\widetilde{u}^\delta_{1/2}$ and $\widetilde{u}^\delta_{3/2}$. They in fact satisfy the linear system of equations

$$h^\alpha \sum_{j=1}^{2} \underbrace{(\omega_{n-j} + w_{nj})}_{=: \, \overline{\omega}_{nj}} k(x_n, x_{j-1/2}) \widetilde{u}^\delta_{n-1/2} = f^\delta_n \text{ for } n = 1, 2, \tag{35}$$

with the notation $\omega_{-1} = 0$. In matrix notation, this linear system of equations can be written as

$$h^\alpha \overbrace{\begin{pmatrix} \overline{\omega}_{11} k(x_1, x_{1/2}) & \overline{\omega}_{12} k(x_1, x_{3/2}) \\ \overline{\omega}_{21} k(x_2, x_{1/2}) & \overline{\omega}_{22} k(x_2, x_{3/2}) \end{pmatrix}}^{=S_h} \begin{pmatrix} \widetilde{u}^\delta_{1/2} \\ \widetilde{u}^\delta_{3/2} \end{pmatrix} = \begin{pmatrix} f^\delta_1 \\ f^\delta_2 \end{pmatrix}. \tag{36}$$

Lemma 4 *The matrix* $S_h \in \mathbb{R}^{2 \times 2}$ *in (36) is regular for sufficiently small values of h, and* $\| {S_h}^{-1} \|_\infty = \mathscr{O}(1)$ *as* $h \to 0$, *where* $\| \cdot \|_\infty$ *denotes the matrix norm induced by the maximum vector norm on* $\mathbb{R}^2$.

Proof We first consider the situation $k \equiv 1$ and denote the matrix S_h by T in this special case. From (4) and (30) it follows

$$\overline{\omega}_{n1} + \overline{\omega}_{n2} = \frac{n^{\alpha}}{\Gamma(\alpha+1)} \qquad \tfrac{1}{2}\overline{\omega}_{n1} + \tfrac{3}{2}\overline{\omega}_{n2} = \frac{n^{\alpha+1}}{\Gamma(\alpha+2)}, \qquad n = 1, 2.$$

Hence the matrix T is regular and does not depend on h.

We next consider the general case for k. Since $k(x,x) = 1$, we have $k(x_n, x_m) \to 1$ as $h \to 0$ uniformly for the four function values of k considered in the matrix S_h. This shows $S_h = T + \Delta_h$ with $\|\Delta_h\|_\infty \to 0$ as $h \to 0$ so that the matrix S_h is regular for sufficiently small values h, with $\|S_h{}^{-1}\|_\infty$ being bounded as $h \to 0$. This completes the proof of the lemma. □

We next consider the error of the modified product midpoint rule at the two grid points $x_{1/2}$ and $x_{3/2}$.

Proposition 1 *Let the conditions of Assumptions 1 and 2 be satisfied. Consider the approximations $\widetilde{u}^{\delta}_{1/2}$ and $\widetilde{u}^{\delta}_{3/2}$ determined by scheme (34) for $n = 1, 2$. Then we have*

$$\max_{n=1,2} |\widetilde{u}^{\delta}_{n-1/2} - u(x_{n-1/2})| = \mathscr{O}(h^p + \frac{\delta}{h^\alpha}) \quad \textit{as} \ \ (h,\delta) \to 0.$$

Proof From (29), (33) and Lemma 3, applied with $\varphi(y) = \varphi_n(y) = k(x_n, y)u(y)$ for $0 \le y \le a$, we obtain the representation

$$\begin{aligned} h^\alpha \sum_{j=1}^{2} \overline{\omega}_{nj}\, k(x_n, x_{j-1/2})\, \tilde{e}^{\delta}_{j-1/2} &= (\widetilde{E}^{\alpha}_{h}\varphi_n)(x_n) + f^{\delta}_n - f(x_n) \\ &= \mathscr{O}(h^{p+\alpha} + \delta) \quad \text{as} \ \ (h,\delta) \to 0, \quad n = 1, 2, \end{aligned}$$

where $\tilde{e}^{\delta}_{j-1/2} = \widetilde{u}^{\delta}_{j-1/2} - u(x_{j-1/2}),\ j = 1, 2$, and the weights $\overline{\omega}_{nj}$ are introduced in (35). Note that Lemmas 1 and 3 imply, for the two integers $n = 1, 2$, that $(\widetilde{E}^{\alpha}_{h}\varphi_n)(x_n) = \mathscr{O}(h^{p+\alpha})$ as $h \to 0$. The proposition now follows from Lemma 4. □

4.5 The Regularizing Properties of the Modified Scheme

Theorem 2 *Let the conditions of Assumptions 1 and 2 be satisfied.*

(a) In the case $\alpha \le 1/2$ we have

$$\max_{n=1,2,\dots,N} |\widetilde{u}^{\delta}_{n-1/2} - u(x_{n-1/2})| = \begin{cases} \mathscr{O}(h^{p-\alpha} + \frac{\delta}{h^\alpha}) \text{ if } \alpha < p \le \alpha + 1, \\ \mathscr{O}(h^{p-1+\alpha} + \frac{\delta}{h^\alpha}) \text{ if } 2 - \alpha < p \le 2. \end{cases}$$

(b) In the case $\alpha \geq 1/2$, $1-\alpha < p \leq 2$ we have

$$\max_{n=1,2,\dots,N} |\widetilde{u}^{\delta}_{n-1/2} - u(x_{n-1/2})| = \mathscr{O}(h^{p-1+\alpha} + \frac{\delta}{h^{\alpha}}) \quad \textit{as} \ \ (h,\delta) \to 0.$$

Proof Let $\tilde{e}^{\delta}_{j-1/2} = \widetilde{u}^{\delta}_{j-1/2} - u(x_{j-1/2})$ for $j = 1,2,\dots,N$. From (29), (32), (33), Lemma 3 and Proposition 1 we obtain the representation

$$\begin{aligned} &h^{\alpha}\sum_{j=1}^{n} \omega_{n-j} k(x_n, x_{j-1/2}) \tilde{e}^{\delta}_{j-1/2} \\ &\qquad = (\widetilde{E}^{\alpha}_{h}\varphi_n)(x_n) + f(x_n) - f^{\delta}_n - h^{\alpha}\sum_{j=1}^{2} w_{nj} k(x_n, x_{j-1/2})\tilde{e}^{\delta}_{j-1/2} \\ &\qquad = (\widetilde{E}^{\alpha}_{h}\varphi_n)(x_n) + \mathscr{O}(h^{p+\alpha} + \delta) = (E^{\alpha}_{h}\widetilde{\varphi}_n)(x_n) + \mathscr{O}(h^{p+\alpha} + \delta) \end{aligned}$$

as $(h,\delta) \to 0$, uniformly for $n = 1,2,\dots,N$, where $\widetilde{\varphi}_n = \varphi_n$, if $p \leq 1$, and $\widetilde{\varphi}_n(y) = \varphi_n(y) - \varphi_n(0) - \varphi_n'(0)y$ for $p > 1$. The theorem now follows by performing the same steps as in the proof of Theorem 1. □

As an immediate consequence of Theorem 2, we can derive regularizing properties of the modified scheme.

Corollary 2 *Let both Assumptions 1 and 2 be satisfied.*

- *If $\alpha \leq 1/2$ and $\alpha < p \leq \alpha + 1$, then choose $h = h(\delta) \sim \delta^{1/p}$. The resulting error estimate is*

$$\max_{n=1,2,\dots,N} |\widetilde{u}^{\delta}_{n-1/2} - u(x_{n-1/2})| = \mathscr{O}(\delta^{1-\alpha/p}) \quad \textit{as} \ \ \delta \to 0.$$

- *Let one of the following two conditions be satisfied: (a) $\alpha \geq 1/2$, $1-\alpha < p \leq 2-\alpha$, or (b) $2-\alpha < p \leq 2$. For $h = h(\delta) \sim \delta^{1/(p-1+2\alpha)}$ we then have*

$$\max_{n=1,2,\dots,N} |\widetilde{u}^{\delta}_{n-1/2} - u(x_{n-1/2})| = \mathscr{O}(\delta^{1-\frac{\alpha}{p-1+2\alpha}}) \quad \textit{as} \ \ \delta \to 0.$$

5 Numerical Experiments

We next present results of some numerical experiments with the linear Abel-type integral equation of the first kind (1). The following example is considered for different values of $0 < \alpha < 1$ and $0 < q \leq 2$:

$$k(x,y) = \frac{1+xy}{1+x^2}, \quad f(x) = \frac{1}{\Gamma(q+2+\alpha)}\frac{x^{q+\alpha}}{1+x^2}(q+1+\alpha+(q+1)x^2), \quad 0 \leq x,y \leq 1, \tag{37}$$

with exact solution (cf. (4))

$$u(y) = \tfrac{1}{\Gamma(q+1)} y^q \quad \text{for } 0 \le y \le 1, \tag{38}$$

so that the conditions in (a)–(c) of Assumption 1 are satisfied with at least $p = q$, provided that $q > c_\alpha$. We present experiments for different values of α and q, sometimes with corrections weights, sometimes without, in order to cover all variants in Corollaries 1 and 2. Here are additional remarks on the numerical tests.

- Numerical experiments with step sizes $h = 1/2^m$ for $m = 5, 6, \ldots, 11$ are employed, respectively.
- For each considered step size h, we consider the noise level $\delta = \delta(h) = ch^{\nu+\alpha}$, where $c = 0.3$, and $\nu = \nu(\alpha, p)$ denotes the rate for exact data, supplied by Theorems 1 and 2. The available error estimate is then of the form $\max_n |u_n^\delta - u(x_n)| = \mathscr{O}(h^\nu) = \mathscr{O}(\delta^{\nu/(\nu+\alpha)})$ as $h \to 0$.
- In the numerical experiments, the perturbations are of the form $f_n^\delta = f(x_n) + \Delta_n$ with uniformly distributed random values Δ_n with $|\Delta_n| \le \delta$.
- In all tables, $\|f\|_\infty$ denotes the maximum norm of the function f.
- Experiments are employed using the programming language OCTAVE.

Example 1 We first consider the situation (37)–(38), with $\alpha = \frac{1}{2}$ and $q = 2$. The conditions in (a)–(c) of Assumption 1 are satisfied with $p = 2$ (also for any $p > 2$ in fact, but then we have saturation). We have $u(0) = u'(0) = 0$, so correction weights are not required here. The provided error estimate, with the choice of $\delta = \delta(h)$ considered in the beginning of this section, is $\max_n |u_n^\delta - u(x_n)| = \mathscr{O}(\delta^{3/4}) = \mathscr{O}(h^{3/2})$. The numerical results are shown in Table 1.

Example 2 We next consider the situation (37)–(38), with $\alpha = 0.9$ and $q = 0.4$ this time. The conditions in (a)–(c) of Assumption 1 are satisfied with $p = 0.4$. Since $p \le 1$, correction weights are not needed here. The expected error estimate, with $\delta = \delta(h)$ as in the beginning of this section, is $\max_n |u_n^\delta - u(x_n)| = \mathscr{O}(\delta^{1/4}) = \mathscr{O}(h^{0.3})$. The numerical results are shown in Table 2.

Example 3 We next consider the situation (37)–(38) with $\alpha = 0.2$ and $q = 0.5$. The conditions in (a)–(c) of Assumption 1 are satisfied with $p = 0.5$ then, and correction

Table 1 Numerical results for Example 1

N	δ	$100\cdot\delta/\|f\|_\infty$	$\max_n \|u_n^\delta - u(x_n)\|$	$\max_n \|u_n^\delta - u(x_n)\| / \delta^{3/4}$
32	$2.9\cdot10^{-4}$	$9.74\cdot10^{-2}$	$2.84\cdot10^{-3}$	1.27
64	$7.3\cdot10^{-5}$	$2.43\cdot10^{-2}$	$1.12\cdot10^{-3}$	1.41
128	$1.8\cdot10^{-5}$	$6.09\cdot10^{-3}$	$3.77\cdot10^{-4}$	1.35
256	$4.6\cdot10^{-6}$	$1.52\cdot10^{-3}$	$1.37\cdot10^{-4}$	1.38
512	$1.1\cdot10^{-6}$	$3.80\cdot10^{-4}$	$5.20\cdot10^{-5}$	1.48
1024	$2.9\cdot10^{-7}$	$9.51\cdot10^{-5}$	$1.89\cdot10^{-5}$	1.53
2048	$7.2\cdot10^{-8}$	$2.38\cdot10^{-5}$	$6.55\cdot10^{-6}$	1.50

Table 2 Numerical results for Example 2

N	δ	$100{\cdot}\delta/\|f\|_\infty$	$\max_n \lvert u_n^\delta - u(x_n)\rvert$	$\max_n \lvert u_n^\delta - u(x_n)\rvert\,/\delta^{1/4}$
32	$4.7{\cdot}10^{-3}$	$6.80{\cdot}10^{-1}$	$1.88{\cdot}10^{-1}$	0.72
64	$2.0{\cdot}10^{-3}$	$2.96{\cdot}10^{-1}$	$1.32{\cdot}10^{-1}$	0.62
128	$8.9{\cdot}10^{-4}$	$1.29{\cdot}10^{-1}$	$1.23{\cdot}10^{-1}$	0.71
256	$3.9{\cdot}10^{-4}$	$5.61{\cdot}10^{-2}$	$9.61{\cdot}10^{-2}$	0.69
512	$1.7{\cdot}10^{-4}$	$2.44{\cdot}10^{-2}$	$8.12{\cdot}10^{-2}$	0.71
1024	$7.3{\cdot}10^{-5}$	$1.06{\cdot}10^{-2}$	$6.77{\cdot}10^{-2}$	0.73
2048	$3.2{\cdot}10^{-5}$	$4.62{\cdot}10^{-3}$	$5.43{\cdot}10^{-2}$	0.72

Table 3 Numerical results for Example 3

N	δ	$100{\cdot}\delta/\|f\|_\infty$	$\max_n \lvert u_n^\delta - u(x_n)\rvert$	$\max_n \lvert u_n^\delta - u(x_n)\rvert\,/\delta^{0.6}$
32	$5.3{\cdot}10^{-2}$	$5.12{\cdot}10^{0}$	$1.18{\cdot}10^{-1}$	0.69
64	$3.8{\cdot}10^{-2}$	$3.62{\cdot}10^{0}$	$8.52{\cdot}10^{-2}$	0.61
128	$2.7{\cdot}10^{-2}$	$2.56{\cdot}10^{0}$	$7.78{\cdot}10^{-2}$	0.69
256	$1.9{\cdot}10^{-2}$	$1.81{\cdot}10^{0}$	$5.89{\cdot}10^{-2}$	0.64
512	$1.3{\cdot}10^{-2}$	$1.28{\cdot}10^{0}$	$5.19{\cdot}10^{-2}$	0.69
1024	$9.4{\cdot}10^{-3}$	$9.05{\cdot}10^{-1}$	$4.20{\cdot}10^{-2}$	0.69
2048	$6.6{\cdot}10^{-3}$	$6.40{\cdot}10^{-1}$	$3.33{\cdot}10^{-2}$	0.68

Table 4 Numerical results for Example 4, without correction weights

N	δ	$100{\cdot}\delta/\|f\|_\infty$	$\max_n \lvert u_n^\delta - u(x_n)\rvert$	$\max_n \lvert u_n^\delta - u(x_n)\rvert\,/\delta^{2/3}$
32	$1.7{\cdot}10^{-3}$	$2.20{\cdot}10^{-1}$	$1.26{\cdot}10^{-2}$	0.90
64	$5.9{\cdot}10^{-4}$	$7.79{\cdot}10^{-2}$	$6.47{\cdot}10^{-3}$	0.92
128	$2.1{\cdot}10^{-4}$	$2.75{\cdot}10^{-2}$	$3.27{\cdot}10^{-3}$	0.94
256	$7.3{\cdot}10^{-5}$	$9.74{\cdot}10^{-3}$	$1.57{\cdot}10^{-3}$	0.89
512	$2.6{\cdot}10^{-5}$	$3.44{\cdot}10^{-3}$	$7.72{\cdot}10^{-4}$	0.88
1024	$9.2{\cdot}10^{-6}$	$1.22{\cdot}10^{-3}$	$3.95{\cdot}10^{-4}$	0.90
2048	$3.2{\cdot}10^{-6}$	$4.30{\cdot}10^{-4}$	$2.06{\cdot}10^{-4}$	0.94

weights are not needed here because of $p \le 1$. The available error estimate is $\max_n |u_n^\delta - u(x_n)| = \mathscr{O}(\delta^{0.6}) = \mathscr{O}(h^{0.3})$. The numerical results are shown in Table 3.

Example 4 Finally we consider the situation (37)–(38) with $\alpha = 0.5$ and $q = 1$. The conditions in (a)–(c) of Assumption 1 are satisfied with any $0.5 < p \le 2$ then, and initial conditions are not satisfied in this case. The presented theory for the product midpoint rule without correction weights suggests that we have $\max_n |u_n^\delta - u(x_n)| = \mathscr{O}(\delta^{2/3}) = \mathscr{O}(h)$. The corresponding numerical results are shown in Table 4.

For the same problem, we also consider the modified version of the product midpoint rule, i.e., correction weights are used this time. The presented theory then yields $\max_n |u_n^\delta - u(x_n)| = \mathscr{O}(\delta^{3/4}) = \mathscr{O}(h^{3/2})$. The related numerical results are shown in Table 5.

Table 5 Numerical results for Example 4, with correction weights

N	δ	$100{\cdot}\delta/\|f\|_\infty$	$\max_n \lvert u_n^\delta - u(x_n) \rvert$	$\max_n \lvert u_n^\delta - u(x_n) \rvert \,/\delta^{3/4}$
32	$2.9{\cdot}10^{-4}$	$3.89{\cdot}10^{-2}$	$2.10{\cdot}10^{-3}$	0.94
64	$7.3{\cdot}10^{-5}$	$9.74{\cdot}10^{-3}$	$6.56{\cdot}10^{-4}$	0.83
128	$1.8{\cdot}10^{-5}$	$2.43{\cdot}10^{-3}$	$2.88{\cdot}10^{-4}$	1.03
256	$4.6{\cdot}10^{-6}$	$6.09{\cdot}10^{-4}$	$8.66{\cdot}10^{-5}$	0.87
512	$1.1{\cdot}10^{-6}$	$1.52{\cdot}10^{-4}$	$3.46{\cdot}10^{-5}$	0.99
1024	$2.9{\cdot}10^{-7}$	$3.80{\cdot}10^{-5}$	$1.22{\cdot}10^{-5}$	0.99
2048	$7.2{\cdot}10^{-8}$	$9.51{\cdot}10^{-6}$	$4.31{\cdot}10^{-6}$	0.99

The last column in each table shows that the theory is confirmed in each of the five numerical experiments.

6 Conclusions

In the present paper we have considered the product midpoint rule for the regularization of algebraic-type weakly singular Volterra integral equations of the first kind with perturbed given right-hand sides. The applied techniques are closely related to those used in Eggermont [8]. The presented results include intermediate continuity and smoothness degrees of the solution of the integral equation in terms of a scale of Hölder spaces. In addition we have given a new proof of the stability estimate for the inverse of the generating sequence, cf. (26), which may be of independent interest. Another topic is the use of correction starting weights to get rid of initial conditions on the solution. Results of some numerical experiments are also given.

7 Appendix 1: Proof of Lemma 2

We next present a proof of estimate (26) for the coefficients of the inverse of the considered generating power series $\sum_{n=0}^{\infty} \omega_n \xi^n$ which differs from that given by Eggermont [8]. Our proof uses Banach algebra theory and may be of independent interest.

7.1 Special Sequence Spaces, and Banach Algebra Theory

We start with the consideration of some sequence spaces in a Banach algebra framework. For an introduction to Banach algebra theory see, e.g., Rudin [26]. The following results can be found in Rogozin [24, 25], and for completeness they are recalled here.

For a sequence of positive real weights $(\sigma_n)_{n\geq 0}$, consider the following norms,

$$\|a\|_{\infty,\sigma} = \sup_{m\geq 0} |a_m|\sigma_m + \sum_{n=0}^{\infty} |a_n|, \qquad \|a\|_1 = \sum_{n=0}^{\infty} |a_n|, \qquad a = (a_n)_{n\geq 0} \subset \mathbb{C},$$

and the spaces

$$\ell^1 = \{\, a = (a_n)_{n\geq 0} \subset \mathbb{C} \mid \|a\|_1 < \infty \,\}, \qquad \ell_\sigma^\infty = \{\, a = (a_n)_{n\geq 0} \subset \mathbb{C} \mid \|a\|_{\infty,\sigma} < \infty \,\},$$
$$c_\sigma^0 = \{\, a \in \ell_\sigma^\infty \mid a_n\sigma_n \to 0 \ \text{ as } n \to \infty \,\}.$$

We obviously have $c_\sigma^0 \subset \ell_\sigma^\infty \subset \ell^1$. By using the canonical identification $a(\xi) = \sum_{n=0}^\infty a_n \xi^n$, the spaces $c_\sigma^0, \ell_\sigma^\infty$ and ℓ^1 can be considered as function algebras on

$$\mathscr{D} = \{\, \xi \in \mathbb{C} \mid |\xi| \leq 1 \,\},$$

the closed disc with center 0 and radius 1. We are mainly interested in positive weights $(\sigma_n)_{n\geq 0}$ which satisfy $\sum_{n=0}^\infty \sigma_n^{-1} < \infty$. In that case, $\sup_{m\geq 0} |a_m|\sigma_m$ for $(a_n)_{n\geq 0} \in \ell_\sigma^\infty$ defines a norm on ℓ_σ^∞ which is equivalent to the given norm $\|\cdot\|_{\infty,\sigma}$. In particular, if $\sigma_0 = 1$ and $\sigma_n = n^\beta$ for $n = 1, 2, \ldots$ $(\beta > 1)$, then ℓ_σ^∞ is the space of sequences $(a_n)_{n\geq 0}$ satisfying $a_n = \mathscr{O}(n^{-\beta})$ as $n \to \infty$. In the sequel we assume that

$$\sigma_n \leq c\sigma_j, \quad \tfrac{n}{2} \leq j \leq n, \quad n \geq 0, \tag{39}$$

holds for some finite constant $c > 0$. We state without proof the following elementary result (cf. [26] for part (a) of the proposition, and [24, 25] for parts (b) and (c)).

Proposition 2 *Let $\sigma_0, \sigma_1, \ldots$ be positive weights satisfying condition (39).*

(a) The space ℓ^1, equipped with convolution $(a * b)_n = \sum_{j=0}^n a_{n-j}b_j, n \geq 0$, for $a, b \in \ell^1$, is a commutative complex Banach algebra, with unit $e = (1, 0, 0, \ldots)$.

(b) The space ℓ_σ^∞ is a subalgebra of ℓ^1, i.e., it is closed with respect to addition, scalar multiplication and convolution. The norm $\|\cdot\|_{\infty,\sigma}$ is complete on ℓ_σ^∞ and satisfies

$$\|a * b\|_{\infty,\sigma} \leq (2c+1)\|a\|_{\infty,\sigma} \cdot \|b\|_{\infty,\sigma}, \qquad a, b \in \ell_\sigma^\infty, \tag{40}$$

where c is taken from estimate (39).

(c) The statements of (b) are also valid for the space c_σ^0 (instead of ℓ_σ^∞), supplied with the norm $\|\cdot\|_{\infty,\sigma}$.

The following proposition is based on the fact that the subalgebra generated by $a(\xi) = \xi = (0, 1, 0, 0, \ldots)$ is dense in the space ℓ^1 and in c_σ^0 as well, i.e., both spaces are single-generated in fact.

Proposition 3 (Rogozin [24]) *Let $\sigma_0, \sigma_1, \ldots$ be positive weights satisfying condition (39). The spaces ℓ^1 and c_σ^0 are inverse-closed, i.e., for each $a \in \ell^1$ with $a(\xi) \neq 0$, $\xi \in \mathscr{D}$, one has $[a(\xi)]^{-1} \in \ell^1$, and for each $a \in c_\sigma^0$ with $a(\xi) \neq 0, \xi \in \mathscr{D}$, one has $[a(\xi)]^{-1} \in c_\sigma^0$.*

For the ℓ_1-case, this is Wiener's theorem, cf., e.g., Rudin [26]. The space ℓ_σ^∞ is not single-generated but still inverse-closed which will be used in the following. The proof is taken from Rogozin [25] and is stated here for completeness.

Proposition 4 (Rogozin [25]) *For positive weights $(\sigma_n)_{n\geq 0}$ satisfying condition (39), the space ℓ_σ^∞ is inverse-closed, i.e., for each $a \in \ell_\sigma^\infty$ with $a(\xi) \neq 0$ for $\xi \in \mathscr{D}$ one has $[a(\xi)]^{-1} \in \ell_\sigma^\infty$.*

Proof Consider $a(\xi) = \sum_{n=0}^\infty a_n \xi^n \in \ell_\sigma^\infty$ with $a(\xi) \neq 0$ for $\xi \in \mathscr{D}$. Then a is invertible in ℓ^1 (cf. Proposition 3), i.e., $1/a(\xi) = \sum_{n=0}^\infty a_n^{(-1)} \xi^n \in \ell^1$. Let us assume contradictory that $1/a(\xi) \notin \ell_\sigma^\infty$. This means that $\limsup_{n\to\infty} |a_n^{(-1)}|\sigma_n = \infty$ and then

$$\kappa_n = \max_{0\leq m\leq n} |a_m^{(-1)}|\sigma_m \to \infty \quad \text{as } n \to \infty, \tag{41}$$

and $\kappa_{n+1} \geq \kappa_n > 0$ for $n = 0, 1, \ldots$. Let $\widetilde{\sigma}_n = \sigma_n/\kappa_n$ for $n = 0, 1, \ldots$. We have

$$0 < \widetilde{\sigma}_n = \frac{\sigma_n}{\kappa_n} \leq \frac{\sigma_n}{\kappa_j} \leq c\frac{\sigma_j}{\kappa_j} = c\widetilde{\sigma}_j, \quad \tfrac{n}{2} \leq j \leq n,$$

so the space $c_{\widetilde{\sigma}}^0 = \{a \in \ell_{\widetilde{\sigma}}^\infty \mid a_n\widetilde{\sigma}_n \to 0 \text{ as } n \to \infty\}$ with $\widetilde{\sigma} = (\widetilde{\sigma}_n)_{n\geq 0}$ is a Banach algebra which is inverse-closed (cf. Propositions 2 and 3).

By assumption we have $\sup_{n\geq 0} |a_n|\sigma_n < \infty$, and then $|a_n|\widetilde{\sigma}_n \to 0$ as $n \to \infty$. From Proposition 3 it then follows

$$|a_n^{(-1)}|\widetilde{\sigma}_n \to 0 \quad \text{as } n \to \infty. \tag{42}$$

However, it follows from (41) that for some infinite subset $\mathbf{N} \subset \mathbb{N}$ we have

$$\kappa_n = |a_n^{(-1)}|\sigma_n \quad \text{for } n \in \mathbf{N}. \tag{43}$$

Otherwise there would exist an $n_1 \geq 1$ with $\kappa_n = \max_{0\leq m\leq n} |a_m^{(-1)}|\sigma_m > |a_n^{(-1)}|\sigma_n$ for $n = n_1, n_1 + 1, \ldots$, which in fact means that $\kappa_{n-1} = \max_{0\leq m\leq n-1} |a_m^{(-1)}|\sigma_m > |a_n^{(-1)}|\sigma_n$ holds, and then $\kappa_n = \kappa_{n-1}$ for $n = n_1, n_1 + 1, \ldots$, a contradiction to (41). From (43) we then get

$$|a_n^{(-1)}|\widetilde{\sigma}_n = |a_n^{(-1)}|\sigma_n/\kappa_n = 1, \quad n \in \mathbf{N},$$

a contradiction to (42). □

7.2 The Power Series $\sum_{n=0}^{\infty}(n+1)^{\alpha}\xi^{n}$

Our analysis continues with a special representation of the power series $\sum_{n=0}^{\infty}(n+1)^{\alpha}\xi^{n}$, and we will make use of the binomial expansion

$$(1-\xi)^{\beta} = \sum_{n=0}^{\infty}(-1)^{n}\tbinom{\beta}{n}\xi^{n} \quad \text{for } \xi \in \mathbb{C},\ |\xi| < 1 \qquad (\beta \in \mathbb{R}), \tag{44}$$

$$(-1)^{n}\tbinom{\beta}{n} = \sum_{s=0}^{m-1} d_{\beta s} n^{-\beta-1-s} + \mathscr{O}(n^{-\beta-1-m}) \quad \text{as } n \to \infty, \tag{45}$$

with certain real coefficients $d_{\beta s}$ for $s = 0, 1, \ldots, m-1$, $m = 0, 1, \ldots$, where $d_{\beta 0} = 1/\Gamma(-\beta)$, $\beta \neq 0, 1, \ldots$, cf. e.g., equation (6.1.47) in Abramowitz and Stegun [1]. We need the following result.

Lemma 5 *For* $0 < \alpha < 1$ *we have, with some coefficients* $r_0, r_1, \ldots,$

$$\frac{1}{\Gamma(\alpha+1)}\sum_{n=0}^{\infty}(n+1)^{\alpha}\xi^{n} = (1-\xi)^{-\alpha-1} r(\xi) \ \textit{for}\ \xi \in \mathbb{C},\ |\xi| < 1, \tag{46}$$

$$\textit{with}\ r(\xi) = \sum_{n=0}^{\infty} r_n \xi^{n}, \quad r(1) = 1, \quad r_n = \mathscr{O}(n^{-\alpha-2})\ \textit{as}\ n \to \infty. \tag{47}$$

Proof We first observe that, for each $m \geq 0$, there exist real coefficients $c_0, \ldots, c_{m-1}$ with

$$\frac{1}{\Gamma(\alpha+1)}\sum_{n=0}^{\infty}(n+1)^{\alpha}\xi^{n} = \sum_{j=0}^{m-1} c_j (1-\xi)^{-\alpha-1+j} + s(\xi) \ \text{ for } \xi \in \mathbb{C},\ |\xi| < 1, \tag{48}$$

with $s(\xi) = \sum_{n=0}^{\infty} s_n \xi^{n}$, where $s_n = \mathscr{O}(n^{\alpha-m})$ as $n \to \infty$, and we have $c_0 = 1$. This follows by comparing the coefficients in the Taylor expansion $\frac{1}{\Gamma(\alpha+1)}(n+1)^{\alpha} = \sum_{t=0}^{m-1} e_t n^{\alpha-t} + \mathscr{O}(n^{\alpha-m})$ with the coefficients in the expansions considered in (44) and (45).

A reformulation of (48) gives, with $m = 4$,

$$\frac{1}{\Gamma(\alpha+1)}\sum_{n=0}^{\infty}(n+1)^{\alpha}\xi^{n} = (1-\xi)^{-\alpha-1}\Big(\sum_{j=0}^{3} c_j (1-\xi)^{j} + (1-\xi)^{\alpha+1} s(\xi)\Big)$$

$$\text{for } \xi \in \mathbb{C},\ |\xi| < 1, \quad \text{with } s(\xi) = \sum_{n=0}^{\infty} s_n \xi^{n}, \quad s_n = \mathscr{O}(n^{\alpha-4}) \quad \text{as } n \to \infty.$$

The statement of the lemma now follows from statement (b) of Proposition 2, applied with $\sigma_0 = 1$ and $\sigma_n = n^{\alpha+2}$ for $n = 1, 2, \ldots$, and from (44), (45) applied with $\beta = \alpha + 1, m = 0$. □

7.3 Asymptotical Behavior of the Coefficients of $[\omega(\xi)]^{-1}$

As a consequence of Lemma 5 we obtain the following representation.

Corollary 3 *For the quadrature weights $\omega_0, \omega_1, \ldots$ considered in (9) we have, with the power series r from (46), (47),*

$$\omega(\xi) = \sum_{n=0}^{\infty} \omega_n \xi^n = (1-\xi)^{-\alpha} r(\xi) \ \textit{for} \ \xi \in \mathbb{C}, \ |\xi| < 1. \tag{49}$$

Proof The two power series $\sum_{n=0}^{\infty}(n+1)^{\alpha}\xi^n$ and $\omega(\xi) = \sum_{n=0}^{\infty} \omega_n \xi^n$ with coefficients as in (9) are obviously related as follows,

$$\sum_{n=0}^{\infty} \omega_n \xi^n = \frac{1-\xi}{\Gamma(\alpha+1)} \sum_{n=0}^{\infty} (n+1)^{\alpha} \xi^n.$$

The representation (46) now implies the statement of the corollary. □

Inverting (49) immediately gives the power series representation

$$\sum_{n=0}^{\infty} \omega_n^{(-1)} \xi^n = (1-\xi)^{\alpha} [r(\xi)]^{-1}, \tag{50}$$

where $\omega_n^{(-1)}$ denote the coefficients of the inverse of the power series $\omega(\xi) = \sum_{n=0}^{\infty} \omega_n \xi^n$, cf. (23).

Below we examine the asymptotic behavior of the coefficients in the power series

$$[r(\xi)]^{-1} = \sum_{n=0}^{\infty} r_n^{(-1)} \xi^n. \tag{51}$$

Lemma 6 *We have $r_n^{(-1)} = \mathscr{O}(n^{-\alpha-2})$ as $n \to \infty$.*

Proof It follows from (47) that the power series r considered in (46) satisfies $r \in \ell_{\sigma}^{\infty}$ for the specific choice $\sigma_0 = 1$ and $\sigma_n = n^{\alpha+2}$ for $n \ge 1$. In addition we have

$$r(\xi) \neq 0 \ \text{ for } \ \xi \in \mathbb{C}, \ |\xi| \le 1, \tag{52}$$

which is proven below. From (52) and Proposition 4 we then obtain $r_n^{(-1)} = \mathscr{O}(n^{-\alpha-2})$ as $n \to \infty$. So it remains to show that (52) holds. For this purpose we consider a reformulation of (49),

$$r(\xi) = (1-\xi)^\alpha \sum_{n=0}^{\infty} \omega_n \xi^n \quad \text{for } \xi \in \mathbb{C},\ |\xi| < 1.$$

We have

$$\Big| \sum_{n=0}^{\infty} \omega_n \xi^n \Big| \ge \frac{1}{2\Gamma(\alpha+1)} \quad \text{for } \xi \in \mathbb{C},\ |\xi| < 1, \tag{53}$$

a proof of (53) is presented in the next section. Since $r(1) \neq 0$ and r is continuous on $\{\, \xi \in \mathbb{C} \mid |\xi| \le 1 \,\}$, estimate (53) then implies (52) as desired, and thus the statement of the lemma is proved. □

Property (52) in fact means that the product midpoint rule is zero-stable; see Cameron and McKee [6] for an introduction of this notation for weakly singular Volterra integral equations.

We are now in a position to continue with the verification of the asymptotical behavior (26) for the coefficients of the power series $[\omega(\xi)]^{-1}$. From the representation (44), (45) with $\beta = \alpha$ it follows that the coefficients in the expansion $(1-\xi)^\alpha = \sum_{n=0}^{\infty} (-1)^n \binom{\alpha}{n} \xi^n$ satisfy $(-1)^n \binom{\alpha}{n} = \mathscr{O}(n^{-\alpha-1})$ as $n \to \infty$. This and Lemma 6 (which in particular means $r_n^{(-1)} = \mathscr{O}(n^{-\alpha-1})$) and part (b) of Proposition 2, applied with $\sigma_0 = 1$ and $\sigma_n = n^{\alpha+1}$ for $n \ge 1$, finally results in the desired estimate (26) for the coefficients of the power series $[\omega(\xi)]^{-1}$.

7.4 The Proof of the Lower Bound (53)

To complete our proof of (26), we need to show that (53) holds. We start with a useful lemma.

Lemma 7 *The quadrature weights $\omega_0, \omega_1, \ldots$ in (9) are positive and satisfy $\sum_{n=0}^{\infty} \omega_n = \infty$. In addition we have*

$$\frac{\omega_{n+1}}{\omega_n} > \frac{\omega_n}{\omega_{n-1}} \quad \textit{for } n = 1, 2, \ldots\ . \tag{54}$$

Proof It follows immediately from the definition that the coefficients $\omega_0, \omega_1, \ldots$ are positive. The identity $\sum_{n=0}^{\infty} \omega_n = \infty$ is obvious, and we next present a proof of the inequality (54). Using the notation

$$f(x) = x^\alpha \quad \text{for } x \ge 0$$

we obtain the following,

$$\frac{\omega_n}{\omega_{n-1}} = \frac{f(n+1)-f(n)}{f(n)-f(n-1)} \overset{(*)}{=} \frac{f'(t_n)}{f'(t_n-1)} = \big(1-\tfrac{1}{t_n}\big)^{1-\alpha} =: h(t_n) \text{ for } n = 1,2,\ldots,$$

with some real number $n < t_n < n+1$. Here, the identity $(*)$ follows from the generalized mean value theorem. The function $h(s)$ is monotonically increasing for $s > 0$ which yields estimate (54). This completes the proof of the lemma. □

For results similar to those in Lemma 7, see Eggermont [8, 10] and Linz [18, Section 10.4]. It follows from Lemma 7 that the conditions of the following lemma are satisfied for $g_n = c\omega_n,\ n = 0,1,\ldots,$ with $c > 0$ arbitrary but fixed.

Lemma 8 (cf. Kaluza [17]; see also Szegö [27], Hardy [14], and Linz [18]) *Let $g_0, g_1, \ldots$ be real numbers satisfying*

$$g_n > 0 \textit{ for } n = 0,1,\ldots, \qquad \frac{g_{n+1}}{g_n} > \frac{g_n}{g_{n-1}} \textit{ for } n = 1,2,\ldots\,. \tag{55}$$

Then the inverse $[g(\xi)]^{-1}$ of the power series $g(\xi) = \sum_{n=0}^{\infty} g_n\xi^n$ can be written as follows,

$$[g(\xi)]^{-1} = c_0 - \sum_{n=1}^{\infty} c_n\xi^n, \tag{56}$$

with coefficients $c_0, c_1, \ldots$ satisfying $c_n > 0$ for $n = 0,1,\ldots$. If moreover $\sum_{n=0}^{\infty} g_n = \infty$ holds and the power series $g(\xi) = \sum_{n=0}^{\infty} g_n\xi^n$ has convergence radius 1, then we have $\sum_{n=1}^{\infty} c_n = c_0$.

Proof Lemma 8 is Theorem 22 on page 68 of Hardy [14]. The proof of $c_n > 0$ for $n = 0,1,\ldots$ is presented there in full detail, and we do not repeat the steps here. However, the proof of $\sum_{n=1}^{\infty} c_n = c_0$ is omitted there, so below we present some details of this proof. Condition (55) and the assumption on the convergence radius of the power series $g(\xi)$ means $g_{n+1}/g_n \to 1$ as $n \to \infty$. The second condition in (55) then implies $0 < g_{n+1} < g_n$ for $n = 0,1,\ldots$. From $c_n \geq 0$ for $n = 0,1,\ldots$ we obtain $g_{n-1}\sum_{j=1}^{n} c_j \leq \sum_{j=1}^{n} g_{n-j}c_j = g_n c_0$ for $n = 1,2,\ldots$. The latter identity follows from the representation (56). Thus

$$\sum_{j=1}^{n} c_j \leq \frac{g_n}{g_{n-1}}c_0 \leq c_0 \text{ for } n = 1,2,\ldots\,.$$

The latter inequality means that $c(\xi) = c_0 - \sum_{j=1}^{\infty} c_j \xi^j$ is absolutely convergent on the closed unit disc $\{ \xi \in \mathbb{C} \mid |\xi| \le 1 \}$ and hence is continuous on this set. This finally gives

$$0 = \lim_{0<x\to 1} \frac{1}{\sum_{j=0}^{\infty} g_j x^j} = c_0 - \lim_{0<x\to 1} \sum_{j=1}^{\infty} c_j x^j = c_0 - \sum_{j=1}^{\infty} c_j.$$

This completes the proof of the lemma. □

The following lemma is closely related to results in Erdős et al. [11]. A detailed proof can be found in [22].

Lemma 9 *Let $c_1, c_2, \ldots$ be a sequence of real numbers satisfying $c_n > 0$ for $n = 1, 2, \ldots,$ and $\sum_{n=1}^{\infty} c_n = \frac{1}{2}$. Then the power series $q(\xi) = \frac{1}{2} - \sum_{n=1}^{\infty} c_n \xi^n$ satisfies $|q(\xi)| < 1$ for each complex number ξ with $|\xi| \le 1$.*

We are now in a position to present a proof of the lower bound (53). In fact, from Lemma 7 it follows that the coefficients of the power series $g(\xi) = 2\Gamma(\alpha+1)\omega(\xi)$ with $\omega(\xi)$ as in (49) satisfy the conditions of Lemma 8, and in addition $g_0 = 2$ holds. This implies that the coefficients of the power series

$$\frac{1}{2\Gamma(\alpha+1)\omega(\xi)} = c_0 - \sum_{n=1}^{\infty} c_n \xi^n$$

satisfy $c_n > 0$ for $n = 0, 1, \ldots$ and $\sum_{n=1}^{\infty} c_n = c_0 = 1/2$. Lemma 9 then implies that $2\Gamma(\alpha+1)|\omega(\xi)| \ge 1$ and thus $|\omega(\xi)| \ge \frac{1}{2\Gamma(\alpha+1)}$ for $\xi \in \mathbb{C}, |\xi| < 1$. This is the desired estimate (53) needed in the proof of Lemma 6.

8 Appendix 2: Proof of Theorem 1

1. We apply the representations (8) and (10) with $\varphi = \varphi_n$, where

$$\varphi_n(y) = k(x_n, y)u(y), \quad 0 \le y \le x_n.$$

Scheme (22) then results in the following,

$$h^\alpha \sum_{j=1}^{n} \omega_{n-j} k(x_n, x_{j-1/2}) e_{j-1/2}^{\delta} = (E_h^\alpha \varphi_n)(x_n) + f_n^\delta - f(x_n) \ \text{ for } n = 1, \ldots, N, \tag{57}$$

where

$$e_{j-1/2}^{\delta} = u_{j-1/2}^{\delta} - u(x_{j-1/2}), \quad j = 1, 2, \ldots, N.$$

2. We next consider a matrix-vector formulation of (57). As a preparation we consider the matrix $A_h \in \mathbb{R}^{N \times N}$ given by

$$A_h = \begin{pmatrix} \omega_0 k_{1,1/2} & 0 & \cdots & \cdots & 0 \\ \omega_1 k_{2,1/2} & \omega_0 k_{2,3/2} & \ddots & & 0 \\ \vdots & \omega_1 k_{3,3/2} & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \omega_{N-1} k_{N,1} & \cdots & \cdots & \omega_1 k_{N,N-3/2} & \omega_0 k_{N,N-1/2} \end{pmatrix}$$

with the notation

$$k_{n,j-1/2} = k(x_n, x_{j-1/2}) \text{ for } 1 \le j \le n \le N.$$

Additionally we consider the vectors

$$\varDelta_h^{\delta} = (e_{j-1/2}^{\delta})_{1 \le j \le N}, \;\; R_h = ((E_h^{\alpha} \varphi_n)(x_n))_{1 \le n \le N}, \;\; F_h^{\delta} = (f_n^{\delta} - f(x_n))_{1 \le n \le N}. \tag{58}$$

Using these notations, the linear system of equations (57) can be written as

$$h^{\alpha} A_h \varDelta_h^{\delta} = R_h + F_h^{\delta}, \quad \text{with } \|F_h^{\delta}\|_{\infty} \le \delta, \tag{59}$$

where $\|\cdot\|_{\infty}$ denotes the maximum norm on $\mathbb{R}^N$. In addition, occasionally we consider a modified error equation which can easily be derived from (59) by applying the matrix D_h to both sides of that equation:

$$h^{\alpha} D_h A_h \varDelta_h^{\delta} = D_h R_h + D_h F_h^{\delta}, \tag{60}$$

where the matrix $D_h \in \mathbb{R}^{N\times N}$ is given by

$$D_h = \begin{pmatrix} \omega_0^{(-1)} & 0 & \cdots & \cdots & 0 \\ \omega_1^{(-1)} & \omega_0^{(-1)} & 0 & & 0 \\ \omega_2^{(-1)} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \omega_{N-1}^{(-1)} & \cdots & \cdots & \omega_1^{(-1)} & \omega_0^{(-1)} \end{pmatrix}. \tag{61}$$

3. For a further treatment of the identity (59) and the modification (60), we next show

$$\|D_h\|_\infty = \mathscr{O}(1), \quad \|(D_hA_h)^{-1}\|_\infty = \mathscr{O}(1), \quad \|{A_h}^{-1}\|_\infty = \mathscr{O}(1) \quad \text{as } h \to 0, \tag{62}$$

where $\|\cdot\|_\infty$ denotes the matrix norm induced by the maximum vector norm on $\mathbb{R}^N$. In fact, the estimate $\|D_h\|_\infty = \mathscr{O}(1)$ as $h \to 0$ follows immediately from the decay of the coefficients of the inverse of the generating function ω, cf. estimate (26). For the proof of the second statement in (62) we use the fact that the matrix D_hA_h can be written in the form $D_hA_h = I_h + K_h$, where $I_h \in \mathbb{R}^{N\times N}$ denotes the identity matrix, and $K_h = (k_{h,n,j}) \in \mathbb{R}^{N\times N}$ denotes some lower triangular matrix which satisfies $\max_{1\le j\le n\le N} |k_{h,n,j}| = \mathscr{O}(h)$ as $h \to 0$, cf. the proof of Lemma 4.2 in Eggermont [9] for more details. We only note that here it is taken into account that the kernel function is uniformly Lipschitz continuous with respect to the first variable, cf. part (c) of Assumption 1. This representation of D_hA_h and the discrete version of Gronwall's inequality now yields $\|(D_hA_h)^{-1}\|_\infty = \mathscr{O}(1)$ as $h \to 0$. The third estimate in (62) follows immediately from the other two estimates considered in (62).

4. In view of (59)–(62), it remains to take a closer look at the representations of the quadrature error considered in Lemma 1. We consider different situations for p and constantly make use of the fact that, for some finite constant $L \ge 0$, we have

$$\varphi_n \in F_L^p[0, x_n] \ \text{ for } n = 1, 2, \ldots, N, \tag{63}$$

cf. Assumption 1.

(i) In the case $p \le 1$ we proceed in two different ways. The first one turns out to be useful for the case $\alpha \le \frac{1}{2}$, while the other one uses partial summation and is useful for the case $\alpha \ge \frac{1}{2}$.

- Our first approach proceeds with (59), and we assume $\alpha < p \le 1$ in this case. We then easily obtain, cf. (16), (63),

$$\|R_h\|_\infty = \max_{1\le n\le N} |(E_h^\alpha \varphi_n)(x_n)| = \mathscr{O}(h^p) \quad \text{as } h \to 0,$$

and then, cf. (59) and (62), $\|\Delta_h^\delta\|_\infty = \mathscr{O}(h^{-\alpha}(h^p + \delta)) = \mathscr{O}(h^{p-\alpha} + \frac{\delta}{h^\alpha})$.

- In our second approach we would like to proceed with (60), and we need to consider the vector $D_h R_h \in \mathbb{R}^N$ in more detail. For this purpose we assume that $1 - \alpha < p \le 1$ holds, and we introduce the notation

$$r_n = (E_h^\alpha \varphi_n)(x_n), \quad n = 1, 2, \ldots, N.$$

Partial summation, applied to the nth entry of $D_h R_h$, gives

$$(D_h R_h)_n = \sum_{j=1}^{n} \omega_{n-j}^{(-1)} r_j = \beta_n r_1 + \sum_{\ell=1}^{n-1} \beta_{n-\ell}(r_{\ell+1} - r_\ell), \tag{64}$$

where

$$0 \le \beta_n := \sum_{\ell=0}^{n-1} \omega_\ell^{(-1)} = -\sum_{\ell=n}^{\infty} \omega_\ell^{(-1)} \text{ for } n = 1, 2, \ldots, \tag{65}$$

cf. Lemma 2. We thus have, cf. again Lemma 2,

$$\beta_n = \mathscr{O}(n^{-\alpha}) \quad \text{as } n \to \infty, \tag{66}$$

and thus

$$\sum_{\ell=1}^{n-1} \beta_\ell = \mathscr{O}(N^{1-\alpha}) = \mathscr{O}(h^{\alpha-1}) \quad \text{as } h \to 0 \tag{67}$$

uniformly for $n = 1, 2, \ldots, N$. Estimate (5), representation (16) and Hölder continuity (63) imply

$$|r_1| = |(E_h^\alpha \varphi_1)(x_1)| = \mathscr{O}(h^{p+\alpha}),$$

and we next consider the differences $r_{\ell+1} - r_\ell$ in more detail. For this purpose we introduce short notation for the interpolation error,

$$\chi_n(y) = \varphi_n(y) - q_h \varphi_n(y) \text{ for } 0 \le y \le x_n, \quad n = 1, 2, \ldots, N.$$

We then have

$$
\begin{aligned}
r_{\ell+1} - r_\ell &= \frac{1}{\Gamma(\alpha)}\Big(\int_0^{x_{\ell+1}} (x_{\ell+1}-y)^{\alpha-1}\chi_{\ell+1}(y)\,dy - \int_0^{x_\ell} (x_\ell - y)^{\alpha-1}\chi_\ell(y)\,dy\Big) \\
&= \frac{1}{\Gamma(\alpha)}\int_{x_\ell}^{x_{\ell+1}} (x_{\ell+1}-y)^{\alpha-1}\chi_{\ell+1}(y)\,dy \\
&\qquad + \frac{1}{\Gamma(\alpha)}\int_0^{x_\ell} (x_{\ell+1}-y)^{\alpha-1}(\chi_{\ell+1}-\chi_\ell)(y)\,dy \\
&\qquad + \frac{1}{\Gamma(\alpha)}\int_0^{x_\ell} ((x_{\ell+1}-y)^{\alpha-1} - (x_\ell - y)^{\alpha-1})\chi_\ell(y)\,dy =: s_1 + s_2 + s_3.
\end{aligned}
$$

We have $s_1 = \mathscr{O}(h^{p+\alpha})$ which easily follows from $\sup_{0\le y\le x_{\ell+1}} |\chi_{\ell+1}(y)| = \mathscr{O}(h^p)$. Moreover, first order Taylor expansions of the kernel k with respect to the first variable at the grid point x_ℓ gives for $x_{j-1} \le y \le x_j$ $(1 \le j \le \ell)$ the following,

$$
\begin{aligned}
(\chi_{\ell+1}-\chi_\ell)(y) &= k(x_{\ell+1}, y)u(y) - k(x_{\ell+1}, x_{j-1/2})u(x_{j-1/2}) \\
&\qquad\qquad - \big\{k(x_\ell, y)u(y) - k(x_\ell, x_{j-1/2})u(x_{j-1/2})\big\} \\
&= \big(\tfrac{\partial k}{\partial x}(x_\ell, y)h + \mathscr{O}(h^2)\big)u(y) - \big(\tfrac{\partial k}{\partial x}(x_\ell, x_{j-1/2})h + \mathscr{O}(h^2)\big)u(x_{j-1/2}) \\
&= h\big(\tfrac{\partial k}{\partial x}(x_\ell, y)u(y) - \tfrac{\partial k}{\partial x}(x_\ell, x_{j-1/2})u(x_{j-1/2})\big) + \mathscr{O}(h^2) = \mathscr{O}(h^{p+1}),
\end{aligned}
$$

and this implies $s_2 = \mathscr{O}(h^{p+1})$. Finally,

$$
\begin{aligned}
|s_3| &\le \frac{L}{\Gamma(\alpha)} h^p \int_0^{x_\ell} (x_\ell - y)^{\alpha-1} - (x_{\ell+1}-y)^{\alpha-1}\,dy \\
&= \frac{L}{\Gamma(\alpha+1)} h^{p+\alpha}(1 + \ell^\alpha - (\ell+1)^\alpha) = \mathscr{O}(h^{p+\alpha}).
\end{aligned}
$$

Summation gives $s_1 + s_2 + s_3 = \mathscr{O}(h^{p+\alpha})$, and (64) finally results in (see also (67))

$$
(D_h R_h)_n = \mathscr{O}(h^{p+\alpha} + h^{\alpha-1}h^{p+\alpha}) = \mathscr{O}(h^{p+2\alpha-1})
$$

uniformly for $n = 1, 2, \ldots, N$. We note that this estimate is useful for $\alpha \ge \frac{1}{2}$ only. We are now in a position to proceed with (60):

$$
\|\Delta_h^\delta\|_\infty = \mathscr{O}\big(h^{-\alpha}\|D_h R_h\|_\infty + \tfrac{\delta}{h^\alpha}\big) = \mathscr{O}(h^{p+\alpha-1} + \tfrac{\delta}{h^\alpha}) \quad \text{as } (h,\delta) \to 0,
$$

where also (62) has been used. This gives the desired result.

(ii) We now proceed with the case $1 < p \le 2$. Preparatory results are given in the present item (ii), and in item (iii) the final steps will be done.

Representation (17) of the integration error gives

$$(E_h^\alpha \varphi_n)(x_n) = h^{\alpha+1} s_n + t_n, \quad \text{with } s_n = \sum_{j=1}^{n} \tau_{n-j} \varphi_n'(x_{j-1/2}),$$
$$t_n = (\mathscr{V}^\alpha(\varphi_n - r_h \varphi_n))(x_n),$$

for $n = 1, 2, \ldots, N$, or, in vector notation (for the definition of R_h see (58))

$$R_h = h^{\alpha+1} S_h + T_h, \quad \text{with } S_h = (s_n)_{n=1,\ldots,N}, \quad T_h = (t_n)_{n=1,\ldots,N}. \tag{68}$$

In view of (59) and (60), we need to consider the four vectors $S_h, D_h S_h, T_h$ and $D_h T_h \in \mathbb{R}^N$ in more detail.

- From the summability of the coefficients τ_s, cf. (20), it immediately follows that $\|S_h\|_\infty = \mathscr{O}(1)$ as $h \to 0$.
- In the case $p > 2 - \alpha$ and $u(0) = u'(0) = 0$, it turns out to be useful to consider the vector $D_h S_h$. Partial summation applied to the nth entry of $D_h S_h$ gives

$$(D_h S_h)_n = \sum_{\ell=1}^{n} \omega_{n-\ell}^{(-1)} s_\ell = \beta_n s_1 + \sum_{\ell=1}^{n-1} \beta_{n-\ell}(s_{\ell+1} - s_\ell), \tag{69}$$

with β_n given by (65). The smoothness property (63), the assumption $u(0) = u'(0) = 0$ and the boundedness $\beta_n = \mathscr{O}(1)$, cf. (66), imply that $\beta_n s_1 = \beta_n \tau_0 \varphi_1'(x_{1/2}) = \mathscr{O}(h^{p-1})$. In addition,

$$\begin{aligned} s_{\ell+1} - s_\ell &= \sum_{j=1}^{\ell+1} \tau_{\ell+1-j} \varphi_{\ell+1}'(x_{j-1/2}) - \sum_{j=1}^{\ell} \tau_{\ell-j} \varphi_\ell'(x_{j-1/2}) \\ &= \tau_\ell \varphi_{\ell+1}'(x_{1/2}) + \sum_{j=1}^{\ell} \tau_{\ell-j} \big(\varphi_{\ell+1}'(x_{j+1/2}) - \varphi_\ell'(x_{j-1/2})\big) = \mathscr{O}(h^{p-1}) \end{aligned}$$

uniformly for $\ell = 1, 2, \ldots, N-1$. The considered partial summation (69) thus finally results in (see also (65), (67))

$$\|D_h S_h\|_\infty = \mathscr{O}(h^{p-1}) + \mathscr{O}(h^{\alpha-1+p-1}) = \mathscr{O}(h^{p+\alpha-2}). \tag{70}$$

- It follows from (15) that $\|T_h\|_\infty = \mathscr{O}(h^p)$ as $h \to 0$. This estimate will be useful in the case $\alpha \le \frac{1}{2}$.
- We next consider the vector $D_h T_h$ in more detail. Partial summation applied to the nth entry of $D_h T_h$ gives

$$(D_h T_h)_n = \sum_{\ell=1}^{n} \omega_{n-\ell}^{(-1)} t_\ell = \beta_n t_1 + \sum_{\ell=1}^{n-1} \beta_{n-\ell}(t_{\ell+1} - t_\ell). \tag{71}$$

We have

$$t_1 = \mathscr{O}(h^{p+\alpha}), \qquad t_{\ell+1} - t_\ell = \mathscr{O}(h^{p+\alpha}),$$

uniformly for $\ell = 1, 2, \ldots, N-1$. This in fact is verified similarly as in the second item of part 4(i) of this proof, this time with second order Taylor expansions of the kernel k as well as first order Taylor expansions of $\frac{\partial k}{\partial y}$ with respect to the first variable, respectively. We omit the simple but tedious computations. This gives

$$\|D_h T_h\|_\infty = \mathscr{O}(h^{p+\alpha}) + \mathscr{O}(h^{\alpha-1+p+\alpha}) = \mathscr{O}(h^{p+2\alpha-1}). \tag{72}$$

This estimate will be useful in the case $\alpha \geq \frac{1}{2}$.

Notice that the second of the four considered items is the only one where the initial condition $u(0) = u'(0) = 0$ is needed.

(iii) We continue with the consideration of the case $1 < p \leq 2$. The results from (ii) allow us to proceed with (59), (60).

- We first consider the case $\alpha \leq \frac{1}{2}, 1 < p \leq \alpha + 1$. The consistency error representations in item (ii) of the present proof yield $\|R_h\|_\infty = \max_{1\leq n \leq N} |(E_h^\alpha \varphi_n)(x_n)| = \mathscr{O}(h^{\alpha+1}\|S_h\|_\infty + \|T_h\|_\infty) = \mathscr{O}(h^{\alpha+1} + h^p) = \mathscr{O}(h^p)$. From the error equation (59) it then follows $\|\Delta_h^\delta\|_\infty = \mathscr{O}(h^{-\alpha}(h^p + \delta)) = \mathscr{O}(h^{p-\alpha} + \delta/h^\alpha)$.
- We next consider the case $\alpha \geq \frac{1}{2}, 1 < p \leq 2 - \alpha$. The integration error estimates obtained in item (ii) yield $\|D_h R_h\|_\infty = \mathscr{O}(h^{\alpha+1}\|S_h\|_\infty + \|D_h T_h\|_\infty) = \mathscr{O}(h^{\alpha+1} + h^{p+2\alpha-1}) = \mathscr{O}(h^{p+2\alpha-1})$, where the first identity in (62) has been applied. From the error equation (60) it then follows $\|\Delta_h^\delta\|_\infty = \mathscr{O}(h^{-\alpha}(h^{p+2\alpha-1} + \delta)) = \mathscr{O}(h^{p-1+\alpha} + \delta/h^\alpha)$.
- Finally we consider the case $2 - \alpha < p \leq 2$ and $u(0) = u'(0) = 0$. The consistency error estimates in item (ii) yield $\|D_h R_h\|_\infty = \mathscr{O}(h^{\alpha+1}\|D_h S_h\|_\infty + \|D_h T_h\|_\infty) = \mathscr{O}(h^{p+2\alpha-1})$. From the error equation (60) we then obtain the estimate $\|\Delta_h^\delta\|_\infty = \mathscr{O}(h^{-\alpha}(h^{p+2\alpha-1} + \delta)) = \mathscr{O}(h^{p-1+\alpha} + \delta/h^\alpha)$. This completes the proof of the theorem.

References

1. M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions*, 10th edn. (National Bureau of Standards, Dover, New York, 1972)
2. R.S. Anderssen, Stable procedures for the inversion of Abel's equation. IMA J. Appl. Math. **17**, 329–342 (1976)
3. H. Brunner, *Collocation Methods for Volterra Integral and Related Functional Differential Equations* (Cambridge University Press, Cambridge, 2004)

4. H. Brunner, P.J. van der Houwen, *The Numerical Solution of Volterra Equations* (Elsevier, Amsterdam, 1986)
5. A.L. Bughgeim, *Volterra Equations and Inverse Problems* (VSP/de Gruyter, Zeist/Berlin, 1999)
6. R.F. Cameron, S. McKee, High accuracy convergent product integration methods for the generalized Abel equation. J. Integr. Equ. **7**, 103–125 (1984)
7. R.F. Cameron, S. McKee, The analysis of product integration methods for Abel's equation using fractional differentiation. IMA J. Numer. Anal. **5**, 339–353 (1985)
8. P.P.B. Eggermont, A new analysis of the Euler-, midpoint- and trapezoidal-discretization methods for the numerical solution of Abel-type integral equations. Technical report, Department of Computer Science, University of New York, Buffalo, 1979
9. P.P.B. Eggermont, A new analysis of the trapezoidal-discretization method for the numerical solution of Abel-type integral equations. J. Integr. Equ. **3**, 317–332 (1981)
10. P.P.B. Eggermont, Special discretization methods for the integral equations of image reconstruction and for Abel-type integral equations. Ph.D. thesis, University of New York, Buffalo, 1981
11. P. Erdős, W. Feller, H. Pollard, A property of power series with positive coefficients. Bull. Am. Math. Soc. **55**, 201–204 (1949)
12. R. Gorenflo, S. Vessella, *Abel Integral Equations* (Springer, New York, 1991)
13. W. Hackbusch, *Integral Equations* (Birkhäuser, Basel, 1995)
14. G.H. Hardy, *Divergent Series* (Oxford University Press, Oxford, 1948)
15. P. Henrici, *Applied and Computational Complex Analysis*, vol. 1 (Wiley, New York, 1974)
16. B. Kaltenbacher, A convergence analysis of the midpoint rule for first kind Volterra integral equations with noisy data. J. Integr. Equ. **22**, 313–339 (2010)
17. T. Kaluza, Über die Koeffizienten reziproker Funktionen. Math. Z. **28**, 161–170 (1928)
18. P. Linz, *Analytical and Numerical Methods for Volterra Equations*, 1st edn. (SIAM, Philadelphia, 1985)
19. Ch. Lubich, Discretized fractional calculus. SIAM J. Math. Anal. **17**(3), 704–719 (1986)
20. Ch. Lubich, Fractional linear multistep methods for Abel–Volterra integral equations of the first kind. IMA J. Numer. Anal. **7**, 97–106 (1987)
21. R. Plato, Fractional multistep methods for weakly singular Volterra equations of the first kind with noisy data. Numer. Funct. Anal. Optim. **26**(2), 249–269 (2005)
22. R. Plato, The regularizing properties of the composite trapezoidal method for weakly singular Volterra integral equations of the first kind. Adv. Comput. Math. **36**(2), 331–351 (2012)
23. R. Plato, The regularizing properties of multistep methods for first kind Volterra integral equations with smooth kernels. Comput. Methods Appl. Math. **17**(1), 139–159 (2017)
24. B.A. Rogozin, Asymptotics of the coefficients in the Levi–Wiener theorems on absolutely convergent trigonometric series. Sib. Math. J. **14**, 917–923 (1973)
25. B.A. Rogozin, Asymptotic behavior of the coefficients of functions of power series and Fourier series. Sib. Math. J. **17**, 492–498 (1976)
26. W. Rudin, *Functional Analysis*, 2nd edn. (McGraw-Hill, New York, 1991)
27. G. Szegö, Bemerkungen zu einer Arbeit von Herrn Fejér über die Legendreschen Polynome. Math. Z. **25**, 172–187 (1926)
28. U. Vögeli, K. Nedaiasl, S. Sauter, A fully discrete Galerkin method for Abel-type integral equations (2016). arXiv:1612.01285
29. R. Weiss, Product integration for the generalized Abel equation. Math. Comput. **26**, 177–190 (1972)
30. R. Weiss, R.S. Anderssen, A product integration method for a class of singular first kind Volterra equations. Numer. Math. **18**, 442–456 (1972)

Heuristic Parameter Choice in Tikhonov Method from Minimizers of the Quasi-Optimality Function

Toomas Raus and Uno Hämarik

Abstract We consider choice of the regularization parameter in Tikhonov method in the case of the unknown noise level of the data. From known heuristic parameter choice rules often the best results were obtained in the quasi-optimality criterion where the parameter is chosen as the global minimizer of the quasi-optimality function. In some problems this rule fails, the error of the Tikhonov approximation is very large. We prove, that one of the local minimizers of the quasi-optimality function is always a good regularization parameter. We propose some algorithms for finding a proper local minimizer of the quasi-optimality function.

1 Introduction

Let $A \in \mathscr{L}(H, F)$ be a linear bounded operator between real Hilbert spaces. We are interested in finding the minimum norm solution u_* of the equation

$$Au = f_*, \qquad f_* \in \mathscr{R}(A). \tag{1}$$

The range $\mathscr{R}(A)$ may be non-closed and the kernel $\mathscr{N}(A)$ may be non-trivial, so in general this problem is ill-posed. As usually in treatment of ill-posed problems, we assume that instead of exact data f_* noisy data $f \in F$ are given. For the solution of the problem $Au = f$ we consider Tikhonov method (see [6, 36]) where regularized solutions in cases of exact and inexact data have corresponding forms

$$u_\alpha^+ = \left(\alpha I + A^*A\right)^{-1} A^*f_*, \qquad u_\alpha = \left(\alpha I + A^*A\right)^{-1} A^*f$$

and $\alpha > 0$ is the regularization parameter.

T. Raus · U. Hämarik (✉)
University of Tartu, Tartu, Estonia
e-mail: toomas.raus@ut.ee; uno.hamarik@ut.ee

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_12

Denote

$$e_1(\alpha) := \left\| u_\alpha^+ - u_* \right\| + \left\| u_\alpha - u_\alpha^+ \right\| . \tag{2}$$

Due to the well-known estimate $\left\| u_\alpha - u_\alpha^+ \right\| \leq \frac{1}{2}\alpha^{-1/2} \left\| f - f_* \right\|$ (see [6, 36]) the error $\left\| u_\alpha - u_* \right\|$ can be estimated by

$$\left\| u_\alpha - u_* \right\| \leq e_1(\alpha) \leq e_2(\alpha, \left\| f - f_* \right\|) := \left\| u_\alpha^+ - u_* \right\| + \frac{1}{2\sqrt{\alpha}} \left\| f - f_* \right\| . \tag{3}$$

We consider choice of the regularization parameter if the noise level for $\left\| f - f_* \right\|$ is unknown. The parameter choice rules which do not use the noise level information are called heuristic rules. Many heuristic rules are proposed, well known are the quasi-optimality criterion [2, 3, 5, 10, 20–22, 25, 35], L-curve rule [16, 17], GCV-rule [8], Hanke-Raus rule [15], Reginska's rule [33], about other rules see [18, 19, 23, 26]. Heuristic rules are numerically compared in [4, 10, 18, 26]. It is also well known that it is not possible to construct heuristic rule guaranteeing convergence $\left\| u_\alpha - u_* \right\| \to 0$ as the noise level goes to zero (see [1]). Nevertheless the heuristic rules give good results in many problems. The problem is that all these rules may fail in some problems and without additional information about the solution, it is difficult to decide, is the obtained parameter reliable or not.

In this article we propose a new strategy for heuristic parameter choice. It is based on analysis of local minimizers of the function $\psi_Q(\alpha) = \alpha \left\| \frac{du_\alpha}{d\alpha} \right\|$, the global minimizer of which on certain interval $[\alpha_M, \alpha_0]$ is taken for parameter in the quasi-optimality criterion. We will call the parameter α_R in arbitrary rule R as pseudooptimal, if

$$\left\| u_{\alpha_R} - u_* \right\| \leq \text{const} \quad \min_{\alpha > 0} e_1(\alpha)$$

and we show that at least one of local minimizers of $\psi_Q(\alpha)$ has this property. Our approach enables to replace the search of the parameter from the interval $[\alpha_M, \alpha_0]$ by search of the proper parameter from the set L_{min} of the local minimizers of the function $\psi_Q(\alpha)$. We consider also the possibility to restrict the set L_{min} to its subset L_{min}^* still containing at least one pseudooptimal parameter. It occurs that in many problems the restricted set L_{min}^* contains only one local minimizer and this is the pseudooptimal parameter. If the set L_{min}^* contains several local minimizers, we consider different algorithms for choice of the proper parameter from the set L_{min}^*.

The plan of this paper is as follows. In Sect. 2 we consider known rules for choice of the regularization parameter, both in case of known and unknown noise level. We will characterize distinctive properties of considered heuristic rules presenting results of numerical experiments on test problems [17]. In Sect. 3 we consider the set L_{min} of local minimizers of the function $\psi_Q(\alpha)$ and prove that this set contains at least one pseudooptimal parameter. In Sect. 4 we show how to restrict the set L_{min} to the set L_{min}^* still containing at least one pseudooptimal parameter. In Sect. 5 we consider the case if the set L^*_{min} contains several elements and we propose some algorithms for finding proper pseudooptimal parameter. In all sections theoretical results and proposed algorithms are illustrated by results of numerical experiments on test problems [17].

2 Rules for the Choice of the Regularization Parameter

An important problem, when applying regularization methods, is the proper choice of the regularization parameter. The choice of the parameter depends on the information about the noise level.

2.1 Parameter Choice in the Case of Known Noise Level

In case of known noise level δ, $\|f - f_*\| \leq \delta$ we use one of so-called δ-rules, where certain functional $d(\alpha)$ and constants $b_2 \geq b_1 \geq b_0$ (b_0 depends on $d(\alpha)$) are chosen and such regularization parameter $\alpha(\delta)$ is chosen which satisfies $b_1\delta \leq d(\alpha) \leq b_2\delta$.

1) Discrepancy principle (DP) [24, 36]:

$$b_1\delta \leq \|Au_\alpha - f\| \leq b_2\delta, \quad b_1 \geq 1.$$

2) Modified discrepancy principle (Raus-Gfrerer rule) [7, 28]:

$$b_1\delta \leq \|B_\alpha\left(Au_\alpha - f\right)\| \leq b_2\delta, \quad B_\alpha := \alpha^{1/2}\left(\alpha I + AA^*\right)^{-1/2}, \quad b_1 \geq 1.$$

3) Monotone error rule (ME-rule) [14, 34]:

$$b_1\delta \leq \frac{\|B_\alpha\left(Au_\alpha - f\right)\|^2}{\|B_\alpha^2\left(Au_\alpha - f\right)\|} \leq b_2\delta, \quad b_1 \geq 1.$$

The name of this rule is justified by the fact that the chosen parameter α_{ME} satisfies

$$\|u_{\alpha_{\text{ME}}} - u_*\| < \|u_\alpha - u_*\| \qquad \forall\alpha > \alpha_{\text{ME}}.$$

Therefore $\alpha_{\text{ME}} \geq \alpha_{opt} := \text{argmin}\|u_\alpha - u_*\|$ and $b_1 = b_2 = 1$ are recommended.

4) Monotone error rule with post-estimation (MEe-rule) [10, 12, 13, 26, 31]. The inequality $\alpha_{ME} \geq \alpha_{opt}$ suggests to use somewhat smaller parameter than α_{ME}. Extensive numerical experiments suggest to take $b_1 = b_2 = 1$, to compute α_{ME} and to use the post-estimated parameter $\alpha_{MEe} := 0.4\alpha_{ME}$. Then typically

$\|u_{\alpha_{\text{MEe}}} - u_*\| / \|u_{\alpha_{\text{ME}}} - u_*\| \in (0.7, 0.9)$. To our best knowledge in case of exact noise level this MEe-rule gives typically best results from all known rules for the parameter choice.

5) Rule R1 [29]: Let $b_2 \geq b_1 \geq 0.325$. Let $d(\alpha) := \alpha^{-1/2} \left\| A^* B_\alpha^2 (Au_\alpha - f) \right\|$. Choose $\alpha(\delta)$ such that $d(\alpha(\delta)) \geq b_1\delta$, but $d(\alpha) \leq b_2\delta$ for all $\alpha \leq \alpha(\delta)$.

   Note that

$$B_\alpha^2 (Au_\alpha - f) = Au_{2,\alpha} - f, \qquad u_{2,\alpha} = \left(\alpha I + A^*A\right)^{-1} (\alpha u_\alpha + A^*f),$$

   where $u_{2,\alpha}$ is the 2-iterated Tikhonov approximation.

6) Balancing principle [4, 9, 26, 27]. This rule has different forms in different papers, in [9] the form

$$b_1\delta \leq \frac{\sqrt{\alpha}\sqrt{q}\|u_\alpha - u_{\alpha/q}\|}{1-q} \leq b_2\delta, \quad b_1 \geq \frac{3\sqrt{6}}{16} \approx 0.459.$$

Typically balancing principle is implemented by computing a sequence of Tikhonov approximations, but in case of a smooth solution much better approximation than single Tikhonov approximation is simple linear combination of Tikhonov approximations with different parameters — the extrapolated approximation (see [9, 11, 26]). See [32] about effective numerical realization of rules 1)–6).

The last five rules are weakly quasioptimal rules (see [30]) for Tikhonov method. If $\|f - f_*\| \leq \delta$, then we have the error estimate (see (3))

$$\left\| u_{\alpha(\delta)} - u_* \right\| \leq C(b_1, b_2) \inf_{\alpha > 0} e_2(\alpha, \delta) = C(b_1, b_2) \inf_{\alpha > 0} \left[\left\| u_\alpha^+ - u_* \right\| + \frac{1}{2\sqrt{\alpha}}\delta \right].$$

The rules for the parameter choice in case of approximately given noise level are proposed and analysed in [12, 13, 26, 31].

2.2 Parameter Choice in the Case of Unknown Noise Level

If the noise level is unknown, then, as shown by Bakushinskii [1], no rule for choosing the regularization parameter can guarantee the convergence of the regularized solution to the exact one as noise level $\|f - f_*\|$ goes to zero. Nevertheless, some heuristic rules are rather popular, because they often work well in practice and because in applied ill-posed problems the exact noise level is often unknown.

A classical heuristic rule is the quasi-optimality criterion. In Tikhonov method it chooses $\alpha = \alpha_Q$ as the global minimizer of the function

$$\psi_Q(\alpha) = \alpha \left\| \frac{du_\alpha}{d\alpha} \right\| = \alpha^{-1} \left\| A^* B_\alpha^2 (Au_\alpha - f) \right\|. \tag{4}$$

In case of the discrete version of the quasi-optimality criterion we choose $\alpha = \alpha_{QD}$ as the global minimizer of the function $\left\| u_\alpha - u_{q\alpha} \right\|$, where $0 < q < 1$.

The Hanke-Raus rule finds the regularization parameter $\alpha = \alpha_{HR}$ as the global minimizer of the function

$$\psi_{HR}(\alpha) = \alpha^{-1/2} \left\| B_\alpha \left(Au_\alpha - f \right) \right\| .$$

In practice the L-curve rule is popular. This rule uses the graph with log-log scale, on x-axis $\|Au_\alpha - f\|$ and on y-axis $\|u_\alpha\|$. The name of the rule is justified by fact that often the points $(\|Au_\alpha - f\|, \|u_\alpha\|)$ have shape similar to the letter L and parameter α_L which corresponds to the "corner point" is often a good parameter. In the literature several concrete rules for choice of the 'corner point' are proposed. One natural rule is proposed in [33] where global minimum point of the function

$$\psi_{RE}(\alpha) = \|Au_\alpha - f\| \, \|u_\alpha\|^\tau ,$$

with $\tau \geq 1$ is used. In numerical experiments below we used this rule with $\tau = 1$.

Some heuristic rules choose the regularization parameter as global minimizer of a function $\alpha^{-1/2} d(\delta)$ with function $d(\delta)$ from some δ-rule 1)–6) from Sect. 2.1 (see [10]). For example, the quasi-optimality criterion and Hanke-Raus rule use functions $d(\delta)$ from the rules 5) (R1) and 2) (modified discrepancy principle) respectively. In [10] heuristic counterpart of rule 3) (ME-rule) is also studied. We call this rule as HME-rule (H means "heuristic counterpart"), here the regularization parameter $\alpha = \alpha_{HME}$ is chosen as the global minimizer of the function

$$\psi_{HME}(\alpha) = \alpha^{-1/2} \frac{\left\| B_\alpha \left(Au_\alpha - f \right) \right\|^2}{\left\| B_\alpha^2 \left(Au_\alpha - f \right) \right\|} .$$

In the following we will find the regularization parameter from the set of parameters

$$\Omega = \left\{ \alpha_j : \alpha_j = q\alpha_{j-1}, \quad j = 1, 2, \ldots, M, \quad 0 < q < 1 \right\}, \tag{5}$$

where α_0, q, α_M are given. In the case if in the discretized problem the minimal eigenvalue λ_{min} of the matrix $A^T A$ is larger than α_M, the heuristic rules above choose parameter α_M, which is generally not a good parameter. The works [21, 22, 25] propose to search the global minimum of the function $\psi_Q(\alpha)$ in the interval $[\max(\alpha_M, \lambda_{min}), \alpha_0]$. We use basically the same approach but consider also local minimizers.

We say that the discretized problem $Au = f$ do not need regularization if

$$e_1(\lambda_{min}) = \min_{\alpha \in \Omega, \alpha \geq \lambda_{min}} e_1(\alpha).$$

If $\lambda_{min} > \alpha_M$ and the discretized problem do not need regularization then α_M is the proper parameter while then it is easy to show the error estimate

$$\|u_{\alpha_M} - u_*\| \leq e_1(\alpha_M) \leq 2 \min_{\alpha \in \Omega} e_1(\alpha).$$

Searching the parameter from the interval $[\max(\alpha_M, \lambda_{min}), \alpha_0]$ means the a priori assumption that the discretized problem needs regularization. Note that if $\lambda_{min} > \alpha_M$, then in general case it is not possible to decide (without additional information about solution or about noise of the data), needs the discretized problem regularization or not. In practice in the case $\lambda_{min} > \alpha_M$ it is meaningful to choose the regularization parameter α_H from the interval $[\lambda_{min}, \alpha_0]$, while then our parameter is not too small. If we have some information about solution or about the noise then this information may help to decide, is α_H or α_M the better final parameter.

Our tests are performed on the well-known set of test problems by Hansen [17]. In all tests we used discretization parameter $n = 100$. Since the performance of rules generally depends on the smoothness p of the exact solution in (1), we complemented the standard solutions u_* of (now discrete) test problems with smoothened solutions $|A|^p u_*, |A| := (A^*A)^{1/2}, p = 2$ (computing the right-hand side as $A(|A|^p u_*)$). After discretization all problems were scaled (normalized) in such a way that the Euclidean norms of the operator and the right-hand side were 1. On the base of exact data f_* we formed the noisy data f, where $\|f - f_*\|$ has values $10^{-1}, 10^{-2}, \ldots, 10^{-6}, f - f_*$ has normal distribution and the components of the noise were uncorrelated. We generated 20 noise vectors and used these vectors in all problems. We search the regularization parameter from the set Ω, where $\alpha_0 = 1, q = 0.95$ and M is chosen so that $\alpha_M \geq 10^{-18} > \alpha_{M+1}$.

Since in model equations the exact solution is known, it is possible to find the regularization parameter α_*, which gives the smallest error in the set Ω. For every rule R the error ratio

$$E = \frac{\|u_{\alpha_R} - u_*\|}{\|u_{\alpha_*} - u_*\|} = \frac{\|u_{\alpha_R} - u_*\|}{\min_{\alpha \in \Omega} \|u_\alpha - u_*\|}$$

describes the performance of the rule R on this particular problem. To compare the rules or to present their properties, the following tables show averages A and maximums M of these error ratios over various parameters of the data set (problems 1–10, smoothness indices p, noise levels δ). We say that the heuristic rule fails if the error ratio $E > 100$. Table 1 contains the results of the previous heuristic rules by problems.

This table shows that the quasi-optimality principle succeeds to choose a proper parameter in almost all problems, except the problem *heat* where this principle fails in 66.7% cases. In contrast to other problems in problem *heat* the maximal ratio $\Lambda = \max_{\lambda_k > \max(\alpha_M, \lambda_n)} \lambda_k / \lambda_{k+1}$ of consecutive eigenvalues $\Lambda = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ of the matrix $A^T A$ in the interval $[\max(\alpha_M, \lambda_n), 1]$ is much larger than in other problems. It means that location of the eigenvalues in the interval $[\max(\alpha_M, \lambda_n), 1]$ is sparse.

Table 1 Averages of error ratios E and failure % (in parenthesis) for heuristic rules, $p = 0$

Problem	Λ	Quasiopt.	HR	HME	Reginska
Baart	1666	1.54	2.58	2.52	1.32
Deriv2	16	1.08	2.07	1.72	35.19 (3.3)
Foxgood	210	1.57	8.36	7.71	36.94 (10.8)
Gravity	4	1.13	2.66	2.32	20.49 (0.8)
Heat	$4 * 10^{29}$	> 100 (66.7)	1.64	1.48	23.40 (4.2)
Ilaplace	16	1.24	1.94	1.81	1.66
Phillips	9	1.09	2.27	1.91	> 100 (44.2)
Shaw	290	1.43	2.34	2.23	1.80
Spikes	1529	1.01	1.03	1.03	1.01
Wing	9219	1.40	1.51	1.51	1.18

The rules of Hanke-Raus and HME did not fail in test problems, but the error of the approximate solution is in most problems approximately two times larger than for parameter chosen by the quasi-optimality principle. The problem in these rules is that they choose too large parameter comparing with the optimal parameter. Reginska's rule may fail in many problems but it has the advantage that it works better than other rules if the noise level is large. The Reginska's rule has average of error ratios of all problems $E = 1.46$ and $E = 3.23$ in cases $\|f - f_*\| = 10^{-1}$ and $\|f - f_*\| = 10^{-2}$ respectively, the Hanke-Raus rule has corresponding averages $E = 3.41$ and $E = 3.50$.

By implementing of all these rules the problem is that without additional information in general case it is difficult to decide, is the obtained parameter good or not. In the following we propose a methodology enabling in many cases to assert that obtained parameter is pseudooptimal.

3 Local Minimum Points of the Function $\psi_Q(\alpha)$

In the following we investigate the function $\psi_Q(\alpha)$ in (4) and show that at least one local minimizer of this function is the pseudooptimal parameter. We need some preliminary results.

Lemma 1 *The function $\psi_Q(\alpha)$ has the estimate (see (2) for notation $e_1(\alpha)$)*

$$\psi_Q(\alpha) \leq e_1(\alpha). \tag{6}$$

Proof The following equalities hold:

$$Au_\alpha - f = A\left(\alpha I + A^*A\right)^{-1}A^*f - f = -\alpha\left(\alpha I + AA^*\right)^{-1}f,$$

$$-\alpha^{-1}A^*B_\alpha^2\left(Au_\alpha - f\right) = \alpha A^*\left(\alpha I + AA^*\right)^{-2}f = \alpha\left(\alpha I + A^*A\right)^{-2}A^*f = \tag{7}$$
$$= \alpha A^*A\left(\alpha I + A^*A\right)^{-2}u_* + \alpha\left(\alpha I + A^*A\right)^{-2}A^*(f - f_*).$$

Now the inequality (6) follows from (4) and the inequalities

$$\alpha\left\|A^*A\left(\alpha I + A^*A\right)^{-2}u_*\right\| \le \alpha\left\|\left(\alpha I + A^*A\right)^{-1}u_*\right\| = \left\|u_\alpha^+ - u_*\right\|.$$

$$\alpha\left\|\left(\alpha I + A^*A\right)^{-2}A^*(f - f_*)\right\| \le \left\|\left(\alpha I + A^*A\right)^{-1}A^*(f - f_*)\right\| = \left\|u_\alpha - u_\alpha^+\right\|. \qquad \square$$

Remark 1 Note that $\lim_{\alpha\to\infty}\psi_Q(\alpha) = 0$, but $\lim_{\alpha\to\infty}e_1(\alpha) = \|u_*\|$. Therefore in the case of too large α_0 this α_0 may be global (or local) minimizer of the function $\psi_Q(\alpha)$. We recommend to take $\alpha_0 = c\,\|A^*A\|\,, c \le 1$ or to minimize the function $\tilde{\psi}_Q(\alpha) := (1 + \alpha/\,\|A^*A\|)\psi_Q(\alpha)$ instead of $\psi_Q(\alpha)$. Due to limit $\lim_{\alpha\to 0}(1 + \alpha/\,\|A^*A\|) = 1$ the function $\tilde{\psi}_Q(\alpha)$ approximately satisfies (6).

Lemma 2 *Denote* $\psi_{QD}(\alpha) = (1-q)^{-1}\left\|u_\alpha - u_{q\alpha}\right\|$. *Then it holds*

$$\psi_Q(\alpha) \le \psi_{QD}(\alpha) \le q^{-1}\psi_Q(q\alpha).$$

Proof We use the equalities (7) and

$$u_\alpha - u_{q\alpha} = \left(\alpha I + A^*A\right)^{-1}A^*f - \left(q\alpha I + A^*A\right)^{-1}A^*f =$$
$$= (q-1)\,\alpha\left(\alpha I + A^*A\right)^{-1}\left(q\alpha I + A^*A\right)^{-1}A^*f.$$

The following inequalities prove the lemma:

$$\psi_Q(\alpha) = \alpha\left\|\left(\alpha I + A^*A\right)^{-2}A^*f\right\| \le \alpha\left\|\left(\alpha I + A^*A\right)^{-1}\left(q\alpha I + A^*A\right)^{-1}A^*f\right\| =$$
$$= \psi_{QD}(\alpha) \le \alpha\left\|\left(q\alpha I + A^*A\right)^{-2}A^*f\right\| = q^{-1}\psi_Q(q\alpha). \qquad \square$$

In the following we define the local minimum points of the function $\psi_Q(\alpha)$ on the set Ω (see (5)).

We say that the parameter $\alpha_k, 0 \le k \le M-1$ is the local minimum point of the sequence $\psi_Q(\alpha_k)$, if $\psi_Q(\alpha_k) < \psi_Q(\alpha_{k+1})$ and in case $k > 0$ there exists index $j \ge 1$

such, that $\psi_Q(\alpha_k) = \psi_Q(\alpha_{k-1}) = \ldots = \psi_Q(\alpha_{k-j+1}) < \psi_Q(\alpha_{k-j})$. The parameter α_M is the local minimum point if there exists index $j \geq 1$ so, that

$$\psi_Q(\alpha_M) = \psi_Q(\alpha_{M-1}) = \ldots = \psi_Q(\alpha_{M-j+1}) < \psi_Q(\alpha_{M-j}).$$

Let the number of the local minimum points be K and denote

$$L_{min} = \left\{\alpha_{min}^{(k)} : \alpha_{min}^{(1)} > \alpha_{min}^{(2)} > \ldots > \alpha_{min}^{(K)}\right\}.$$

The parameter $\alpha_k, 0 < k < M$ is the local maximum point of the sequence $\psi_Q(\alpha_k)$ if $\psi_Q(\alpha_k) > \psi_Q(\alpha_{k+1})$ and there exists index $j \geq 1$ so, that

$$\psi_Q(\alpha_k) = \psi_Q(\alpha_{k-1}) = \ldots = \psi_Q(\alpha_{k-j+1}) > \psi_Q(\alpha_{k-j}).$$

We denote by $\alpha_{max}^{(k)}$ the local maximum point between the local minimum points $\alpha_{min}^{(k+1)}$ and $\alpha_{min}^{(k)}$, $1 \leq k \leq K-1$. Denote $\alpha_{max}^{(0)} = \alpha_0, \alpha_{max}^{(K)} = \alpha_M$. Then by the construction

$$\alpha_{max}^{(0)} \geq \alpha_{min}^{(1)} > \alpha_{max}^{(1)} > \ldots > \alpha_{max}^{(K-1)} > \alpha_{min}^{(K)} \geq \alpha_{max}^{(K)}.$$

Theorem 1 *The following estimates hold for the local minimum points of the function $\psi_Q(\alpha)$:*

1.

$$\min_{\alpha \in L_{min}} \|u_\alpha - u_*\| \leq q^{-1} C \min_{\alpha_M \leq \alpha \leq \alpha_0} e_1(\alpha), \tag{8}$$

where

$$C := 1 + \max_{1 \leq k \leq K} \max_{\alpha_j \in \Omega, \alpha_{max}^{(k)} \leq \alpha_j \leq \alpha_{max}^{(k-1)}} T\left(\alpha_{min}^{(k)}, \alpha_j\right) \leq 1 + c_q \ln\left(\frac{\alpha_0}{\alpha_M}\right),$$

$$T(\alpha, \beta) := \frac{\left\|u_\alpha - u_\beta\right\|}{\psi_Q(\beta)}, \qquad c_q := \left(q^{-1} - 1\right) / \ln q^{-1} \to 1 \text{ } if q \to 1.$$

2. Let $u_* = |A|^p v$, $\|v\| \leq \rho, p > 0$ and $\alpha_0 = 1$. If $\delta_0 := \sqrt{\alpha_M} \leq \|f - f_*\|$, then

$$\min_{\alpha \in L_{min}} \|u_\alpha - u_*\| \leq c_p Q^{\frac{1}{p+1}} \max\{\ln \frac{\|f - f_*\|}{\delta_0}, |\ln \|f - f_*\||\} \|f - f_*\|^{\frac{p}{p+1}}, 0 < p \leq 2. \tag{9}$$

Proof For arbitrary parameters $\alpha \geq 0, \quad \beta \geq 0$ the inequalities

$$\|u_\alpha - u_*\| \leq \left\|u_\alpha - u_\beta\right\| + \left\|u_\beta - u_*\right\| \leq T(\alpha, \beta)\psi_Q(\beta) + e_1(\beta)$$

and (6) lead to the estimate

$$\|u_\alpha - u_*\| \leq (1 + T(\alpha, \beta))\, e_1(\beta). \tag{10}$$

It is easy to see that

$$\min_{\alpha_j \in \Omega} e_1(\alpha_j) \leq q^{-1} \min_{\alpha_M \leq \alpha \leq \alpha_0} e_1(\alpha), \tag{11}$$

while in case $q\alpha \leq \alpha' \leq \alpha$ we have $e_1(\alpha') \leq q^{-1} e_1(\alpha)$.

Let $\alpha_{j*} = \alpha_0 q^{j*}$ be the global minimum point of the function $e_1(\alpha)$ on the set of the parameters Ω. Then $\alpha_{j*} \in [\alpha_{max}^{(k)}, \alpha_{max}^{(k-1)}]$ for some k, $1 \leq k \leq K$. Denote $u_j = u_{\alpha_j}$ and $u_{kmin} = u_{\alpha_{min}^{(k)}}$. Then using (10) we can estimate

$$\|u_{kmin} - u_*\| \leq \left(1 + T(\alpha_{min}^{(k)}, \alpha_{j*})\right) e_1(\alpha_{j*}) \leq$$

$$\left(1 + \max_{\alpha_{max}^{(k)} \leq \alpha_j \leq \alpha_{max}^{(k-1)}} T(\alpha_{min}^{(k)}, \alpha_j)\right) \min_{\alpha_j \in \Omega} e_1(\alpha_j).$$

Since we do not know to which interval $[\alpha_{max}^{(k)}, \alpha_{max}^{(k-1)}]$ the parameter α_{j*} belongs, we take maximum of T over all intervals, $1 \leq k \leq K$. Using also (11) we obtain the estimate (8).

Now we show that $C \leq 1 + c_q \ln\left(\frac{\alpha_0}{\alpha_M}\right)$. At first we estimate $T(\alpha_{min}^{(k)}, \alpha_j)$ in the case if $\alpha_{min}^{(k)} \leq \alpha_j \leq \alpha_{max}^{(k-1)}$. Then Lemma 2 enables to estimate

$$\left\|u_{kmin} - u_j\right\| \leq \Sigma_{j \leq i \leq kmin-1} \|u_i - u_{i+1}\| \leq q^{-1}(1-q)\, \Sigma_{j \leq i \leq kmin-1} \psi_Q(\alpha_{i+1})$$

and

$$T(\alpha_{min}^{(k)}, \alpha_j) = \frac{\left\|u_{kmin} - u_j\right\|}{\psi_Q(\alpha_j)} \leq q^{-1}(1-q)\, \Sigma_{j \leq i \leq kmin-1} \frac{\psi_Q(\alpha_{i+1})}{\psi_Q(\alpha_j)} \leq$$

$$(q^{-1} - 1)(kmin - j) \leq (q^{-1} - 1)M = \frac{(q^{-1} - 1)}{\ln q^{-1}} \ln \frac{\alpha_0}{\alpha_M} = c_q \ln \frac{\alpha_0}{\alpha_M}.$$

If $\alpha_{max}^{(k)} \leq \alpha_j \leq \alpha_{min}^{(k)}$, then analogous estimation of $T(\alpha_{min}^{(k)}, \alpha_j)$ gives the same result.

For source-like solution $u_0 - u_* = |A|^p\, v$, $\|v\| \leq \rho$, $p > 0$ the error estimate

$$\min_{\alpha_M \leq \alpha \leq \alpha_0} e_1(\alpha) \leq c_p \rho^{1/(p+1)} \|f - f_*\|^{p/(p+1)}, 0 < p \leq 2$$

is well-known (see [6, 36]), the relations

$$\ln \frac{\alpha_0}{\alpha_M} = \ln \delta_0^{-2} \leq 4 \max\left\{\ln \frac{\|f - f_*\|}{\delta_0}, |\ln \|f - f_*\||\right\}$$

lead to the estimate (9). □

Table 2 Results for the set L_{min}, $p = 0$

	ME	MEe	DP	Best of L_{min}		$\lvert L_{min}\rvert$		Apost. C	
Problem	Aver E	Aver E	Aver E	Aver E	Max E	Aver	Max	Aver	Max
Baart	1.43	1.32	1.37	1.23	2.51	6.91	8	3.19	3.72
Deriv2	1.09	1.08	1.28	1.08	1.34	2.00	2	3.54	4.49
Foxgood	1.98	1.42	1.34	1.47	6.19	3.63	6	3.72	4.16
Gravity	1.40	1.13	1.16	1.13	1.83	1.64	3	3.71	4.15
Heat	1.19	1.03	1.05	1.12	2.36	3.19	5	3.92	4.50
Ilaplace	1.33	1.21	1.26	1.20	2.56	2.64	5	4.84	6.60
Phillips	1.27	1.02	1.02	1.06	1.72	2.14	3	3.99	4.66
Shaw	1.37	1.24	1.28	1.19	2.15	4.68	7	3.48	4.43
Spikes	1.01	1.00	1.01	1.00	1.02	8.83	10	3.27	3.70
Wing	1.16	1.13	1.15	1.09	1.38	5.20	6	3.07	3.72
Total	1.32	1.16	1.19	1.16	6.19	4.09	10	3.67	6.60

The results of numerical experiments for local minimizers $\alpha \in L_{min}$ of the function $\psi_Q(\alpha)$ are given in Table 2. For comparison the results of δ-rules with $\delta = \|f - f_*\|$ are added to the columns 2–4. Columns 5 and 6 contain respectively the averages and maximums of error ratios E for the best local minimizer $\alpha \in L_{min}$. The results show that the Tikhonov approximation with the best local minimizer $\alpha \in L_{min}$ is even more accurate than with the best δ-rule parameter α_{MEe}. Columns 7 and 8 contain the averages and maximums of cardinalities $|L_{min}|$ of sets L_{min} (number of elements of these sets). Note that number of local minimizers depends on parameter q (for smaller q the number of local minimizers is smaller) and on length of minimization interval determined by the parameter α_M. The number of local minimizers is smaller also for larger noise size. Columns 9 and 10 contain the averages and maximums of values of constant C in the a posteriori error estimate (8). The value of C and error estimate (8) allow to assert, that in test problems [17] the choice of α as the best local minimizer in L_{min} guarantees that error of the Tikhonov approximation has the same order as $\min_{\alpha_M \le \alpha \le \alpha_0} e_1(\alpha)$. Note that average and maximum of error ratio $E1 = \|u_{\alpha_R} - u_*\| / \min_{\alpha \in \Omega} e_1(\alpha)$ for the best local minimizer α_R over all problems were 0.84 and 1.39 (for the MEe-rule corresponding error ratios were 0.85 and 1.69).

4 Restricted Set of the Local Minimizers of the Function $\psi_Q(\alpha)$

We will restrict the set L_{min} using two phases. In the first phase we remove from L_{min} local minimizers in interval, where the function $\|B_\alpha (Au_\alpha - f)\|$ decreases only a little bit. On the second phase we remove from set obtained on the first phase these local minimizers for which the function $\psi_Q(\alpha)$ for decreasing α-values has only small growth before the next decrease.

1. Denote $\delta_M := \|B_{\alpha_M}(Au_{\alpha_M} - f)\|$ and by $\alpha = \alpha_{MD}$ the parameter for which $\|B_\alpha(Au_\alpha - f)\| = b\delta_M, \quad b > 1$. Denote $\alpha_{MDQ} := \min\left(\alpha_{MD}, \alpha_Q\right)$, where $\alpha_Q \in L_{min}$ is the global minimizer of the function $\psi_Q(\alpha)$ on the set Ω. Let $\alpha_{max}^{(k_0)} \le \alpha_{MDQ} < \alpha_{max}^{(k_0-1)}$ for some $k_0, 1 \le k_0 \le K$. Then the set of local minimizers what we obtain on the first phase of restriction, has the form $L_{min}^0 = \left\{\alpha_{min}^{(k)} : 1 \le k \le k_0\right\}$. In the case $\alpha_{max}^{(k_0)} \le \alpha_{MDQ} \le \alpha_{min}^{(k_0)}$ we change denotation to $\alpha_{max}^{(k_0)} := \alpha_{min}^{(k_0)}$.
2. We remove from the set L_{min}^0 these local minimizers $\alpha_{min}^{(k)}$ and following maximizers $\alpha_{max}^{(k)}$, which satisfy the following conditions:

$$\alpha_{min}^{(k)} \ne \alpha_{max}^{(k)}; \qquad \frac{\psi_Q(\alpha_{max}^{(k)})}{\psi_Q(\alpha_{min}^{(k)})} \le c_0; \qquad \frac{\psi_Q(\alpha_{min}^{(k)})}{\min_{j\le k}\psi_Q(\alpha_{min}^{(j)})} \le c_0,$$

where $c_0 > 1$ is some constant. We denote by

$$L_{min}^* := \left\{\overline{\alpha}_{min}^{(k)} : \overline{\alpha}_{min}^{(1)} > \overline{\alpha}_{min}^{(2)} > \ldots > \overline{\alpha}_{min}^{(k_*)}\right\}$$

the set of minimizers remained in L_{min}^0 and denote the remained maximizers by $\overline{\alpha}_{max}^{(k)} : \overline{\alpha}_{max}^{(0)} > \overline{\alpha}_{min}^{(1)} > \ldots > \overline{\alpha}_{max}^{(k_*)}$. According to this algorithm the following inequalities hold:

$$\overline{\alpha}_{max}^{(0)} \ge \overline{\alpha}_{min}^{(1)} > \overline{\alpha}_{max}^{(1)} > \ldots > \overline{\alpha}_{max}^{(k_*-1)} > \overline{\alpha}_{min}^{(k_*)} \ge \overline{\alpha}_{max}^{(k_*)}.$$

Note that if α_M is the global minimizer of the function $\psi_Q(\alpha)$ then $\alpha_M \in L_{min}^*$. But in case $\alpha_{MD} < \alpha_Q$ the global minimizer of the function $\psi_Q(\alpha)$ may not belong to the set L_{min}^*. For the restricted set of local minimizers the following theorem hold.

Theorem 2 *The following estimates hold for the local minimum points of the set* L_{min}^**:*

1.

$$\min_{\alpha\in L_{min}^*} \|u_\alpha - u_*\| \le \max\left\{q^{-1}C_1 \min_{\alpha_M\le\alpha\le\alpha_0} e_1(\alpha), C_2(b)\min_{\alpha_M\le\alpha\le\alpha_0} e_2(\alpha,\delta_*)\right\}, \tag{12}$$

where

$$C_1 := 1 + \max_{1\le k\le k_*} \max_{\alpha_j\in\Omega, \overline{\alpha}_{max}^{(k)}\le\alpha_j\le\overline{\alpha}_{max}^{(k-1)}} T\left(\overline{\alpha}_{min}^{(k)}, \alpha_j\right) \le 1 + c_0c_q \ln\left(\frac{\alpha_0}{\overline{\alpha}_{max}^{(k_*)}}\right) \tag{13}$$

and $\delta_* = \max(\delta_M, \|f - f_*\|)$, $C_2(b) = b + 2$.

2. Let $u_* = |A|^p\, v$, $\|v\| \le \rho$, $p > 0$, $\alpha_0 = 1$. *If* $\delta_0 := \sqrt{\alpha_M} \le \|f - f_*\|$, *then*

$$\min_{\alpha \in L^*_{min}} \|u_\alpha - u_*\| \le c_0 c_p \ln \frac{\|f - f_*\|}{\delta_0} \rho^{\frac{1}{p+1}} \left|\ln \|f - f_*\|\right| \|f - f_*\|^{\frac{p}{p+1}}, 0 < p \le 2. \tag{14}$$

Proof Due to the inequality $\delta_* \ge \|f - f_*\|$ the global minimizer of the function $e_2(\alpha, \delta_*)$ is greater or equal to the global minimizer of the function $e_1(\alpha)$. Denote $\overline{\alpha} := \overline{\alpha}^{(k_*)}_{min}$, let α_* be the global minimizer of the function $e_2(\alpha, \delta_*)$ and α_j* be the global minimizer of the function $e_1(\alpha)$ on the set Ω. We consider separately the cases a) $\alpha_j* \ge \overline{\alpha}$, b) $\alpha_j* \le \overline{\alpha} \le \alpha_*$, c) $\alpha_* \le \overline{\alpha}$.

In the case a) we get the estimate

$$\min_{\alpha \in L^*_{min}} \|u_\alpha - u_*\| \le q^{-1} C_1 \min_{\alpha_M \le \alpha \le \alpha_0} e_1(\alpha) \tag{15}$$

analogically to the proof of Theorem 1, but use for the estimation of $T(\alpha^{(k)}_{min}, \alpha_j)$ the inequality $\Sigma_{j \le i \le kmin-1} \frac{\psi_Q(\alpha_{i+1})}{\psi_Q(\alpha_j)} \le c_0 M$.

In the case b) we estimate

$$\|u_{\overline{\alpha}} - u_*\| \le \left\|u^+_{\alpha_*} - u_*\right\| + 0.5\alpha_j*^{-1/2} \|f - f_*\| \le \min_{\alpha \in \Omega} e_1(\alpha) + \min_{\alpha} e_2(\alpha, \delta_*). \tag{16}$$

In the case c) we have $\overline{\alpha} \le \alpha_{MD}$ and therefore also $\|B_{\overline{\alpha}}(Au_{\overline{\alpha}} - f)\| \le b\delta_M \le b\delta_*$. Now we can prove analogically to the proof of the weak quasioptimality of the modified discrepancy principle [30] that under assumption $\alpha_* \le \overline{\alpha}$ the error estimate

$$\|u_{\overline{\alpha}} - u_*\| \le C_2(b) \min_{\alpha_M \le \alpha \le \alpha_0} e_2(\alpha, \delta_*) \tag{17}$$

holds. Now the assertion 1 of Theorem 2 follows from the inequalities (15)–(17). The proof of assertion 2 is analogical to the proof of Theorem 1. □

We recommend to choose the constant b from the interval [1.5; 2] and coefficient c_0 from the interval [1.5; 3]. In all following numerical examples $b = c_0 = 2$. The numerical experiments show that the set L^*_{min} contains in many test problems only one local minimizer and this is a good regularization parameter. In Table 3 for the test problems [17] the results are given for the set L^*_{min}. The columns 2–7 contain the averages and maximums of the error ratio E for the best parameter from the set L^*_{min}, the average and maximum of numbers $|L^*_{min}|$ of elements of L^*_{min} and averages and maximums of the constants C_1 in the error estimate. The last column of the table contains % of cases, where the set L^*_{min} contained only one element or two elements one of which was α_M. Tables 2 and 3 show that for the best parameter from the set L^*_{min} the error ratio E is smaller than for parameter from the ME-rule. Table 3 shows

Table 3 Results about the set L^*_{min}, $p = 0$

	Best of L^*_{min}		$\lvert L^*_{min}\rvert$		Apost. C_1		$\lvert L^*_{min}\rvert = 1$
Problem	Aver E	Max E	Aver	Max	Aver	Max	%
Baart	1.40	2.91	1.41	3	6.38	7.93	60.8
Deriv2	1.08	1.34	2.00	2	3.54	4.49	100
Foxgood	1.57	6.69	1.00	1	4.39	4.92	100
Gravity	1.14	2.15	1.00	1	3.02	3.95	100
Heat	1.12	2.36	2.05	3	5.08	5.38	0
Ilaplace	1.23	2.56	1.00	1	4.68	6.68	100
Phillips	1.06	1.72	2.10	3	3.97	4.66	90.0
Shaw	1.39	3.11	1.16	2	5.89	8.06	84.2
Spikes	1.01	1.03	1.64	3	10.07	11.82	55.0
Wing	1.30	1.84	2.18	4	3.03	6.63	1.7
Total	1.23	6.69	1.55	4	5.01	11.82	69.2

also that in test problems *foxgood, gravity* and *ilaplace* the set L^*_{min} contains only one element and this a good parameter. Due to small values of C_1 the chosen parameter is pseudooptimal. Note that average and maximum of the error ratio $E1$ for the best local minimizer α_R from L^*_{min} over all problems were 0.88 and 1.61 respectively.

5 Choice of the Regularization Parameter from the Set L^*_{min}

Now we give algorithm for choice of the regularization parameter from the set L^*_{min}.

1. If the set L^*_{min} contains only one parameter, we take this for the regularization parameter. On the base of Theorem 2 we know (we can compute also the a posteriori coefficient C_1), that this parameter is reliable.
2. If the set L^*_{min} contains two parameters one of which is α_M, we take for the regularization parameter another parameter $\alpha \neq \alpha_M$. This parameter is good under the assumption that this problem needs regularization.
3. If the set L^*_{min} contains after possible elimination of α_M more than one parameter, we may use for parameter choice the following algorithms.

a) Let α_Q, α_{HR} be global minimizers of the functions $\psi_Q(\alpha)$, $\psi_{HR}(\alpha)$ respectively on the interval $[\max(\alpha_M, \lambda_{min}), \alpha_0]$. Let $\alpha_{Q1} := \max\big(\alpha_Q, \alpha_{HR}\big)$. Choose from the set L^*_{min} the largest parameter α, which is smaller or equal to α_{Q1}.

b) Let α_{RE} be the global minimizer of the function $\psi_{RE}(\alpha)$ on the interval $[\max(\alpha_M, \lambda_{min}), \alpha_0]$. Let α_{Q2} be the global minimizer of the function $\psi_Q(\alpha)$ on the interval $[\alpha_{RE}, \alpha_0]$. Choose from the set L^*_{min} the largest parameter α, which is smaller or equal to α_{Q2}.

c) For the parameters from L^*_{min} we compute value $R(\alpha) = \frac{\psi_{HR}(\alpha)}{\|u_\alpha\|}$ which we consider as the rough estimate for the relative error $\frac{\|u_\alpha - u_*\|}{\|u_*\|}$ under assumption that parameter α is near to the optimal parameter. We choose for the regularization parameter the smallest parameter α_* from the set L^*_{min}, which satisfies the condition $R(\alpha_*) \le C^* \min_{\alpha \in L^*_{min}, \alpha > \alpha_*} R(\alpha)$. We recommend to choose the constant C^* from the interval $5 \le C^* \le 10$. In the numerical experiments we used $C^* = 5$.

Note that these algorithms are motivated by experience that global minimizers of functions $\psi_{HR}(\alpha)$, $R(\alpha_*)$ are typically too large parameters. Therefore we choose smaller parameter under condition, that in case, if the optimal local minimizer is larger than chosen parameter, the chosen parameter is still pseudooptimal. Choice of value of constant C in algorithm c) was suggested by numerical experiments.

The results of the numerical experiments for different algorithms for the parameter choice are given in Table 4. The results for all three algorithms are very similar and the average of the error ratio is even smaller than for α from the ME-rule. In the case if the set L^*_{min} contained more than three parameters, in 68.1% of cases all three algorithms gave the same parameter and in 92.7% of cases the parameters from algorithms b) and c) coincided. We changed also the parameters $b \in [1.5; 2]$ and $c_0 \in [1.5; 3]$, but the overall average of the ratio E changed less than 2%.

The proposed algorithms for parameter choice are complicated (formation of the set L^*_{min}) but they enable to estimate also the reliability of the chosen parameter and propose alternative parameters if the set L^*_{min} contains several local minimizers. If some information about solution or noise is available, it may help to find from the set L^*_{min} better parameter than algorithms a)–c) find. If the purpose is only parameter choice, simpler rules below may be used (parameters α_{Q1} and α_{Q2} are defined in algorithm a), b)).

Table 4 Averages and maximums of error ratios E in case of different heuristic algorithms, $p = 0$

	Algorithm a)		Algorithm b)		Algorithm c)	
Problem	Aver E	Max E	Aver E	Max E	Aver E	Max E
Baart	1.83	3.63	1.61	2.91	1.61	2.91
Deriv2	1.08	1.34	1.08	1.34	1.08	1.34
Foxgood	1.57	6.69	1.57	6.69	1.57	6.69
Gravity	1.14	2.15	1.14	2.15	1.14	2.15
Heat	1.12	2.36	1.12	2.36	1.12	2.36
Ilaplace	1.23	2.56	1.23	2.56	1.23	2.56
Phillips	1.06	1.72	1.06	1.72	1.06	1.72
Shaw	1.48	3.64	1.45	3.64	1.45	3.64
Spikes	1.01	1.03	1.01	1.03	1.01	1.03
Wing	1.50	1.86	1.38	2.04	1.32	1.84
Total	1.30	6.69	1.26	6.69	1.26	6.69

1. We choose for the regularization parameter the smallest local minimizer $\alpha_{min}^{(k_*)}$ of the function $\psi_Q(\alpha)$ which satisfies the following conditions:

$$\frac{\psi_Q(\alpha_{max}^{(k)})}{\psi_Q(\alpha_{min}^{(k)})} \leq c_0, \qquad k = k_0, k_0+1, \ldots, k_*-1; \tag{18}$$

$$\frac{\psi_Q(\alpha_{min}^{(k)})}{\min_{j\leq k} \psi_Q(\alpha_{min}^{(j)})} \leq c_0, \qquad k = k_0, k_0+1, \ldots, k_*, \tag{19}$$

   where k_0 is the index for which $\alpha_{min}^{(k_0)} \leq \alpha_{Q1} \leq \alpha_{max}^{(k_0-1)}$.
2. We choose for the regularization parameter the smallest local minimizer $\alpha_{min}^{(k_*)}$ of the function $\psi_Q(\alpha)$ satisfying conditions (18), (19) where k_0 is index for which $\alpha_{min}^{(k_0)} \leq \alpha_{Q2} \leq \alpha_{max}^{(k_0-1)}$.

   These rules give in test problems [17] the same results as the algorithms a) and b) respectively.

Table 5 gives results of the numerical experiments in the case of smooth solution, $p = 2$. The table shows that in case of smooth solution the number of local minimizers in L_{min} and number of elements L_{min}^* are smaller than in case $p = 0$. If the set L_{min}^* contains several elements, then the algorithms a) and c) gave the same parameter, which was always the best parameter from L_{min}^* with smallest error. In case of algorithm b) the overall average of the ratio E was 1.25. In all problems except the problem *wing* the heuristic rule gave parameter where the average of error was smaller than by parameter from the ME-rule, and only 10% larger than by parameter from the MEe-rule (both ME-rule and the MEe rule used the exact noise level).

Table 5 Results of the numerical experiments, $p = 2$

	ME	MEe	Best of L_{min}	$\lvert L_{min}\rvert$	Best of L_{min}^*	$\lvert L_{min}^*\rvert$	$\lvert L_{min}^*\rvert = 1$
Problem	Aver E	Aver E	Aver E	Aver	Aver E	Aver	%
Baart	1.86	1.19	1.18	4.74	1.41	1.02	98.3
Deriv2	1.10	1.19	1.03	2.00	1.03	2.00	100
Foxgood	1.56	1.13	1.14	2.08	1.20	1.00	100
Gravity	1.33	1.05	1.09	1.72	1.11	1.00	100
Heat	1.13	1.12	1.05	2.10	1.05	2.10	0
Ilaplace	1.47	1.06	1.11	2.73	1.11	1.00	100
Phillips	1.26	1.06	1.04	2.10	1.04	2.10	90
Shaw	1.37	1.06	1.11	3.72	1.22	1.01	99.2
Spikes	1.85	1.12	1.19	4.78	1.31	1.00	100
Wing	1.67	1.14	1.22	4.53	1.73	1.01	99.2
Total	1.46	1.11	1.12	3.05	1.22	1.32	88.7

We finish the paper with the following conclusion. For the heuristic choice of the regularization parameter we recommend to choose the parameter from the set of local minimizers of the function $\psi_Q(\alpha)$. Proposed algorithm enables to restrict this set and in many problems the restricted set contains only one element, this parameter is the pseudooptimal parameter.

Acknowledgements The authors are supported by institutional research funding IUT20-57 of the Estonian Ministry of Education and Research.

References

1. A.B. Bakushinskii, Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. Comput. Math. Math. Phys. **24**(4), 181–182 (1984)
2. F. Bauer, S. Kindermann, The quasi-optimality criterion for classical inverse problems. Inverse Prob. **24**(3), 035002 (2008)
3. F. Bauer, S. Kindermann, Recent results on the quasi-optimality principle. J. Inverse Ill-Posed Prob. **17**(1), 5–18 (2009)
4. F. Bauer, M.A. Lukas, Comparing parameter choice methods for regularization of ill-posed problems. Math. Comput. Simul. **81**(9), 1795–1841 (2011)
5. F. Bauer, M. Reiss, Regularization independent of the noise level: an analysis of the quasi-optimality. Inverse Prob. **24**, 055009 (2008)
6. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*. Mathematics and Its Applications, vol. 375 (Kluwer, Dordrecht, 1996)
7. H. Gfrerer, An a posteriori parameter choice for ordinary and iterated Tikhonov regularization of ill-posed problems leading to optimal convergence rates. Math. Comput. **49**(180), 507–522 (1987)
8. G.H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics **21**(2), 215–223 (1979)
9. U. Hämarik, T. Raus, About the balancing principle for choice of the regularization parameter. Numer. Funct. Anal. Optim. **30**(9–10), 951–970 (2009)
10. U. Hämarik, R. Palm, T. Raus, On minimization strategies for choice of the regularization parameter in ill-posed problems. Numer. Funct. Anal. Optim. **30**(9–10), 924–950 (2009)
11. U. Hämarik, R. Palm, T. Raus, Extrapolation of Tikhonov regularization method. Math. Model. Anal. **15**(1), 55–68 (2010)
12. U. Hämarik, R. Palm, T. Raus, Comparison of parameter choices in regularization algorithms in case of different information about noise level. Calcolo **48**(1), 47–59 (2011)
13. U. Hämarik, R. Palm, T. Raus, A family of rules for parameter choice in Tikhonov regularization of ill-posed problems with inexact noise level. J. Comput. Appl. Math. **36**(2), 221–233 (2012)
14. U. Hämarik, U. Kangro, R. Palm, T. Raus, U. Tautenhahn, Monotonicity of error of regularized solution and its use for parameter choice. Inverse Prob. Sci. Eng. **22**(1), 10–30 (2014)
15. M. Hanke, T. Raus, A general heuristic for choosing the regularization parameter in ill-posed problems. SIAM J. Sci. Comput. **17**(4), 956–972 (1996)
16. P.C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev. **34**(4), 561–580 (1992)
17. P.C. Hansen, Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems. Numer. Algorithms **6**(1), 1–35 (1994)
18. M.E. Hochstenbach, L. Reichel, G. Rodriguez, Regularization parameter determination for discrete ill-posed problems. J. Comput. Appl. Math. **273**, 132–149 (2015)

19. B. Jin, D. Lorenz, Heuristic parameter-choice rules for convex variational regularization based on error estimates. SIAM J. Numer. Anal. **48**(3), 1208–1229 (2010)
20. S. Kindermann, Convergence analysis of minimization-based noise level-free parameter choice rules for linear ill-posed problems. Electron. Trans. Numer. Anal. **38**, 233–257 (2011)
21. S. Kindermann, Discretization independent convergence rates for noise level-free parameter choice rules for the regularization of ill-conditioned problems. Electron. Trans. Numer. Anal. **40**, 58–81 (2013)
22. S. Kindermann, A. Neubauer, On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. Inverse Prob. Imag. **2**(2), 291–299 (2008)
23. S. Lu, P. Mathe, Heuristic parameter selection based on functional minimization: optimality and model function approach. Math. Comput. **82**(283), 1609–1630 (2013)
24. V.A. Morozov, On the solution of functional equations by the method of regularization. Soviet Math. Dokl. **7**, 414–417 (1966)
25. A. Neubauer, The convergence of a new heuristic parameter selection criterion for general regularization methods. Inverse Prob. **24**, 055005 (2008)
26. R. Palm, Numerical comparison of regularization algorithms for solving ill-posed problems. PhD thesis, University of Tartu, 2010, http://hdl.handle.net/10062/14623
27. S.V. Pereverzev, E. Schock, On the adaptive selection of the parameter in the regularization of ill-posed problems. SIAM J. Numer. Anal. **43**(5), 2060–2076 (2005)
28. T. Raus, On the discrepancy principle for solution of ill-posed problems with non-selfadjoint operators. Acta et comment. Univ. Tartu. **715**, 12–20 (1985) [In Russian]
29. T. Raus, About regularization parameter choice in case of approximately given error bounds of data. Acta et comment. Univ. Tartu. **937**, 77–89 (1992)
30. T. Raus, U. Hämarik, On the quasioptimal regularization parameter choices for solving ill-posed problems. J. Inverse Ill-Posed Prob. **15**(4), 419–439 (2007)
31. T. Raus, U. Hämarik, New rule for choice of the regularization parameter in (iterated) Tikhonov method. Math. Model. Anal. **14**(2), 187–198 (2009)
32. T. Raus, U. Hämarik, On numerical realization of quasioptimal parameter choices in (iterated) Tikhonov and Lavrentiev regularization. Math. Model. Anal. **14**(1), 99–108 (2009)
33. T. Reginska, A regularization parameter in discrete ill-posed problems. SIAM J. Sci. Comput. **17**(3), 740–749 (1996)
34. U. Tautenhahn, U. Hämarik, The use of monotonicity for choosing the regularization parameter in ill-posed problems. Inverse Prob. **15**(6), 1487–1505 (1999)
35. A.N. Tikhonov, V.B. Glasko, Y. Kriksin, On the question of quasioptimal choice of a regularized approximation. Sov. Math. Dokl. **20**, 1036–1040 (1979)
36. G.M. Vainikko, A.Y. Veretennikov, *Iteration Procedures in Ill- Posed Problems* (Nauka, Moscow, 1986) [In Russian]

Modification of Iterative Tikhonov Regularization Motivated by a Problem of Identification of Laser Beam Quality Parameters

Teresa Regińska and Kazimierz Regiński

Abstract Presented is a new method for finding an approximate minimum of a real function given on a discrete set of points where its values are given with some errors. The applied approach is a certain modification of the iterative Tikhonov regularization. The essence of the presented method is to reduce the initial problem to that of finding an approximation of the function in a class of functions whose minimum can easily be calculated. The presented method is motivated by a problem of identification of laser beam quality parameters, however the scope of its applicability is quite general.

1 Introduction

In this paper we propose a new method for finding an approximate minimum of a function f given on a discrete set of points. The actually available data $f^{\delta}(z)$ for $z \in I := [s_1, s_2]$ will be contaminated with noise for which we here use a deterministic model, i.e.,

$$|f^{\delta}(z) - f(z)| \leq \delta(z). \tag{1}$$

Let $F(v; z)$ denote a real continuous function of several variables $v \in D(F) \subset R^d$, and $z \in I$. Let us consider a set

$$\mathscr{F} = \{f \in C(I) : \exists v \in D(F)\, f(z) = F(v; z) \text{ for } z \in I\}.$$

T. Regińska (✉)
Institute of Mathematics of the Polish Academy of Sciences, Warsaw, Poland
e-mail: reginska@impan.pl

K. Regiński
Institute of Electron Technology, Warsaw, Poland
e-mail: reginski@ite.waw.pl

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_13

We assume that

1. for any $v \in D(F)$ $F(v;\cdot)$ is convex and its minimum is easily computable;
2. $F(v_1;\cdot) = F(v_2;\cdot)$ if and only if $v_1 = v_2$;
3. for the exact f there exists $v^\dagger$: $F(v^\dagger;\cdot) = f(\cdot)$, i.e., $f \in \mathscr{F}$.
4. for any $z \in I$ a noisy $f^\delta(z)$ can be computed or measured.

The case when f is not in $\mathscr{F}$ but there exists $\tilde{f} \in \mathscr{F}$ such that $\|\tilde{f} - f\|_\infty \le \epsilon$ for sufficiently small ϵ can also be considered. Then $\tilde{v}^\dagger$ will denote the solution of $F(\tilde{v}^\dagger;\cdot) = \tilde{f}(\cdot)$.

So, the problem of minimizing f is replaced by the following one:

Problem 1 Find $v^\delta \in D(F)$ such that $F(v^\delta;z) \sim f^\delta(z)$ on a chosen set of points z and $\min_{z\in I} F(v^\delta, z) \sim \min_{z\in I} f(z)$.

The problem above is generated by the problem of determining the axial profile of the laser beam. In the case of narrow beam one can assume that this profile is a hyperbola type [9, 19]. The final aim of the proposed approach is to approximate laser beam quality parameters such as the waist of an axial profile of the beam and its position on the axis of beam by corresponding values of a hyperbola which well approximate the given noisy points of axial profile. The problem for the axial profile of the laser beam is described in Sect. 2.

The problem of finding v^δ could be formulated as a system of nonlinear equations

$$F(v;z_i) = f^\delta(z_i), \; i = 1,\dots,N \tag{2}$$

for a fixed set of points $\{z_i\}$. In [15] we proposed to approximate $v^\dagger$ by v^δ_α given by the Tikhonov regularization method. This method is widely used especially in connection with the output least squares formulation of parameter estimation problems (see [1, 5]). However, a small distance between $f^\delta(z_i)$ and $f(z_i)$ for z_i from the given a priori set of points $\{z_i\}$ does not guarantee a small distance between $\arg\min_{z\in I} F(v^\delta_\alpha, z)$ and $\arg\min_{z\in I} f(z)$.

It should be stressed that it is not clear how to choose a priori an appropriate set of points $\{z_i\}$. Therefore we propose an iteration method in which the set of points changes in every step of iteration; to the fixed set of points $\{z_i^0\}_{i=1}^r$ we add additional one which is defined successively during iterations. So, let

$$Z_n := \{z_1^0, \cdots, z_r^0, z_n\}.$$

We abbreviate the notation by introducing the operator

$$F_n(\cdot) := \left(F(\cdot;z_1^0), \cdots, F(\cdot;z_r^0), F(\cdot;z_n)\right), \qquad F_n : D(F) \subset R^d \to R^{r+1} \tag{3}$$

and the vector

$$f_n^\delta = (f^\delta(z_1^0), \dots, f^\delta(z_r^0), f^\delta(z_n)). \tag{4}$$

The norms and inner products in R^d and R^{r+1} will be denoted by $\|\cdot\|$ and $<\cdot,\cdot>$; they can always be identified from the context in which they appear. Consider the equation

$$F_n(v) = f_n \tag{5}$$

with a noisy right hand side f_n^δ, satisfying

$$\|f_n^\delta - f_n\| \le \delta_n \le \delta. \tag{6}$$

Generally, for the noisy right hand side f_n^δ the solution in the classical sense does not exist. So, as a solution of the problem (5) we chose the concept of so called v^*-minimum norm least-squares solution, i.e. the least squares solution of a minimal distance to a fixed v^* [5]. Available a priori information about the location of least-squares solutions of (5) has to enter into the selection of v^*.

The starting point of our approach is the iterated Tikhonov method [6] for solving $F(v) = f$. This regularization method is defined by

$$v_{n+1}^\delta = \arg\min_{v\in D(F)} \left\{\| F(v) - f^\delta\|^2 + \beta\|v - v_n^\delta\|^2\right\}.$$

Our modification consists in introducing different operators at different steps. We propose in this article the following regularization method:

Let us denote an initial guess for $n = 0$

$$v_0^\delta = v^*. \tag{7}$$

Step $n + 1$:

- $$z_{n+1}^\delta = \arg\min_{z\in I} F(v_n^\delta; z); \tag{8}$$

- $$F_{n+1}(v) = \left(F(\cdot; z_1^0), \cdots, F(\cdot; z_r^0), F(\cdot; z_{n+1})\right); \tag{9}$$

- $$\Phi_{n+1}(v) := \| F_{n+1}(v) - f_{n+1}^\delta\|^2 + \beta\|v - v_n^\delta\|^2; \tag{10}$$

- $$v_{n+1}^\delta = \arg\min_{v\in D(F)} \Phi_{n+1}(v) \text{ for } n \ge 0. \tag{11}$$

Here β is an appropriately chosen number (see (26) below). The number r is chosen a priori and it depends on F. For instance, in the case of a hyperbola (18) we can take $r = 3$ (see Sect. 5). We apply an a-posteriori stopping rule which employs the discrepancy principle, i.e., the iteration is stopped after n_* steps with

$$n_* = \min\{n \in N : n > 0 \text{ and } \| F_n(v_n^\delta) - f_n^\delta \| \leq \tau\delta\}, \tag{12}$$

where $\tau > 1$. The article is outlined as follows. In Sect. 2 we present the underlying mathematical model of laser beam and the problem of identification of laser beam quality parameters. In Sect. 3 disadvantages of the Tikhonov method applied to the problem stated in Introduction are considered. In Sect. 4 we analyze the proposed method, we formulate basic assumptions and derive a convergence result. Section 5 contains remarks concerning applicability of the method to the problem underlaying. Section 6 is devoted to final remarks and conclusions.

2 Problem Formulation for a Laser Beam

The laser beam is an electromagnetic field, i.e. from a mathematical point of view a vector field defined on R^3. Usually we measure only one component of this field, e.g. one component of the electric field u. (For technical details of measurements, see e.g. [11, 17, 18].) A simplified mathematical model for a collimated laser beam leads to the Cauchy problem for the Helmholtz equation on u. This problem is ill-posed. Hence, its numerical treatment requires the application of special regularization methods. For references to the extensive literature on the subject one may refer to [7]. The Cauchy problem for the Helmholtz equation can be considered on different bounded or unbounded domains, but the boundary conditions are always given only on a part of the boundary. It is considered mainly on an infinite strip or on a rectangle or a cuboid. Methods for an infinite strip based on the equivalent operator equation in the frequency space were considered for instance in [7, 14–16, 20, 21]. For a treatment of rectangular or cuboidal domains one may refer to [2, 12, 13, 22].

For formulating a mathematical model we follow the papers [14–16] where the domain is an infinite strip and where the numerical analysis of a spectral type regularization is presented for the reconstruction of the field. Consider the Helmholtz equation

$$\begin{aligned} \Delta u + k^2 u = 0 \text{ for } (x, y, z) \in \Omega = R^2 \times (0, s), \\ u(\cdot, \cdot, z) \in L^2(R^2) \text{ for } z \in (0, s), \end{aligned} \tag{13}$$

where u is a component of the electric field, $\Delta u = u_{xx} + u_{yy} + u_{zz}$ and $k > 0$ is the wave number. We assume that the boundary conditions are given only on

$\Gamma = R^2 \times s$, i.e. on a part of the boundary $\partial\Omega$:

$$\begin{aligned} u &= g, \text{ on } \Gamma, \\ \frac{\partial u}{\partial z} &= h, \text{ on } \Gamma \end{aligned} \tag{14}$$

while on $\Gamma_0 = R^2 \times 0$, a source condition $\|u(\cdot, 0)\|_{L^p} \leq E$ is assumed.

In practical applications, instead of defining the laser beam as a vector field, we often describe it as a geometrical object. In the present paper we consider the axial profile defined by the radii of the beam at the points z denoted by $f(z)$. In general the beam is not axially symmetric, i.e. it has different radii in directions x and y, but for simplicity we restrict further considerations to the x, z plane.

In the literature there are many definitions of the radii of the beam (cf. [9, 19]), but the most popular one uses the second moment of the power of radiation:

$$f(z) = \frac{P_{2,x}(z)}{P(z)} \tag{15}$$

where

$$P_{2,x}(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 u^2(x, y, z)dxdy, \tag{16}$$

and

$$P(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2(x, y, z)dxdy \tag{17}$$

is the total intensity of the beam on the cross-section at the point z.

In [15], the regularized solutions of the Cauchy problem for the Helmholtz equation (13) (14) are employed to approximate the axial profile. In order to find approximate beam quality parameters such as the waist of the axial profile of the beam and its position, the axial profile has been approximated by a hyperbola. It is reasonable, because in the case of a narrow beam the exact f is a hyperbola (see [9, 10, 19]).

Therefore, in the case of axial profiles of laser beams we chose

$$F(v; z) := \sqrt{a^2 + \frac{a^2}{b^2}(z - c)^2}, \tag{18}$$

where $v = (a, b, c)$ and

- c is the center of the hyperbola,
- a denotes the transverse semi-axis,
- b denotes the conjugate semi-axis.

The central equation of the hyperbola is generated by parameters $v = (a, b, c)$

$$\frac{x^2}{a^2} - \frac{(z-c)^2}{b^2} = 1.$$

Now $F_{n+1}(v)$ in (9) transforms $D(F) \subset R^3$ into R^{r+1} where

$$D(F) := [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}] \times [\underline{c}, \overline{c}]. \tag{19}$$

We assume, that we are able to indicate a proper $D(F)$ for the considered model. Moreover, we assume that $[\underline{c}, \overline{c}] \subseteq [s_1, s_2] \subset (0, s)$.

In the paper we will consider two cases:

1. for the exact $f(z)$ there exist the unique vector $v^\dagger = (a^\dagger, b^\dagger, c^\dagger)$ such that

$$f(z) = F(v^\dagger; z) \text{ for } z \in (s_1, s_2). \tag{20}$$

2. there exists $g \in \mathscr{F}$ such that for any $z \in I$, $|f(z) - g(z)| \leq \epsilon$ for sufficiently small ϵ and

$$g(z) = F(v^\dagger; z) \text{ for } z \in (s_1, s_2). \tag{21}$$

The aim of consideration is to find the minimum of axial profile and the point where it is attained. In the case when f is the branch of hyperbola then its minimum equals $a^\dagger$ and it is attained at the point $z = c^\dagger$. In the second case $a^\dagger$ is the minimum of g and can be considered as an approximation of the minimum of f.

The finite set of noisy data $f_n^\delta(z)$ (4) can be obtained in the following way:

- An approximate value of $f^\delta(z)$ is obtained by quadratures applied to numerical computation of a value of the functional (15) acting on the electric field $u(\cdot z)$ (on the beam cross-section at the point z).
- In practice, the available $u(\cdot, z)$ is always corrupted by noise. Instead of the exact function in (15) we have measured $u^\delta(\cdot, z)$ or numerically computed $u^\delta_{\alpha(\delta)}$.
- If on the beam cross-section at a point z there are no measurements of the field, we can reconstruct this field by solving an appropriate problem for the Helmholtz equation:

- if $z_i \in (s_1, s_2)$ is less than the smallest point $\underline{z}$ for which we have measurement data, then we have to deal with an ill-posed Cauchy problem for the Helmholtz equation on $R^2 \times (0, \underline{z})$ and for a field reconstruction we have to use some regularization method.
- if $z_i \in (\underline{z}, s_2)$ than we can formulate a well posed boundary value problem for the Helmholtz equation and obtain approximate field by a stable numerical method (see [14]).

3 Tikhonov Regularization

In this section we consider the system of equations (2) with a fixed set of points $\{z_i\}_{i=1}^N$. In Tikhonov regularization method a solution of (2) is approximated by a solution of the minimization problem

$$v_\alpha^\delta = \arg \min_{v \in D(F)} \left\{ \| F_N(v) - f_N^\delta \|^2 + \alpha \| v - v^* \|^2 \right\}. \tag{22}$$

By the assumption on F, F_N is continuous and compact and $D(F)$ is closed and convex. Thus the minimization problem (22) admits a solution (see [6, Chap. 10.2]). However, since F_N is nonlinear, the solution will not be unique, in general.

The problem of solving (22) is stable in the sense of continuous dependence of the solutions on the data f^δ (see Theorem 10.2 in [6]). The following convergence result follows from [6, Theorem. 10.3]:

Theorem 1 *Let $f \in \mathscr{F}$, the solution $v^\dagger$ of (2) be unique, and $\|f_N^\delta - f_N\| \leq \delta$. If $\alpha(\delta)$ is such that*

$$\alpha(\delta) \to 0 \textit{ and } \frac{\delta^2}{\alpha(\delta)} \to 0 \textit{ as } \delta \to 0,$$

then

$$\lim_{\delta \to 0} v_{\alpha(\delta)}^\delta = v^\dagger.$$

However, this theoretical asymptotic result is not very useful in practical situation when we have to deal with given measurement errors. Moreover a small distance between f_N^δ and f_N does not guarantee a small distance between $\arg\min_{z \in I} F(v_\alpha^\delta, z)$ and $\arg\min_{z \in I} f(z)$.

4 Modification of Iterated Tikhonov Regularization

In this section we assume that $f \in F$ and $F \in C^2(D(F) \times I)$. Since

$$F_n'(v)^* \left(F_n(v) - f_n^\delta \right) + \beta (v - v_{n-1}^\delta)$$

is the gradient of the Tikhonov functional $\Phi_n(v)$ defined by (10), minimization (11) corresponds to the iteration

$$v_n^\delta = v_{n-1}^\delta - \frac{1}{\beta} F_n'(v_n^\delta)^* \left(F_n(v_n^\delta) - f_n^\delta \right). \tag{23}$$

First, let us prove that under some assumptions on the function F, v_n^δ is well defined, i.e., the Tikhonov functional $\Phi_n(v)$ in (10) has unique minimizer in a closed ball $B_\rho(v^*)$ of radius ρ around v^*.

Lemma 1 *Let ρ, M, L, be such that*

1. $\| F_v'(v,z)\| \leq M$ *for* $v \in B_\rho(v^*)$, $z \in I$;
2. $\forall z \in I$, $\| F_v'(v_1,z) - F_v'(v_2,z)\| \leq L\|v_1 - v_2\|$ *for* $v_1, v_2 \in B_\rho(v^*)$.

If $\beta > M^2 + LC$ *where* $C = \sup\{\| F_n(v) - f_n^\delta\| : v \in B_\rho(v^*), \|f_n - f_n^\delta\| \leq \delta\}$, *then* v_n^δ *is uniquely defined.*

Proof Suppose that $v_1,\ v_2 \in B_\rho(v^*)$ are minimizers of Φ_n. Using (23) we get

$$\|v_1 - v_2\|^2 = \frac{1}{\beta} < F_n'(v_2)^* \left(F_n(v_2) - f_n^\delta\right) - F_n'(v_1)^* \left(F_n(v_1) - f_n^\delta\right), v_1 - v_2 >$$

$$= \frac{1}{\beta} < \left(F_n(v_2) - f_n^\delta\right), \left(F_n'(v_2) - F_n'(v_1)\right)(v_1 - v_2) >$$

$$+\frac{1}{\beta} < (F_n(v_2) - F_n(v_1)), F_n'(v_1)(v_1 - v_2) > .$$

Since the estimations 1-2 hold in particular for $z \in Z_n$, we conclude that corresponding estimations hold for F_n and its Fréchet-derivative F_n'. Hence

$$\|v_1 - v_2\|^2 \leq \frac{1}{\beta}(CL + M^2)\|v_1 - v_2\|^2,$$

It means that for β sufficiently large, $\|v_1 - v_2\| = 0$.

Let us observe that (23) is a modification of the nonlinear Landweber iteration [8] in which the operator F_n changes at each step. Iterations with different operators occur in Kaczmarz type methods (see [3, 4]). The convergence analysis presented below follows the lines of the proof of convergence of an iterated Tikhonov-Kaczmarz method established in [4].

Lemma 2 *For* $n \geq 0$

$$\| F_{n+1}(v_{n+1}^\delta) - f_{n+1}^\delta\|^2 \leq \| F_{n+1}(v_n^\delta) - f_{n+1}^\delta\|^2.$$

Proof v_{n+1} is a minimizer of (10), thus

$$\| F_{n+1}(v_{n+1}^\delta) - f_{n+1}^\delta\|^2 \leq \Phi_{n+1}(v_{n+1}^\delta) \leq \Phi_{n+1}(v_n^\delta).$$

Thus, Lemma follows from the equality

$$\Phi_{n+1}(v_n^\delta) = \| F_{n+1}(v_n^\delta) - f_{n+1}^\delta\|^2.$$

Lemma 3 *Let us assume that for the exact f there exists the unique vector $v^\dagger$ such that $f(\cdot) = F(v^\dagger;\cdot)$. If*

$$v^* \in B_{\rho/4}(v^\dagger) \subset D(F), \tag{24}$$

$$v_n^\delta \in B_{\rho/4}(v^\dagger), \tag{25}$$

$$\beta \geq \left(\frac{4\delta_{n+1}}{\rho}\right)^2, \tag{26}$$

and

$$\|f_{n+1}^\delta - f_{n+1}\| \leq \delta_{n+1}$$

then

$$v_{n+1}^\delta \in B_\rho(v^*)$$

and

$$v_{n+1}^\delta \in B_\rho(v^\dagger).$$

Proof From (10) and (11) it follows that

$$\beta\|v_{n+1}^\delta - v_n^\delta\|^2 \leq \Phi_{n+1}(v_{n+1}^\delta) \leq \Phi_{n+1}(v^\dagger) = \|f_{n+1} - f_{n+1}^\delta\|^2 + \beta\|v^\dagger - v_n^\delta\|^2,$$

since $F_{n+1}(v^\dagger) = f_{n+1}$. Thus by (25) and (26)

$$\|v_{n+1}^\delta - v_n^\delta\| \leq \frac{\delta_{n+1}}{\sqrt{\beta}} + \|v^\dagger - v_n^\delta\| \leq \frac{\rho}{2}.$$

It means that for

$$\|v_{n+1}^\delta - v^\dagger\| \leq \|v_{n+1}^\delta - v_n^\delta\| + \|v_n^\delta - v^\dagger\| \leq \frac{\rho}{2} + \frac{\rho}{4}.$$

Moreover, by (24)

$$\|v_{n+1}^\delta - v^*\| \leq \|v_{n+1}^\delta - v^\dagger\| + \|v^\dagger - v^*\| \leq \rho$$

which ends the proof.

The proof of monotonicity of the method i.e., $\|v_{n+1}^\delta - v^\dagger\| \leq \|v_n^\delta - v^\dagger\|$, can be done like the proof of monotony property for an iterated Tikhonov-Kaczmarz method presented in [4, Proposition 1]. In particular, it is based on an auxiliary lemma which for (11) takes the following form

Lemma 4 *Let the assumptions of Lemma 3 be satisfied. If for some $\eta < 1$*

$$| F(u;z) - F(v;z) - F'(v;z)(u-v)| \leq \eta | F(u;z) - F(v;z)| \tag{27}$$

holds for any $u, v \in B_\rho(v^*)$ and for $n > 0$, then

$$\|v^\delta_{n+1} - v^\dagger\|^2 - \|v^\delta_n - v^\dagger\|^2 \leq \frac{2}{\beta} C^\delta_{n+1}\left((\eta-1)C^\delta_{n+1} + (\eta+1)\delta_{n+1}\right), \tag{28}$$

where

$$C^\delta_{n+1} := \| F_{n+1}(v^\delta_{n+1}) - f^\delta_{n+1}\|.$$

Proof The assumption (27) holds for $z \in Z_{n+1}$ and for $u = v^\delta_{n+1}$ and $v = v^\dagger$, since they belong to $B_\rho(v^*)$ according to Lemma 3. Thus

$$\begin{aligned} \| F_{n+1}(v^\delta_{n+1}) - F_{n+1}(v^\dagger) - F'_{n+1}(v^\delta_{n+1})(v^\delta_{n+1} - v^\dagger)\| \leq \\ \eta \| F_{n+1}(v^\delta_{n+1}) - F_{n+1}(v^\dagger)\|. \end{aligned} \tag{29}$$

We have

$$\begin{aligned} &\|v^\delta_{n+1} - v^\dagger\|^2 - \|v_n - v^\dagger\|^2 \leq 2 < v^\delta_{n+1} - v^\dagger, v^\delta_{n+1} - v^\delta_n > \\ &= \frac{2}{\beta} < F_{n+1}(v^\delta_{n+1}) - f^\delta_{n+1}, -F'_{n+1}(v^\delta_{n+1})(v^\delta_{n+1} - v^\dagger) \pm F_{n+1}(v^\dagger) \pm F_{n+1}v^\delta_{n+1}) > \\ &\leq \frac{2}{\beta} C^\delta_{n+1} \eta \| F_{n+1}(v^\delta_{n+1}) - F_{n+1}(v^\dagger) \pm f^\delta_{n+1} \pm f_{n+1}\| \\ &+ \frac{2}{\beta} < F_{n+1}(v^\delta_{n+1}) - f^\delta_{n+1}, F_{n+1}(v^\dagger) - f^\delta_{n+1} - (F_{n+1}(v^\delta_{n+1}) - f^\delta_{n+1}) > \\ &\leq \frac{2}{\beta} C^\delta_{n+1}\left((\eta-1)C^\delta_{n+1} + (\eta+1)\|f^\delta_{n+1} - f_{n+1}\|\right). \end{aligned}$$

This implies the assertion.

Theorem 2 *If the assumptions of Lemmas 3 and 4 are satisfied and the constant τ in the stopping rule (12) is such that*

$$\tau > \frac{1+\eta}{1-\eta}, \tag{30}$$

then

$$\|v^\delta_{n+1} - v^\dagger\| \leq \|v^\delta_n - v^\dagger\| \ \textit{for}\ n < n_*.$$

Proof In the first step of iteration we compute v_1. If for $n = 1$ the stopping criterion (12) is not satisfied (i.e.: $1 < n_*$) then

$$C_1^\delta = \| F_1(v_1^\delta) - f_1^\delta \| > \tau\delta.$$

Moreover, Lemma 3 guarantees that $v_1 \in B_\rho(v^\dagger)$. Thus from Lemma 4 and from the assumption (30) it follows that

$$\|v_1^\delta - v^\dagger\|^2 - \|v_0^\delta - v^\dagger\|^2 \leq \frac{2}{\beta}\delta C_1^\delta \left((\eta - 1)\tau + (\eta + 1)\right) < 0.$$

In particular $v_1^\delta \in B_{\rho/4}(v^\dagger)$. Now, it follows by induction that Lemma 4 is applicable for all $0 < n < n_*$, which ends the proof of Theorem.

Corollary 1 *Let the assumptions of Theorem 2 hold. Then the stopping index n_* in (12) is finite and*

$$n_* \leq 1 + \frac{\beta}{\tau\delta^2} \frac{1}{(1-\eta)\tau - (1+\eta)} \|v_0^\delta - v^\dagger\|^2.$$

Proof Taking into account that $C_n^\delta > \tau\delta$ for $0 < n < n_*$, and

$$\left((\eta - 1)C_n^\delta + (\eta + 1)\delta\right) \leq (\eta - 1)\tau\delta + (\eta + 1)\delta \leq \tau\delta \left(\frac{\eta + 1}{\tau} - 1 + \eta\right),$$

from (28) we get

$$\frac{2}{\beta}(C_n^\delta)^2 \left(-\frac{\eta + 1}{\tau} + 1 - \eta\right) \leq -\|v_n^\delta - v^\dagger\|^2 + \|v_{n-1}^\delta - v^\dagger\|^2.$$

Adding up these inequalities for n from 1 through $n_* - 1$ we obtain

$$\tau^2\delta^2(n_* - 1) \leq \sum_{k=1}^{n_*-1} (C_k^\delta)^2 \leq \frac{\beta\tau}{(1-\eta)\tau - (1+\eta)} \|v_0^\delta - v^\dagger\|^2$$

which ends the proof.

Consider now the iteration process (23) for the exact data. The iterates are denoted by v_n in contrast with v_n^δ in the noisy case.

Corollary 2 *If $\delta = 0$, then from Theorem 2 and Lemma 4 it follows that*

$$\sum_{k=n_0}^{\infty} \| F_{k+1}(v_{k+1}) - f_{k+1} \|^2 \leq \frac{\beta}{2(1-\eta)} \|v^* - v^\dagger\|^2. \tag{31}$$

Proof For $\delta = 0$, $n_* = \min\{n > 0 : \| F_n(v_n) - f_n\| = 0\}$. If n_* is finite then $v_{n_*} = v^\dagger$ and then from (10) $v_n = v^\dagger$ for any $n > n_*$. If $n_* = \infty$ then for any $n > 0$, $\| F_n(v_n) - f_n\| > 0$. Thus (28) can be rewritten as

$$\frac{2}{\beta}(1-\eta)\| F_n(v_n) - f_n\|^2 \le \|v_{n-1} - v^\dagger\|^2 - \|v_n - v^\dagger\|^2$$

and holds for any $n > 0$. By adding these inequalities for all $n > 0$ we get (31).

Convergence of the method can be obtained in a similar way as it was done for the Landweber iteration in [6, 8] or for iterated Tikhonov-Kaczmarz method by De Cesaro et al. in [4].

Theorem 3 *If the assumptions of Theorem 2 are satisfied, then for the exact data the iteration v_n converges to $v^\dagger$ as $n \to \infty$.*

Proof The proof is based on Corollary 2 and monotonicity of $e_n = \|v_n - v^\dagger\|$. The proof follows the lines of the proof of [6, Theorem 11.4] or [4, Theorem 3.2].

Theorem 4 *Let the assumptions of Theorem 2 be satisfied and let n_* $(= n_*(\delta))$ be chosen according to the stopping rule (12). Then*

$$v^\delta_{n*} \longrightarrow v^\dagger \text{ as } \delta \longrightarrow 0.$$

Proof Because of the stability of nonlinear Tikhonov regularization (see [6, Theorem 10.2]), we have continuous dependence of v^δ_n on the data f^δ_n for any fixed iteration index n. Now, the proof is analogous to the proof of [8, Theorem 2.6] and will be omitted.

The next result shows that the considered method can be used for finding approximate value of $z^\dagger := \arg\min_{z\in I} f(z)$.

Theorem 5 *Let the assumptions of Theorem 4 be satisfied and let n_* $(= n_*(\delta))$ be chosen according to the stopping rule (12). If for $v \in B_\rho(v^*)$ and $z \in I$*

1. *$F(v;\cdot) \in C^2(I)$, and $c_0 > 0$ is such that $F''_z(v;z) \ge c_0$;*
2. *M_2 is such that $\|\frac{\partial^2}{\partial v \partial z} F(v;z)\| \le M_2$*

then

$$|z^\delta_{n_*+1} - z_{n_*+1}| \le \frac{M_2}{c_0}\|v^\delta_{n*} - v_{n_*}\| \tag{32}$$

and

$$z^\delta_{n_*} \longrightarrow z^\dagger \text{ as } \delta \longrightarrow 0.$$

Proof Let $\delta = 0$ and $z_{n+1} := \arg\min_{z\in I} F(v_n; z)$. Applying Taylor's theorem to $F'_z(v_n;\cdot)$ we get

$$F'_z(v_n; z^\dagger) = (z^\dagger - z_{n+1}) F''_z(v_n; \tilde{z})$$

for some real number $\tilde{z}$ between $z^\dagger$ and z_{n+1}. From this, it follows that

$$|z^\dagger - z_{n+1}| \le \frac{1}{c_0} |\, F_z'(v_n; z^\dagger)| \longrightarrow 0 \text{ as } n \longrightarrow \infty, \tag{33}$$

since the assumptions on F and Theorem 3 are satisfied.

Now, let us consider the noisy case. Let $n = n_*(\delta)$. Then

$$F_z'(v_n; z_{n+1}) = 0 \;\text{ and }\; F_z'(v_n^\delta; z_{n+1}^\delta) = 0.$$

Utilizing Taylor's theorem we have

$$0 = F_z'(v_n^\delta; z_{n+1}^\delta) = F_z'(v_n^\delta; z_{n+1}) + (z_{n+1}^\delta - z_{n+1}) F_z''(v_n^\delta; \tilde{z})$$

for some $\tilde{z}$, and thus

$$|z_{n+1}^\delta - z_{n+1}| \le \frac{1}{c_0} |\, F_z'(v_n^\delta; z_{n+1})|. \tag{34}$$

Similarly, for some $\tilde{v}$ from a neighborhood of v_n,

$$0 = F_z'(v_n; z_{n+1}) = F_z'(v_n^\delta; z_{n+1}) + \frac{\partial}{\partial v} F_z'(\tilde{v}; z_{n+1})(v_n^\delta - v_n).$$

Hence

$$|\, F_z'(v_n^\delta; z_{n+1})| \le M_2 \| v_n^\delta - v_n \|. \tag{35}$$

Now, (32) follows from (34) and (35). The convergence for $\delta \to 0$ follows from Theorems 3, 4 and (33).

5 Application to a Laser Beam Profile

Let $r = 3$ in (3). If $f^\delta(z_1), f^\delta(z_2), f^\delta(z_3)$ are not collinear, the equation

$$F_3 v = f_3^\delta$$

has a unique solution which can be chosen as the initial guess v^* of iteration process. If $f^\delta(z_1), f^\delta(z_2), f^\delta(z_3)$ are collinear, we define

$$v^* = \arg \min_{v \in D(F)} \left\{ \| \, F_3(v) - f_3^\delta \|_3^2 + \beta \| v - v_0 \|^2 \right\},$$

where v_0 is an a priori information about the location of $v^\dagger$. Let subsequent elements of the sequence (11) be denoted as

$$v_n^\delta = (a_n^\delta, b_n^\delta, c_n^\delta)$$

according to (18). If v_n^δ is computed, then without any additional computation we have

$$\arg\min_{z\in I} F(v_n^\delta; z) = c_n^\delta \quad \text{and} \quad \min_{z\in I} F(v_n^\delta; z) = a_n^\delta. \tag{36}$$

Moreover, if v_n^δ approximates $v^\dagger$, then the minimum of $f(z)$ as well as its localization are approximated by a_n^δ and c_n^δ, respectively, with the same accuracy.

If n_* $(= n_*(\delta))$ is chosen according to the stopping rule (12), then, in particular

$$|\, F(v_{n*}^\delta; c_{n*-1}^\delta) - f^\delta(c_{n*-1}^\delta)| \le \tau\delta. \tag{37}$$

The crucial point for stating the monotonicity (Theorem 2) and convergence (Theorem 4) is the question whether the local tangential cone condition (27) is satisfied. This problem is still open. The remaining constants appearing in lemmas and theorems of Sect. 4 can be computed and their values depend on a choice of $D(F)$ and I. For example, if $D(F) = [0.05, 0.15] \times [5, 15] \times [15, 25]$, then $\| F_n''(v)\| \le 1$.

6 Conclusion

A new method for finding an approximate minimum of a real function f given on a discrete set of points (where its values are given with some errors) has been presented. The initial problem is replaced by that of finding d parameters (denoted in the paper by $v \in R^d$) such that $F(v, \cdot)$ approximates $f(\cdot)$. Here F is appropriately chosen and $F(v, \cdot)$ are functions whose minima can easily be calculated. The method for finding v is a certain modification of the iterative Tikhonov regularization. Our modification consists in introducing different operators at different steps of iteration. Namely, we propose an iteration method in which the set of points (where noisy data f^δ are taken) changes at every step of iteration; to the fixed set of points we add additional one which is defined successively during iterations. In the paper the convergence of the method is proved but the rate of convergence is still an open problem.

As a motivation and illustration of the theory, the problem of identification of laser beam quality parameters has been used. This theory can serve as a basis of numerical programs in different applications, which will be a subject of forthcoming papers.

References

1. G. Chavent, *Nonlinear Least Squares for Inverse Problems* (Springer, Berlin, 2009)
2. H. Cheng, C.-L. Fu, X.-L. Feng, An optimal filtering method for the Cauchy problem of the Helmholtz equation. Appl. Math. Lett. **24**(6), 958–964 (2011)
3. A. De Cezaro, M. Haltmeier, A. Leitão, O. Scherzer, On Steepest-Descent-Kaczmarz methods for regularizing systems of nonlinear ill-posed equations. Appl. Math. Comput. **202**, 596–607 (2008)
4. A. De Cezaro, J. Baumeister, A. Leitão, Modified iterated Tikhonov methods for solving systems of nonlinear ill-posed equations. Inverse Prob. Imag. **5**(1), 1–17 (2011)
5. H.W. Engl, K. Kunisch, A. Neubauer, Convergence rates for Tikhonov regularization of non-linear ill-posed problems. Inverse Prob. **5**, 523–540 (1989)
6. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems* (Kluwer Academic, Dordrecht, 1996)
7. X.-L. Feng, C.-L. Fu, H. Cheng, A regularization method for solving the Cauchy problem for the Helmholtz equation. Appl. Math. Model. **35**, 3301–3315 (2011)
8. B. Kaltenbacher, A. Neubauer, O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. Radon Series on Computational and Applied Mathematics, vol. 6 (Walter de Gruyter GmbH, Berlin, 2008)
9. H. Kogelnik, T. Li, Laser beams and resonators. Appl. Opt. **5**, 1550–1567 (1966)
10. G.F. Marshall, *Handbook of Optical and Laser Scanning* (Marcel Dekker, New York, 2004)
11. E. Pruszyńska-Karbownik, K. Regiński, B. Mroziewicz, M. Szymański, P. Karbownik, K. Kosiel, A. Szerling, Analysis of the spatial distribution of radiation emitted by MIR quantum cascade lasers, in *Proceedings of SPIE 8702, Laser Technology 2012: Progress in Lasers*, 87020E (2013)
12. H.H. Qin, T. Wei, Two regularization methods for the Cauchy problems of the Helmholtz equation. Appl. Math. Model. **34**(4), 947–967 (2010)
13. T. Regińska, Regularization methods for a mathematical model of laser beams. Eur. J. Math. Comput. Appl. **1**(2), 39–49 (2014)
14. T. Regińska, K. Regiński, Approximate solution of a Cauchy problem for the Helmholtz equation. Inverse Prob. **22**, 975–989 (2006)
15. T. Regińska, K. Regiński, Regularization strategy for determining laser beam quality parameters, J. Inverse Ill-Posed Prob. **23**(6), 657–671 (2015)
16. T. Regińska, U. Tautenhahn, Conditional stability estimates and regularization with applications to Cauchy problems for the Helmholtz equation. Numer. Funct. Anal. Optim. **30**, 1065–1097 (2009)
17. K. Regiński, B. Mroziewicz, E. Pruszyńska-Karbownik, Goniometric method of measuring the spatial distribution of intensity of radiation emitted by quantum cascade lasers. Elektronika **10**, 48–51 (2011) [in Polish]
18. T.S. Ross, *Laser Beam Quality Metrics* (SPIE Press, Bellingham, 2013)
19. A.E. Siegman, *Lasers* (University Science Books, Mill Valey, CA, 1986)
20. X.-T. Xiong, A regularization method for a Cauchy problem of the Helmholtz equation. J. Comput. Appl. Math. **233**(8), 1723–1732 (2010)
21. X.-T. Xiong, C.-L. Fu, Two approximate methods of a Cauchy problem for the Helmholtz equation. Comput. Appl. Math. **26**(2), 285–307 (2007)
22. H.W. Zhang, H.H. Qin, T. Wei, A quasi-reversibility regularization method for the Cauchy problem of the Helmholtz equation. Int. J. Comput. Math. **88**(4), 839–850 (2011)

Tomographic Terahertz Imaging Using Sequential Subspace Optimization

Anne Wald and Thomas Schuster

Abstract Terahertz tomography aims for reconstructing the complex refractive index of a specimen, which is illuminated by electromagnetic radiation in the terahertz regime, from measurements of the resulting (total) electric field outside the object. The illuminating radiation is reflected, refracted, and absorbed by the object. In this work, we reconstruct the complex refractive index from tomographic measurements by means of regularization techniques in order to detect defects such as holes, cracks, and other inclusions, or to identify different materials and the moisture content. Mathematically, we are dealing with a nonlinear parameter identification problem for the two-dimensional Helmholtz equation, and solve it with the Landweber method and sequential subspace optimization. The article concludes with some numerical experiments.

1 Introduction

Terahertz (THz) tomography refers to the nondestructive testing of dielectric objects by illuminating them by electromagnetic radiation with a frequency of about 0.1–10 THz and measuring the resulting total electric field, see, e.g., [6, 11]. In the electromagnetic spectrum, THz radiation is located between infrared and microwave radiation. We are thus dealing with a frequency range for which a neglection of either the wave character or the ray character is not convenient. Radiation with a lower frequency has a prominent wave character, and the description of its propagation in space leads to typical scattering problems as in RADAR and microwave tomography, or, in the case of mechanical waves, ultrasound tomography [7]. Conversely, radiation with a higher frequency, such as X-radiation or gamma radiation, can be treated by purely considering its ray character: the rays travel along straight lines, which is used in classical computerized tomography [15].

A. Wald · T. Schuster (✉)
Saarland University, Saarbrücken, Germany
e-mail: anne.wald@num.uni-sb.de; thomas.schuster@num.uni-sb.de

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_14

The inverse problem of two-dimensional THz tomography has already been addressed by Tepe et al. in [20], where the algebraic reconstruction technique (ART) has been modified to using refracted ray paths and taking into account reflection losses. The mathematical model is based on the Radon transform along these refracted ray paths. In order to complement this work, we now want to treat an approach that originates from scattering theory. Note that our approach allows an easy inclusion of the rather complex geometry of the incident Gaussian beam in the model.

Outline Section 2 is devoted to a detailed analysis of the forward model of THz tomography. We start by a deduction of the Helmholtz equation as mathematical model for the propagation of time-harmonic terahertz waves using Maxwell's equations as a starting point (Sect. 2.1). Section 2.1 also contains a motivation of the Gaussian beam as physical model for terahertz beams. As boundary values we use the Robin condition mimicking the Sommerfeld radiation condition for bounded domains. Existence and uniqueness of a weak solution of the arising boundary value problem is proven in Sect. 2.2. Since we intend to use iterative methods for solving the inverse problem, we need Fréchet differentiability. We give an explicit representation of the Fréchet derivative of the scattering map (Sect. 2.3) and of its adjoint (Sect. 2.4). Finally we construct the so-called observation operator which contains the details of the measurement process (Sect. 2.5). This completes the analytical model of THz tomography. Section 3 contains then implementations and numerical experiments for solving the inverse imaging problem by using Landweber's method as well as regularizing sequential subspace optimization (RESESOP). The reconstruction techniques are described in Sect. 3.1 and the numerical verification is contained in Sect. 3.2.

2 An Analysis of the Forward Operator in THz Tomography

We begin by introducing the physical basics and the notation. The overall goal is a description of the nonlinear forward operator F, which turns out to be the composition of a scattering operator S, the trace operator γ, and a suitable observation operator Q. The resulting nonlinear inverse problem

$$F(m) = y, \tag{1}$$

where m represents the complex refractive index and y are the measured data, is solved iteratively by the Landweber method and an adapted sequential subspace optimization method that originates from the techniques developed in [21] for nonlinear inverse problems, which are themselves inspired by subspace techniques from the theory of linear inverse problems [14, 18, 19]. To this end, expressions of the form $F'(m)^*(F(m)-y)$ need to be evaluated. We present all necessary tools and analyze the occurring operators in detail.

2.1 Physical Basics and Notations

Generally, the propagation of electromagnetic waves in space is described by Maxwell's equations, which are coupled first order partial differential equations for the electric field $\mathbf{E} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{C}, (t, \mathbf{x}) \mapsto \mathbf{E}(t, \mathbf{x})$, and the induction field $\mathbf{B} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{C}, (t, \mathbf{x}) \mapsto \mathbf{B}(t, \mathbf{x})$. If the wave travels through dielectric, non-magnetic media, where the dielectric permittivity $\epsilon : \mathbb{R}^3 \to \mathbb{R}$ satisfies

$$\nabla \cdot (\epsilon(\mathbf{x})\mathbf{E}(t, \mathbf{x})) \approx \epsilon(\mathbf{x})\nabla \cdot \mathbf{E}(t, \mathbf{x}),$$

and the magnetic permeability $\mu = \mu_0$ is constant, the electric field $\mathbf{E}$ (and also the induction field $\mathbf{B}$) solves the wave equation

$$\Delta \mathbf{E}(t, \mathbf{x}) - \epsilon(\mathbf{x})\mu_0 \frac{\partial^2}{\partial t^2}\mathbf{E}(t, \mathbf{x}) = 0.$$

If there are neither sinks, nor sources. In our case, we use time-harmonic electromagnetic waves $\mathbf{E}(t, \mathbf{x}) = \mathbf{u}(\mathbf{x})e^{i\omega t}$ with a fixed frequency ω, such that a separation of variables yields the *Helmholtz equation*

$$\Delta \mathbf{u}(\mathbf{x}) + \tilde{k}^2 \mathbf{u}(\mathbf{x}) = 0$$

with $\tilde{k}^2 = \omega^2 \epsilon(\mathbf{x})\mu_0$. In vacuum, we have $\epsilon(\mathbf{x})\mu_0 = c_0^{-2}$, where c_0 is the speed of light in free space.

In absorbing, anisotropic media, we use a complex electric permittivity $\tilde{\epsilon}(\mathbf{x}) = \epsilon_1(\mathbf{x}) + i\epsilon_2(\mathbf{x})$ to model the absorption losses and write

$$\tilde{k} = k_0 \tilde{n} = k_0(n + i\kappa),$$

where $\tilde{n}$ is the *complex refractive index*, n is the refractive index and κ is the extinction coefficient. The Helmholtz equation thus reads

$$\Delta \mathbf{u}(\mathbf{x}) + k_0^2 \tilde{n}^2 \mathbf{u}(\mathbf{x}) = 0. \tag{2}$$

Note that the objects in question usually consist of plastics or ceramics (suitable materials are discussed in [9]), which have a very small extinction coefficient in the THz range. This implies a high penetration depth, explaining the relevance of THz radiation in nondestructive testing.

The radiation that is used to illuminate the object is a time-harmonic, electromagnetic Gaussian beam $\mathbf{u}_G : \mathbb{R}^3 \to \mathbb{C}^3$ with a fixed wave number $k_0 > 0$. An essential property of Gaussian beams is that they propagate in a certain direction, in our case in y-direction. A mathematical description of Gaussian beams is given as an approximate solution of the *paraxial Helmholtz equation*

$$\left(\frac{\partial^2}{\partial x^2} + 2ik_0\frac{\partial}{\partial y} + \frac{\partial^2}{\partial z^2} + 2k_0^2\right)\mathbf{u}(\mathbf{x}) = 0,$$

which is derived from the Helmholtz equation $\Delta\mathbf{u}+k_0^2\mathbf{u} = 0$ in vacuum by assuming that the change in the amplitude of the electric field in the direction y of propagation is small compared to orders of the wave length. Each component u_G of the electric field $\mathbf{u}$ is expressed in cylindric coordinates by

$$u_\mathrm{G}(\mathbf{x}) = a_0\frac{W_0}{W(y)}\exp\left(-\frac{r^2}{W^2(y)}\right)\exp\left(i\frac{k_0 y - \phi(y) + k_0 r^2}{2R(y)}\right), \tag{3}$$

where $r = \sqrt{x^2+z^2}$ is the radial component, a_0 is the amplitude at the origin, $W_0 = W(0)$ the *beam waist*, the function $\phi(y) = \arctan\left({}^{y}/_{y_0}\right)$ the *Gouy phase*, and $y_0 \in \mathbb{R}$. The factor

$$R(y) = y\left(1 + \frac{y_0^2}{y^2}\right)$$

defines the radius of curvature of the wavefronts. The function

$$W(y) = W_0\sqrt{1 + \frac{y^2}{y_0^2}}$$

is called the *spot size parameter*. A full derivation and discussion is to be found in [17].

Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded domain with a C^1-boundary $\partial\Omega$. The incident field, which is denoted by $\mathbf{u}_\mathrm{i}$, is thus given analytically by (3). It approximately solves the Helmholtz equation (2). The total field $\mathbf{u}_\mathrm{t}$ fulfills

$$\Delta\mathbf{u}_\mathrm{t} + k_0^2\tilde{n}^2\mathbf{u}_\mathrm{t} = 0$$

in Ω and is obtained by the superposition principle as the sum

$$\mathbf{u}_\mathrm{t} = \mathbf{u}_\mathrm{i} + \mathbf{u}_\mathrm{sc}, \tag{4}$$

where $\mathbf{u}_\mathrm{sc}$ is the scattered electric field. Note that the low extinction coefficient motivates the use of the superposition principle.

The superposition principle (4) and the Helmholtz equation (2) now yield the scattered field $\mathbf{u}_\mathrm{sc}$ as the solution of the inhomogeneous Helmholtz equation

$$\Delta\mathbf{u}_\mathrm{sc} + k_0^2\tilde{n}^2\mathbf{u}_\mathrm{sc} = k_0^2(1-\tilde{n}^2)\mathbf{u}_\mathrm{i}.$$

In the context of a (numerical) solution of the above partial differential equation, we have to impose suitable boundary conditions on $\mathbf{u}_{\mathrm{sc}}$. The correct physical choice are radiation conditions such as the Sommerfeld radiation condition. Since we are required to work on a bounded domain, we use *Robin boundary conditions* that approximate the Sommerfeld radiation condition and are sometimes referred to as *first order scattering boundary conditions.* They are given by

$$\frac{\partial \mathbf{u}_{\mathrm{sc}}}{\partial \mathbf{n}} - ik_0 \mathbf{u}_{\mathrm{sc}} = 0.$$

By $\frac{\partial}{\partial \mathbf{n}}$ we denote the partial directional derivative in the direction of the outward normal vector $\mathbf{n}$ of the boundary.

Since we are interested in two-dimensional THz tomography, we make some further simplifications. First, we only aim at reconstructing the object in the x-y-plane, such that we define $\tilde{n}$ as a complex function on Ω and neglect the influence of the object outside this plane. Second, the use of polarization filters in the receivers allows us to restrict our considerations to the z-component of the electric field $\mathbf{u}_{\mathrm{t}}$. Given the complex refractive index in Ω, the z-component of the total electric field, denoted by u_{t}, is consequently obtained by solving

$$\begin{aligned} \Delta u_{\mathrm{sc}} + k_0^2 \tilde{n}^2 u_{\mathrm{sc}} &= k_0^2 (1 - \tilde{n}^2) u_{\mathrm{i}} \quad \text{in } \Omega, \\ \frac{\partial u_{\mathrm{sc}}}{\partial \mathbf{n}} - ik_0 u_{\mathrm{sc}} &= 0 \quad \text{on } \partial\Omega, \\ u_{\mathrm{sc}} + u_{\mathrm{i}} &= u_{\mathrm{t}} \quad \text{in } \overline{\Omega}, \end{aligned}$$

where u_{i} and u_{sc} are the respective z-components of the incident and scattered field.

Remark 1 The propagation of THz radiation is barely influenced by the presence of air. Consequently, the complex refractive index $\tilde{n}_{\mathrm{air}}$ of air fulfills $\tilde{n}_{\mathrm{air}} \approx 1$ and $1 - \tilde{n}_{\mathrm{air}}^2 \approx 0$.

Generally, the complex refractive index $\tilde{n} : \Omega \to \mathbb{C}$ is a bounded, complex-valued function on Ω. Note that the function $\tilde{n}$ contains the same information as the function $1 - \tilde{n}^2$. This motivates the following definition.

Definition 1 Let $m : \Omega \to \mathbb{C}$ be the bounded, complex-valued function given by

$$m(\mathbf{x}) = 1 - \tilde{n}^2(\mathbf{x})$$

for all $\mathbf{x} = (x, y) \in \Omega$.

Remark 2 Obviously, we have $m \in L^\infty(\Omega)$. Since m vanishes outside the object, we also have

$$m \in L^2_{\mathrm{comp}}(\Omega) := \left\{ f \in L^2(\Omega) \ : \ \mathrm{supp}(f) \subseteq \Omega \right\},$$

where the support of $f \in L^2(\Omega)$ is defined as

$$\operatorname{supp}(f) := \Omega \setminus \bigcup \{U \subseteq \Omega \,:\, U \text{ open}, f|_U = 0 \text{ a.e.}\}.$$

In summary, the inclusion

$$m \in L^\infty(\Omega) \cap L^2_{\text{comp}}(\Omega) \subseteq L^2(\Omega)$$

allows us to treat m as an element of the Hilbert space $L^2(\Omega)$, which is necessary for the convergence of our reconstruction algorithms. For the subsequent analysis, we will, however, exploit that $m \in L^\infty(\Omega)$. Furthermore, we define

$$L^{2,\infty}_{\text{comp}}(\Omega) := L^\infty(\Omega) \cap L^2_{\text{comp}}(\Omega).$$

From now on, we silently abuse notation and refer to m as the complex refractive index of the test object.

2.2 Existence and Uniqueness of a Weak Solution

Up to now, we have discussed the physical model for the propagation of the THz beam through an object with complex refractive index m. The first part of our forward operator is thus given as the parameter-to-solution mapping, which maps m to the total electric field

$$u_{\text{t}} = u_{\text{i}} + u_{\text{sc}} \quad \text{in } \overline{\Omega}, \tag{5}$$

where u_{i} is the incident Gaussian beam and the scattered field u_{sc} solves the boundary value problem

$$\Delta u_{\text{sc}} + k_0^2(1-m)u_{\text{sc}} = k_0^2 m u_{\text{i}} \quad \text{in } \Omega, \tag{6}$$

$$\frac{\partial u_{\text{sc}}}{\partial \mathbf{n}} - ik_0 u_{\text{sc}} = 0 \quad \text{on } \partial\Omega. \tag{7}$$

In a first step, we establish the existence and uniqueness of a weak solution u of (6), (7), i.e., we show that there is a unique $u \in H^1(\Omega)$, which solves the respective *variational problem*

$$a(u,v) = b(v) \quad \text{for all } v \in H^1(\Omega), \tag{8}$$

where the sesquilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{C}$ is defined by

$$a(u,v) := (\nabla u, \nabla v)_{L^2(\Omega)} - k_0^2\left((1-m)u, v\right)_{L^2(\Omega)} - ik_0(u,v)_{L^2(\partial\Omega)}, \tag{9}$$

and the linear functional $b : H^1(\Omega) \to \mathbb{C}$ by

$$b(v) := -k_0^2 (m u_{\mathrm{i}}, v)_{L^2(\Omega)}. \tag{10}$$

Remark 3 The restriction of $u \in H^1(\Omega)$ to $\partial\Omega$ is to be understood in the context of the *trace operator*

$$\gamma : H^1(\Omega) \to L^2(\partial\Omega),\ u \mapsto u|_{\partial\Omega},$$

which is well-defined if $\partial\Omega$ is of class C^1. For reference, see, e.g., [2, 8].

Our approach is inspired by the analysis of the *inverse medium problem* by Bao and Li in [3, 4]. We begin with a uniqueness result.

Theorem 1 *For any $m \in L^{2,\infty}_{\mathrm{comp}}(\Omega)$ with real part $m_{\mathrm{r}} := \mathrm{Re}(m)$ and imaginary part $m_{\mathrm{i}} := \mathrm{Im}(m) \le 0$ there is at most one solution to the variational scattering problem* (8)*.*

Proof We consider the variational problem (8) for $v = u$, such that $a(u,u) = b(u)$. Due to the linearity of the elliptic partial differential equation it suffices to show that $u = 0$ in case there is no incident field, i.e., $u_{\mathrm{i}} = 0$. We then have

$$a(u,u) = \int_\Omega \nabla u \cdot \nabla \overline{u}\, \mathrm{d}\mathbf{x} - k_0^2 \int_\Omega (1-m) u \cdot \overline{u}\, \mathrm{d}\mathbf{x} - i k_0 \int_{\partial\Omega} u \cdot \overline{u}\, \mathrm{d}s_{\mathbf{x}} = 0.$$

We write $m = m_{\mathrm{r}} + i m_{\mathrm{i}}$ and obtain

$$i k_0^2 \int_\Omega m_{\mathrm{i}} u \cdot \overline{u}\, \mathrm{d}\mathbf{x} = i k_0 \int_{\partial\Omega} u \cdot \overline{u}\, \mathrm{d}s_{\mathbf{x}}$$

for the imaginary part of the previous equation. We thus have

$$\|u\|^2_{L^2(\partial\Omega)} = k_0 \int_\Omega m_{\mathrm{i}} |u|^2\, \mathrm{d}\mathbf{x} \le 0,$$

as $m_{\mathrm{i}}(\mathbf{x}) \le 0$ for all $\mathbf{x} \in \Omega$. This yields $u|_{\partial\Omega} = 0$, such that our boundary condition now reads $\frac{\partial u}{\partial \mathbf{n}} = 0$ (Neumann boundary conditions) as well as $m_{\mathrm{i}} \cdot |u|^2 = 0$ a.e., which implies $m_{\mathrm{i}} \cdot u = 0$ a.e.. Hence, it remains to show that there is at most one solution of the Neumann boundary value problem

$$\begin{aligned} \Delta u + k_0^2(1 - m_{\mathrm{r}}) u &= 0 \qquad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{11}$$

Corollary 8.2 from [10] yields $u = 0$ in (11) on Ω and consequently, we have $u = 0$ on $\overline{\Omega}$. □

Remark 4 The condition $m_{\mathrm{i}} \leq 0$ holds naturally due to $m = 1 - (n + i\kappa)^2 = 1 - n^2 + \kappa^2 - i \cdot 2n\kappa$ and $n \geq 1$, $\kappa \geq 0$. More precisely, it suffices that $m_{\mathrm{i}} \leq 0$ almost everywhere in Ω.

Having established the uniqueness of a solution of the variational problem (8), we now have to prove its existence. In the following, we will always assume that the complex refractive index m satisfies

$$\mathrm{Im}(m) \leq 0,$$

such that we can apply Theorem 1.

Throughout this section, let $c_j > 0, j \in \mathbb{N}$, be positive constants.

Theorem 2 *Let Ω be a bounded domain with C^1-boundary $\partial\Omega$, $k_0 \in \mathbb{R}^+$ a nonnegative constant and $u_{\mathrm{i}} \in H^1(\Omega)$ the incident field. If $m \in L^{2,\infty}_{\mathrm{comp}}(\Omega)$, the variational problem* (8) *possesses a unique, weak solution $u \in H^1(\Omega)$ satisfying*

$$\|u\|_{H^1(\Omega)} \leq C_1 \|m\|_{L^\infty(\Omega)} \, \|u_{\mathrm{i}}\|_{L^2(\Omega)} \tag{12}$$

for some constant $C_1 = C_1(k_0, \Omega) > 0$.

Proof We split the sesquilinear form a from (9) in two sesquilinear forms $a_1, a_2 : H^1(\Omega) \times H^1(\Omega) \to \mathbb{C}$, where

$$a_1(v_1, v_2) = (\nabla v_1, \nabla v_2)_{L^2(\Omega)} - ik_0(v_1, v_2)_{L^2(\partial\Omega)},$$

and

$$a_2(v_1, v_2) = -\big((1 - m)v_1, v_2\big)_{L^2(\Omega)},$$

such that

$$a = a_1 + k_0^2 a_2.$$

Note that a_2 can be defined on $L^2(\Omega) \times L^2(\Omega)$ as well.

We first show that a_1 is bounded and coercive. From

$$\begin{aligned} |a_1(v_1, v_2)| &\leq |v_1|_{H^1(\Omega)} \cdot |v_2|_{H^1(\Omega)} + k_0 \, \|v_1\|_{L^2(\partial\Omega)} \cdot \|v_2\|_{L^2(\partial\Omega)} \\ &\leq \|v_1\|_{H^1(\Omega)} \cdot \|v_2\|_{H^1(\Omega)} + c_1 k_0 \, \|v_1\|_{H^1(\Omega)} \cdot \|v_2\|_{H^1(\Omega)} \\ &\leq c_2 k_0 \|v_1\|_{H^1(\Omega)} \cdot \|v_2\|_{H^1(\Omega)} \end{aligned}$$

we obtain the boundedness of a_1. We have used the semi-norm $|\cdot|_{H^1(\Omega)}$ on $H^1(\Omega)$, given by

$$|v|^2_{H^1(\Omega)} = \int_\Omega \nabla v \cdot \nabla \overline{v} \, \mathrm{d}\mathbf{x}$$

and satisfying $|v|_{H^1(\Omega)} \leq \|v\|_{H^1(\Omega)}$, and the trace theorem (for reference, see, e.g., [1, 8]). The constant $c_2 > 0$ depends only on Ω.

The coercivity of a_1 is obtained by estimating

$$\begin{aligned}\left|\left(a_1(v,v)\right)\right| &= \left(|v|^4_{H^1(\Omega)} + k_0^2\|v\|^4_{L^2(\partial\Omega)}\right)^{1/2} \\ &\geq c_3\left(|v|^2_{H^1(\Omega)} + k_0\|v\|^2_{L^2(\partial\Omega)}\right) \\ &\geq c_4 k_0\|v\|^2_{H^1(\Omega)},\end{aligned}$$

using the equivalence of the Euclidean norm and the ℓ^1-norm on $\mathbb{R}^2$,

$$\left\|\left(|v|^2_{H^1(\Omega)}, k_0\|v\|^2_{L^2(\partial\Omega)}\right)^T\right\|_2 \geq c_3\left\|\left(|v|^2_{H^1(\Omega)}, k_0\|v\|^2_{L^2(\partial\Omega)}\right)^T\right\|_1,$$

and a norm equivalence that can be found in [2, p. 214]. The constant c_4 depends only on Ω (also [2, p. 214]).

In the following, we denote by

$$\Phi : H^1(\Omega) \to (H^1(\Omega))^*,\ v \mapsto (v,\cdot)_{H^1(\Omega)}$$

the isometric Riesz isomorphism (see, e.g., [5] and [22], Theorem V.3.6). The Lax-Milgram lemma yields the existence of an isomorphism $T : H^1(\Omega) \to H^1(\Omega)$ with $\|T\|_{H^1(\Omega)\to H^1(\Omega)} \leq c_2 k_0$ and $\|T^{-1}\|_{H^1(\Omega)\to H^1(\Omega)} \leq (c_4 k_0)^{-1}$, which satisfies

$$a_1(u,v) = (Tu,v)_{H^1(\Omega)}$$

for all $u, v \in H^1(\Omega)$ (this operator is *associated* to a_1, see [5, 16]). Now consider the mapping

$$\mathscr{B} : L^2(\Omega) \to (H^1(\Omega))^*,\ s \mapsto a_2(s,\cdot),$$

which is well-defined and the mapping $a_2(s,\cdot)$ is antilinear for $s \in L^2(\Omega)$. For $w \in H^1(\Omega)$, we write $\mathscr{B}s[w] = a_2(s,w)$. For all $s \in L^2(\Omega)$ and $v \in H^1(\Omega)$, we have

$$\begin{aligned}|a_2(s,v)| &= \left|\left((1-m)s, v\right)_{L^2(\Omega)}\right| \leq \|1-m\|_{L^\infty(\Omega)}\|s\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\ &\leq \|1-m\|_{L^\infty(\Omega)}\|s\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}.\end{aligned}$$

Consequently, $a_2(s,\cdot)$ is continuous and we have

$$\|a_2(s,\cdot)\|_{(H^1(\Omega))^*} \leq \|1-m\|_{L^\infty(\Omega)}\|s\|_{L^2(\Omega)}.$$

This estimate also yields the boundedness of the linear mapping $\mathscr{B}$ with

$$\|\mathscr{B}\|_{L^2(\Omega)\to(H^1(\Omega))^*} \leq \|1-m\|_{L^\infty(\Omega)}.$$

We now define the linear operator

$$\widetilde{\mathscr{A}} := T^{-1}\Phi^{-1}\mathscr{B} : L^2(\Omega) \to H^1(\Omega).$$

Consider the operator

$$\mathscr{A} : L^2(\Omega) \to H^1(\Omega) \hookrightarrow L^2(\Omega),\ s \mapsto \widetilde{\mathscr{A}}s.$$

Since $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$, $\mathscr{A} : L^2(\Omega) \to L^2(\Omega)$ is compact as a composition of a compact and a bounded linear operator. Note that $\mathscr{A}(L^2(\Omega)) \subseteq H^1(\Omega)$. We obtain for every $s \in L^2(\Omega)$ the estimate

$$\begin{aligned}\|\mathscr{A}s\|_{H^1(\Omega)} = \|\widetilde{\mathscr{A}}s\|_{H^1(\Omega)} &\leq \|T^{-1}\|_{H^1(\Omega)\to H^1(\Omega)} \cdot \|\Phi^{-1}\mathscr{B}s\|_{H^1(\Omega)}\\ &\leq (c_4k_0)^{-1}\|\mathscr{B}s\|_{(H^1(\Omega))^*} \leq (c_4k_0)^{-1}\|1-m\|_{L^\infty(\Omega)}\|s\|_{L^2(\Omega)}\end{aligned}$$

and compute

$$\begin{aligned}a_1(\mathscr{A}s, w) = a_1\big(\widetilde{\mathscr{A}}s, w\big) = a_1\big(T^{-1}\Phi^{-1}\mathscr{B}s, w\big) &= \big(\Phi^{-1}\mathscr{B}s, w\big)_{H^1(\Omega)}\\ &= \big(\Phi(\Phi^{-1}\mathscr{B}s)\big)[w] = \mathscr{B}s[w] = a_2(s, w).\end{aligned}$$

It is easily verified that $\mathscr{A}$ is unique having this property.

By $I : L^2(\Omega) \to L^2(\Omega)$, we denote the identity mapping in $L^2(\Omega)$. In the next step, we show that for every $k_0 > 0$ the operator $I + k_0^2\mathscr{A}$ is injective.

Let $s \in \mathscr{N}\big(I + k_0^2\mathscr{A}\big) \subseteq L^2(\Omega)$. Then we have $s = -k_0^2\mathscr{A}s \in H^1(\Omega)$ and thus

$$\begin{aligned}a_1(s, s) + k_0^2a_2(s, s) &= a_1\big(-k_0^2\mathscr{A}s, s\big) + k_0^2a_2(s, s)\\ &= -k_0^2a_1\big(\mathscr{A}s, s\big) + k_0^2a_2(s, s)\\ &= -k_0^2a_2(s, s) + k_0^2a_2(s, s) = 0.\end{aligned}$$

Our uniqueness result, Theorem 1, now yields $s = 0$. Hence, the operator $I + k_0^2\mathscr{A}$ is injective.

Consider now the (antilinear) functional $b \in (H^1(\Omega))^*$, see (10), and let $u \in H^1(\Omega)$. Using the definitions of a_1, a_2, and the operator $\mathscr{A}$, we see that our original variational problem of finding $u \in H^1(\Omega)$,

$$a_1(u, v) + k_0^2a_2(u, v) = b(v) \qquad \text{for all } v \in H^1(\Omega)$$

is equivalent to finding $u \in H^1(\Omega)$, such that

$$b(v) = a_1\bigl(u + k_0^2 \mathscr{A} u, v\bigr) = \bigl(T\bigl(u + k_0^2 \mathscr{A}\bigr), v\bigr)_{H^1(\Omega)}$$

for all $v \in H^1(\Omega)$. This yields $\Phi\bigl(T\bigl(I + k_0^2 \mathscr{A}\bigr)u\bigr) = b$ and we finally obtain $u = \bigl(I + k_0^2 \mathscr{A}\bigr)^{-1} T^{-1} \Phi^{-1}(b)$, such that our variational problem has at least one solution $u \in H^1(\Omega)$. Now put

$$\tilde{u} := T^{-1} \Phi^{-1}(b)$$

and we see that

$$b(v) = \bigl(\Phi T \tilde{u}\bigr)[v] = (T\tilde{u}, v) = a_1(\tilde{u}, v) \qquad \text{for all } v \in H^1(\Omega).$$

Since $\mathscr{A}$ is compact and $I + k_0^2 \mathscr{A}$ is injective, the Fredholm alternative is applicable and yields the existence of a unique $u \in H^1(\Omega)$, such that

$$\bigl(I + k_0^2 \mathscr{A}\bigr)u = \tilde{u}, \tag{13}$$

and the boundedness of the inverse of $I + k_0^2 \mathscr{A}$, i.e.,

$$\bigl\| \bigl(I + k_0^2 \mathscr{A}\bigr)^{-1} \bigr\|_{H^1(\Omega) \to H^1(\Omega)} \le c_5, \tag{14}$$

where $c_5 = c_5(k_0)$ depends on the wave number k_0. We estimate

$$\begin{aligned}
\|u\|_{H^1(\Omega)} &\le \bigl\| I + k_0^2 \mathscr{A} \bigr\|^{-1}_{H^1(\Omega)} \|\tilde{u}\|_{H^1(\Omega)} \\
&\le c_5 \bigl\| T^{-1} \bigr\|_{H^1(\Omega) \to H^1(\Omega)} \bigl\| \Phi^{-1} b \bigr\|_{H^1(\Omega)} \\
&\le c_5 (c_4 k_0)^{-1} \bigl\| \Phi^{-1} b \bigr\|_{H^1(\Omega)}
\end{aligned}$$

and together with the boundedness of b, which we derive from

$$\begin{aligned}
\|b\|_{(H^1(\Omega))^*} &= \sup_{\|v\|_{H^1(\Omega)}=1} |b(v)| = \sup_{\|v\|_{H^1(\Omega)}=1} |k_0^2 (m u_{\mathrm{i}}, v)| \\
&\le \sup_{\|v\|_{H^1(\Omega)}=1} k_0^2 \|m\|_{L^\infty(\Omega)} \|u_{\mathrm{i}}\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\
&\le \sup_{\|v\|_{H^1(\Omega)}=1} k_0^2 \|m\|_{L^\infty(\Omega)} \|u_{\mathrm{i}}\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)} \\
&\le k_0^2 \|m\|_{L^\infty(\Omega)} \|u_{\mathrm{i}}\|_{L^2(\Omega)},
\end{aligned}$$

we finally arrive at

$$\begin{aligned}\|u\|_{H^1(\Omega)} &\leq c_5(c_4k_0)^{-1} \cdot k_0^2\|m\|_{L^\infty(\Omega)}\|u_\mathrm{i}\|_{L^2(\Omega)}\\ &= c_4^{-1}c_5 \cdot k_0\|m\|_{L^\infty(\Omega)}\|u_\mathrm{i}\|_{L^2(\Omega)}\\ &= C_1\|m\|_{L^\infty(\Omega)}\|u_\mathrm{i}\|_{L^2(\Omega)}\end{aligned}$$

where $C_1 = c_4^{-1}c_5 \cdot k_0$ depends on k_0 and Ω. □

We now define the first part of our forward operator, the so-called *scattering operator*.

Definition 2 Let $u_\mathrm{i} \in H^1(\Omega)$. We define the mapping

$$S : \mathscr{D}(S) \to H^1(\Omega),\ m \mapsto S(m) := u_\mathrm{t},$$

where

$$\mathscr{D}(S) \subseteq \Big\{m \in L^{2,\infty}_{\mathrm{comp}}(\Omega) : \|m\|_{L^\infty(\Omega)} \leq M \ \text{ and } \ \mathrm{Im}(m) \leq 0\Big\}$$

for some fixed $M > 0$, such that $u_\mathrm{t} = u_\mathrm{i} + u_\mathrm{sc}$ and u_sc is the solution of the variational problem (8) of the boundary value problem (6), (7).

Note, that S is well-defined according to our previous results. The analysis of S is concluded by a continuity result. In the following, we always refer to weak solutions of the occurring boundary value problems.

Lemma 1 *Let $m_1, m_2 \in \mathscr{D}(S)$. Then S is Lipschitz continuous on $\mathscr{D}(S)$, i.e., we have*

$$\|S(m_1) - S(m_2)\|_{H^1(\Omega)} \leq C_2 \,\|m_1 - m_2\|_{L^\infty(\Omega)}\, \|u_\mathrm{i}\|_{L^2(\Omega)}\,, \tag{15}$$

where $C_2 = C_2\,(k_0, \Omega) > 0$ and $u_\mathrm{i} \in H^1(\Omega)$ is the incident field.

Proof We set $u_{(1)} := S(m_1) - u_\mathrm{i}$ and $u_{(2)} := S(m_2) - u_\mathrm{i}$, such that

$$\Delta u_{(j)} + k_0^2\left(1 - m_j\right)u_{(j)} = k_0^2 m_j u_\mathrm{i}, \quad j = 1, 2. \tag{16}$$

From Theorem 2 we deduce that

$$\left\|u_{(j)}\right\|_{H^1(\Omega)} \leq C_1\|m_j\|_{L^\infty(\Omega)}\, \|u_\mathrm{i}\|_{L^2(\Omega)}\,. \tag{17}$$

By subtracting Eq. (16) for $j = 2$ from the one for $j = 1$ and by setting $w := u_{(1)} - u_{(2)}$ we obtain

$$\Delta w + k_0^2\,(1 - m_1)\,w = k_0^2(m_1 - m_2)\left(u_\mathrm{i} + u_{(2)}\right).$$

Note that w satisfies the Robin boundary condition of our scattering problem, such that we can apply Theorem 2. We thus have

$$\|w\|_{H^1(\Omega)} \leq C_1 \|m_1 - m_2\|_{L^\infty(\Omega)} \left\|u_\mathrm{i} + u_{(2)}\right\|_{L^2(\Omega)} .$$

Combining this estimate with $\left\|u_{(2)}\right\|_{L^2(\Omega)} \leq \left\|u_{(2)}\right\|_{H^1(\Omega)}$ and (17) for $j = 2$ yields

$$\begin{aligned}
\|S(m_1) - S(m_2)\|_{H^1(\Omega)} &= \left\|u_{(1)} - u_{(2)}\right\|_{H^1(\Omega)} \\
&\leq C_1 \|m_1 - m_2\|_{L^\infty(\Omega)} \left\|u_\mathrm{i} + u_{(2)}\right\|_{L^2(\Omega)} \\
&\leq C_1 \|m_1 - m_2\|_{L^\infty(\Omega)} \left(\|u_\mathrm{i}\|_{L^2(\Omega)} + \left\|u_{(2)}\right\|_{L^2(\Omega)}\right) \\
&\leq C_1 \big(1 + C_1 \|m_2\|_{L^\infty(\Omega)}\big) \|m_1 - m_2\|_{L^\infty(\Omega)} \|u_\mathrm{i}\|_{L^2(\Omega)} \\
&\leq C_1 \big(1 + C_1 M\big) \|m_1 - m_2\|_{L^\infty(\Omega)} \|u_\mathrm{i}\|_{L^2(\Omega)} .
\end{aligned}$$

By setting $C_2 = C_2\,(k_0, \Omega) := C_1\big(1 + C_1 M\big)$ we obtain the continuity estimate (15).
□

2.3 The Linearized Scattering Problem

In the numerical reconstruction of the complex refractive index, the linearization of the scattering map S in $m \in \mathscr{D}(S)$ plays an important role. Before proving the Fréchet differentiability of S, we define a further parameter-to-solution operator that will prove useful for our further investigations.

Definition 3 For a fixed $m \in \mathscr{D}(S)$ and the respective total field $u_\mathrm{t} := S(m)$, let $T_m : \mathscr{D}(S) \to H^1(\Omega)$ be the operator that maps $h \in \mathscr{D}(S)$ to the unique (weak) solution of the boundary value problem

$$\Delta w + k_0^2(1 - m)w = k_0^2 h \cdot u_\mathrm{t} \qquad \text{in } \Omega, \tag{18}$$

$$\frac{\partial w}{\partial \mathbf{n}} - ik_0 w = 0 \qquad \text{on } \partial\Omega. \tag{19}$$

Let us for now assume that S is Gâteaux differentiable in an open neighborhood of $m \in \mathscr{D}(S)$. The Gâteaux differentiability in m yields the existence of the limit

$$\lim_{\alpha \to 0} \frac{\big(S(m + \alpha h) - S(m)\big)}{\alpha}, \qquad h \in \mathscr{D}(S).$$

The boundary value problem (18), (19) is obtained from the original scattering problem (6), (7) by considering, for $m \in \mathscr{D}(S)$, the perturbed boundary value problem

$$\Delta u_{\mathrm{sc,h}} + k_0^2\big(1-(m+\alpha h)\big)u_{\mathrm{sc,h}} = k_0^2(m+\alpha h)u_{\mathrm{i}} \qquad \text{in } \Omega,$$
$$\frac{\partial u_{\mathrm{sc,h}}}{\partial \mathbf{n}} - ik_0 u_{\mathrm{sc,h}} = 0 \qquad \text{on } \partial\Omega,$$

where $u_{\mathrm{sc,h}} := S(m+\alpha h) - u_{\mathrm{i}}$. As before, we define $u_{\mathrm{sc}} := S(m) - u_{\mathrm{i}}$ and note that both fields u_{sc} and $u_{\mathrm{sc,h}}$ fulfill the Robin boundary condition (7).

Note that $u_{\mathrm{sc}}, u_{\mathrm{sc,h}} \in H^1(\Omega)$, which follows from our analysis of the scattering map S. We subtract the Helmholtz equation for u_{sc} from the one for $u_{\mathrm{sc,h}}$ and obtain

$$\Delta\big(S(m+\alpha h) - S(m)\big) + k_0^2(1-m)\big(S(m+\alpha h) - S(m)\big) = k_0^2\big(S(m+\alpha h)\big)\alpha h.$$

We divide this expression by α and consider the weak formulation (8) of this partial differential equation,

$$\int_\Omega \nabla u_\alpha \cdot \nabla \overline{v}\,\mathrm{d}\mathbf{x} - k_0^2 \int_\Omega (1-m)u_\alpha \cdot \overline{v}\,\mathrm{d}\mathbf{x} - ik_0 \int_{\partial\Omega} u_\alpha \cdot \overline{v}\,\mathrm{d}s_{\mathbf{x}} = -k_0^2 \int_\Omega m u_{\mathrm{i}} \cdot \overline{v}\,\mathrm{d}\mathbf{x}, \tag{20}$$

where we replaced u by

$$u_\alpha := \frac{\big(S(m+\alpha h) - S(m)\big)}{\alpha}.$$

We postulated the existence of the limit $\lim_{\alpha\to 0} u_\alpha$ and assume $|\alpha| \le \overline{\alpha}$ for some $\overline{\alpha} > 0$. Note that due to our previous findings, we estimate

$$\begin{aligned}\|\nabla u_\alpha\|_{L^2(\Omega)} &= |u_\alpha|_{H^1(\Omega)} \le \|u_\alpha\|_{H^1(\Omega)} \\ &\le c(k_0, \overline{\alpha}, \Omega, \|m\|_{L^\infty(\Omega)})\|h\|_{L^\infty(\Omega)}\|u_{\mathrm{i}}\|_{L^2(\Omega)}.\end{aligned}$$

As a consequence,

$$\sup_{|\alpha|\le\overline{\alpha}} \big(\nabla u_\alpha \cdot \nabla \overline{v}\big) \in L^1(\Omega).$$

Obviously, our previous analysis also yields

$$\sup_{|\alpha|\le\overline{\alpha}} \big((1-m)u_\alpha \cdot \overline{v}\big) \in L^1(\Omega), \quad \sup_{|\alpha|\le\overline{\alpha}} \big(u_\alpha \cdot \overline{v}\big) \in L^1(\partial\Omega), \quad \sup_{|\alpha|\le\overline{\alpha}} \big(m u_{\mathrm{i}} \cdot \overline{v}\big) \in L^1(\Omega).$$

We let α tend to zero in (20). The dominated convergence theorem (see, e.g., [12]) now allows us to interchange limit and integration. Additionally, the continuity of the operator $\nabla : H^1(\Omega) \to L^2(\Omega)$, which follows from $\|\nabla u\|_{L^2(\Omega)} = |u|_{H^1(\Omega)} \le \|u\|_{H^1(\Omega)}$, allows a further interchange of limit and differentiation. We thus derived a variational formulation of

$$\Delta \lim_{\alpha\to 0} u_\alpha + k_0^2(1-m)\lim_{\alpha\to 0} u_\alpha = k_0^2 \lim_{\alpha\to 0} \big(S(m+\alpha h)\big)h$$

with Robin boundary conditions. Since S is continuous, the right-hand side converges to $k_0^2 S(m)h$ for $\alpha \to 0$. Now define

$$w := \lim_{\alpha\to 0} u_\alpha = \lim_{\alpha\to 0} \frac{\big(S(m+\alpha h) - S(m)\big)}{\alpha},$$

which satisfies the inhomogeneous Helmholtz equation (18). This yields a candidate for the Fréchet derivative of the scattering map S in m. We will further investigate the linear mapping T_m, prove that S is Fréchet differentiable and show that its Fréchet derivative coincides with the operator T_m.

Similar techniques as applied in the previous section allow us to prove the existence of a weak solution $w \in H^1(\Omega)$ of the boundary value problem (18), (19), such that T_m is well-defined on $\mathscr{D}(S)$. Also, we can deduce that T_m is bounded and therefore continuous. We will skip the proof as it is similar to the proof of Theorem 2 and the subsequent statements.

Lemma 2 *Let $m \in \mathscr{D}(S)$ and $u_\mathrm{i} \in H^1(\Omega)$ be fixed. The operator*

$$T_m : \mathscr{D}(S) \to H^1(\Omega),\ T_m h := w,$$

where w is the unique weak solution of (18), (19), is linear and bounded. For $h \in \mathscr{D}(S)$, we have

$$\|T_m h\|_{H^1(\Omega)} \le C_3 \|h\|_{L^\infty(\Omega)} \cdot \|u_\mathrm{i}\|_{L^2(\Omega)}, \tag{21}$$

where $C_3 := C_1(1 + C_1 M) > 0$ depends on k_0, Ω.

Let us now consider the mapping

$$\widetilde{T} : \mathscr{D}(S) \to L\big(\mathscr{D}(S), H^1(\Omega)\big),\ m \mapsto \widetilde{T}(m) := T_m.$$

This operator maps a bounded, compactly supported function m to the respective linear operator T_m. We can formulate a continuity result for this mapping.

Lemma 3 *Let $m_1, m_2, h \in \mathscr{D}(S)$ and $u_\mathrm{i} \in H^1(\Omega)$. The mapping $\widetilde{T}$ fulfills*

$$\big\|\widetilde{T}(m_1)h - \widetilde{T}(m_2)h\big\|_{H^1(\Omega)} \le C_4 \|m_1 - m_2\|_{L^\infty(\Omega)} \cdot \|h\|_{L^\infty(\Omega)} \cdot \|u_\mathrm{i}\|_{L^2(\Omega)}, \tag{22}$$

where $C_4 = C_4\big(k_0, \Omega\big) > 0$.

Proof For $w_j := \widetilde{T}(m_j)h = T_{m_j}h, j = 1, 2$, we have

$$\Delta w_j + k_0^2(1 - m_j)w_j = k_0^2 h \cdot S(m_j).$$

By subtracting these two equations from each other, we obtain

$$\Delta(w_1 - w_2) + k_0^2(1 - m_1)(w_1 - w_2) = k_0^2 h\big(S(m_1) - S(m_2)\big) + k_0^2(m_1 - m_2)w_2.$$

Using the previous arguments again, we obtain

$$\begin{aligned}\|w_1 - w_2\|_{H^1(\Omega)} \leq k_0 \big(&\|h\|_{L^\infty(\Omega)} \cdot \|S(m_1) - S(m_2)\|_{H^1(\Omega)} \\ &+ \|m_1 - m_2\|_{L^\infty(\Omega)} \cdot \|w_2\|_{H^1(\Omega)}\big) .\end{aligned}$$

Finally, we use Lemmas 1 and 2 to further estimate

$$\|w_1 - w_2\|_{H^1(\Omega)} \leq k_0(C_2 + C_3)\, \|h\|_{L^\infty(\Omega)} \cdot \|m_1 - m_2\|_{L^\infty(\Omega)} \cdot \|u_{\mathrm{i}}\|_{L^2(\Omega)}$$

and set $C_4 := k_0(C_2 + C_3)$. □

Theorem 3 *The operator S from Definition 2 is Fréchet differentiable with respect to $m \in \mathscr{D}(S)$. The Fréchet derivative in $m \in \mathscr{D}(S)$ is the linear operator*

$$S'(m) : \mathscr{D}(S) \to H^1(\Omega),\ h \mapsto S'(m)h = w,$$

where $w \in H^1(\Omega)$ solves the linearized boundary value problem

$$\Delta w + k_0^2(1 - m)w = k_0^2 u_{\mathrm{t}} \cdot h \qquad \text{in } \Omega, \tag{23}$$

$$\frac{\partial w}{\partial \mathbf{n}} - ik_0 w = 0 \qquad \text{on } \partial\Omega. \tag{24}$$

The function $u_{\mathrm{t}} := S(m)$ is the weak solution of the scattering problem (5)–(7).

We thus have $S'(m)h = T_m h$ for all $h \in \mathscr{D}(S)$ and $S'(m)$ is continuous.

Proof If there is a positive constant $C_5 = C_5(k_0, \Omega)$, such that the estimate

$$\|S(m + h) - S(m) - T_m h\|_{H^1(\Omega)} \leq C_5 \|h\|^2_{L^\infty(\Omega)} \cdot \|u_{\mathrm{i}}\|_{L^2(\Omega)} \tag{25}$$

holds for $m, h \in \mathscr{D}(S)$, we are done.

Define the functions

$$\begin{aligned} u_{\mathrm{sc}} &:= S(m) - u_{\mathrm{i}}, \\ u_{\mathrm{sc,h}} &:= S(m + h) - u_{\mathrm{i}}, \\ w &:= T_m h, \end{aligned}$$

which fulfill the Robin boundary condition and

$$\begin{aligned}\Delta u_{\mathrm{sc}} + k_0^2\big(1-m\big)u_{\mathrm{sc}} &= k_0^2 m u_{\mathrm{i}},\\ \Delta u_{\mathrm{sc,h}} + k_0^2\big(1-(m+h)\big)u_{\mathrm{sc,h}} &= k_0^2\big(m+h\big)u_{\mathrm{i}},\\ \Delta w + k_0^2\big(1-m\big)w &= k_0^2 h\big(u_{\mathrm{sc}}+u_{\mathrm{i}}\big).\end{aligned}$$

From these equations, we obtain for $v := u_{\mathrm{sc,h}} - u_{\mathrm{sc}} - w = S(m+h) - S(m) - T_m h$ the Helmholtz equation

$$\Delta v + k_0^2(1-m)v = k_0^2 h\big(u_{\mathrm{sc,h}} - u_{\mathrm{sc}}\big),$$

and since v satisfies the Robin boundary condition (7), we estimate

$$\begin{aligned}\|v\|_{H^1(\Omega)} &\le C_1 \, \|h\|_{L^\infty(\Omega)} \cdot \|u_{\mathrm{sc,h}} - u_{\mathrm{sc}}\|_{L^2(\Omega)}\\ &\le C_1 \, \|h\|_{L^\infty(\Omega)} \cdot \|u_{\mathrm{sc,h}} - u_{\mathrm{sc}}\|_{H^1(\Omega)}\\ &\le C_1 C_2 \, \|h\|^2_{L^\infty(\Omega)} \cdot \|u_{\mathrm{i}}\|_{L^2(\Omega)} \,,\end{aligned}$$

where we used Lemma 1. Resubstituting v again, we finally have shown

$$\|S(m+h) - S(m) - T_m h\|_{H^1(\Omega)} \le C_5 \, \|h\|^2_{L^\infty(\Omega)} \cdot \|u_{\mathrm{i}}\|$$

with $C_5 = C_5(k_0, \Omega) = C_1 C_2 > 0$. The continuity of $S'(m)$ is a direct result of the boundedness of T_m. □

Lemma 4 *For $m_1, m_2 \in \mathscr{D}(S)$, the operator S fulfills the estimate*

$$\begin{aligned}&\|S(m_1) - S(m_2) - S'(m_1)(m_1 - m_2)\|_{L^2(\Omega)}\\ &\qquad \le C_6 \cdot \|S(m_1) - S(m_2)\|_{L^2(\Omega)},\end{aligned} \tag{26}$$

where $C_6 = C_6\big(k_0, \Omega, M\big)$.

Proof Let $u_1 := S(m_1) - u_{\mathrm{i}}$, $u_2 := S(m_2) - u_{\mathrm{i}}$ and $w := S'(m_1)(m_1 - m_2)$ satisfying

$$\begin{aligned}\Delta u_1 + k_0^2(1-m_1)u_1 &= k_0^2 m_1 u_{\mathrm{i}},\\ \Delta u_2 + k_0^2(1-m_2)u_2 &= k_0^2 m_2 u_{\mathrm{i}},\\ \Delta w + k_0^2(1-m_1)w &= k_0^2(m_1 - m_2)(u_1 + u_{\mathrm{i}}).\end{aligned}$$

Additionally, u_1, u_2, w, and consequently $u_1 - u_2 - w$ obey the Robin boundary condition. In particular, we have $u_1, u_2, w \in H^1(\Omega) \subseteq L^2(\Omega)$ according to our previous results.

Subtracting the equations for u_2 and w from the one for u_1 yields

$$\Delta(u_1 - u_2 - w) + k_0^2(1 - m_1)(u_1 - u_2 - w) = k_0^2(m_1 - m_2)(u_2 - u_1).$$

As before, we now estimate

$$\begin{aligned}\|u_1 - u_2 - w\|_{L^2(\Omega)} &\le \|u_1 - u_2 - w\|_{H^1(\Omega)} \\ &\le C_1 \cdot \|m_1 - m_2\|_{L^\infty(\Omega)} \cdot \|u_1 - u_2\|_{L^2(\Omega)} \\ &\le C_1 \cdot 2M \cdot \|u_1 - u_2\|_{H^1(\Omega)},\end{aligned}$$

which is equivalent to (26) for $C_6 := 2C_1M$ due to our definitions of u_1, u_2, and w. □

Remark 5 For the constant C_6 in the estimate (26) holds

$$C_6(k_0, \Omega, M) < 1 \tag{27}$$

for M sufficiently small. Lemma 4 then states the validity of the *tangential cone condition* for S,

$$\|S(m_1) - S(m_2) - S'(m_1)(m_1 - m_2)\|_{L^2(\Omega)} \le c_{\text{tc}}\|S(m_1) - S(m_2)\|_{L^2(\Omega)} \tag{28}$$

with a constant $c_{\text{tc}} < 1$ for all $m_1, m_2 \in \mathscr{D}(S)$. Note further, that the value of C_6 in (26) depends in particular on the wave number k_0 and has to be adapted when working with different frequencies. Finally we like to emphasize that the estimate (26) is valid in $H^1(\Omega)$ as well, according the proof of Lemma 4.

2.4 The Adjoint Linearized Scattering Operator on the Boundary

We define the composition of the operators γ and $S'(m)$, $m \in \mathscr{D}(S)$, by

$$\mathscr{T}_m : \mathscr{D}(S) \to L^2(\partial\Omega), \; h \mapsto \gamma S'(m)h.$$

Let $\sigma \in L^2(\partial\Omega)$. We want to find a function $\delta m \in L^2(\Omega)$, such that

$$\mathscr{T}_m^* \sigma = \delta m. \tag{29}$$

For that purpose, we will consider the standard L^2-inner product, such that for $\eta \in L^2(\Omega)$ we have

$$(\delta m, \eta)_{L^2(\Omega)\times L^2(\Omega)} = \left(\mathscr{T}_m^*\sigma, \eta\right)_{L^2(\Omega)\times L^2(\Omega)} = (\sigma, \mathscr{T}_m\eta)_{L^2(\partial\Omega)\times L^2(\partial\Omega)} .$$

Theorem 4 *There exists a $\phi \in H^1(\Omega)$, such that*

$$\mathscr{T}_m^* \sigma = k_0^2 \cdot \overline{S(m)} \cdot \phi, \tag{30}$$

where $m \in \mathscr{D}(S)$. The function ϕ is uniquely determined as the weak solution of the adjoint problem

$$\Delta\phi + k_0^2(1-\overline{m})\phi = 0 \qquad \text{in } \Omega, \tag{31}$$

$$\frac{\partial\phi}{\partial\mathbf{n}} + ik_0\phi = -\sigma \qquad \text{on } \partial\Omega. \tag{32}$$

Proof Let $w := S'(m)h = T_m h$. Consider the inner product $(\cdot,\cdot)_{L^2(\Omega)\times L^2(\Omega)}$ of Eq. (23) with some $\phi \in H^1(\Omega)$,

$$\int_\Omega \Delta w\overline{\phi}\ \mathrm{d}\mathbf{x} + k_0^2 \int_\Omega (1-m)w\overline{\phi}\ \mathrm{d}\mathbf{x} = k_0^2 \int_\Omega h \cdot u_\mathrm{t}\overline{\phi}\ \mathrm{d}\mathbf{x}, \tag{33}$$

where $u_\mathrm{t} := S(m)$ denotes the solution of the direct scattering problem (5)–(7).

With partial integration, we obtain from the first term

$$\int_\Omega \Delta w\overline{\phi}\ \mathrm{d}\mathbf{x} = \int_{\partial\Omega} \frac{\partial w}{\partial\mathbf{n}}\overline{\phi}\ \mathrm{d}s_\mathbf{x} - \int_{\partial\Omega} w\frac{\partial\overline{\phi}}{\partial\mathbf{n}}\ \mathrm{d}s_\mathbf{x} + \int_\Omega w\Delta\overline{\phi}\ \mathrm{d}\mathbf{x}.$$

By applying the boundary condition (19) of the linearized problem, this yields

$$\int_\Omega \Delta w\overline{\phi}\ \mathrm{d}\mathbf{x} = \int_{\partial\Omega} w \cdot \overline{\left(-ik_0\phi - \frac{\partial\phi}{\partial\mathbf{n}}\right)}\ \mathrm{d}s_\mathbf{x} + \int_\Omega w\Delta\overline{\phi}\ \mathrm{d}\mathbf{x}.$$

The second term of (33) is rewritten as

$$k_0^2 \int_\Omega (1-m)w\overline{\phi}\ \mathrm{d}\mathbf{x} = \int_\Omega w\overline{\big(k_0^2(1-\overline{m})\phi\big)}\ \mathrm{d}\mathbf{x}.$$

By setting $\mathscr{T}_m^*\sigma = k_0^2\overline{u_\mathrm{t}}\phi$, the right-hand side of (33) yields

$$k_0^2 \int_\Omega h \cdot u_\mathrm{t}\overline{\phi}\ \mathrm{d}\mathbf{x} = (\mathscr{T}_m h, \sigma)_{L^2(\partial\Omega)\times L^2(\partial\Omega)} = \int_{\partial\Omega} w\overline{\sigma}\ \mathrm{d}s_\mathbf{x}.$$

Summarizing the above results we obtain

$$\int_\Omega w \cdot \overline{\big(\Delta\phi + k_0^2(1-\overline{m})\phi\big)}\ \mathrm{d}\mathbf{x} = 0$$

and

$$\int_{\partial\Omega} w \cdot \overline{\left(-ik_0\phi - \frac{\partial\phi}{\partial\mathbf{n}} - \sigma\right)}\, \mathrm{d}s_{\mathbf{x}} = 0.$$

As the weak solution w of the linearized scattering problem does generally not vanish on either Ω or its boundary $\partial\Omega$, ϕ fulfills the postulated boundary value problem from Theorem 4. The existence and uniqueness of a weak solution ϕ of (31), (32) can be derived by similar calculations as in the preceding section. □

2.5 *The Observation Operator*

For our numerical reconstructions, we have to specify the observation operator and find an expression for its adjoint. We assume there are $N \in \mathbb{N}$ receivers that measure the total electric field on the boundary of Ω. In order to increase the amount of data points, the tomograph is rotated around the test object in $J \in \mathbb{N}$ equidistant steps $\theta_j = (j-1)\frac{2\pi}{J}, j = 1, \dots, J$. The emitter also serves as a receiver. However, the incident field u_{i}, and correspondingly also the scattered field u_{sc}, the total field u_{t}, the scattering operator S, and the observation operator Q depend on the position j of the tomograph. We indicate this dependence by an additional index j.

The receivers' surfaces are denoted by E_ν^j for $\nu = 1, \dots, N$. A sketch of the setup is given in Fig. 1.

Definition 4 For each $j = 1, \dots, J$, the observation operator Q^j maps the (restricted) electric field $\gamma u^j \in L^2(\partial\Omega)$ to the measured values $y^j = \left(y_1^j, \dots, y_N^j\right)^T \in \mathbb{C}^N$ by $y^j = Q^j(\gamma u^j)$, where

$$\begin{aligned} Q^j : L^2(\partial\Omega) &\to \mathbb{C}^N, \\ \varphi &\mapsto \left(\int_{\partial\Omega^j} e_\nu^j(\mathbf{x})\varphi(\mathbf{x})\, \mathrm{d}s_{\mathbf{x}}\right)_{\nu=1,\dots,N}. \end{aligned} \tag{34}$$

The functions $e_\nu^j : \partial\Omega \to [0,\infty)$, $\nu = 1, \dots, N$, are called the *sensor characteristics*.

Obviously, the observation operator Q^j is linear and bounded. The adjoint $(Q^j)^*$ of Q^j is given by $(Q^j)^* : \mathbb{C}^N \to L^2(\partial\Omega)$, where

$$\beta = (\beta_\nu)_{\nu=1,\dots,N} \mapsto (Q^j)^*\beta = \sum_{\nu=1}^{N} \beta_\nu e_\nu^j. \tag{35}$$

Together with our previous results, we have derived a complete representation of the forward operator $F^j = Q^j\gamma S^j$, its Fréchet derivative $(F^j)'(m) = Q^j\gamma(S^j)'(m)$

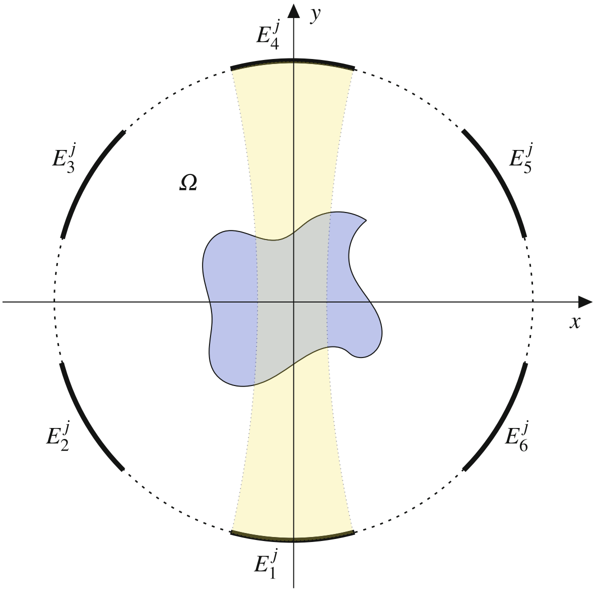


Fig. 1 Sketch of a THz tomograph with $N = 6$ receivers E_ν^j, $\nu = 1, \ldots, 6$, where the receiver E_1^j also serves as an emitter

in $m \in \mathscr{D}(S)$ and the adjoint of $(F^j)'(m)$. In summary, the inverse problem of THz tomography is formulated as the collection of operator equations

$$F^j(m) = y^j, \qquad j = 1, \ldots, J,$$

and consists in the reconstruction of m from (noisy) data $y = (y_\nu^j)_{j=1,\ldots,J,\nu=1,\ldots,N}$.

3 Numerical Reconstruction of the Complex Refractive Index from Simulated Data

This section deals with numerical reconstructions of tomographic THz data, where we use iterative solvers as the Landweber method and sequential subspace techniques (RESESOP). Nonlinear inverse problems (1) are usually solved iteratively, and for both methods it is essential to have gradients

$$g(m) := \big(F'(m)\big)^* \big(F(m) - y^\delta\big)$$

of the functional $\frac{1}{2}\|F(x) - y^\delta\|^2$ available. Due to our previous results, we are able to determine the gradient $g^j(m) = \big((F^j)'(m)\big)^* \big(F^j(m) - y^{j,\delta}\big)$ in $m \in \mathscr{D}(S)$ for each position $j = 1, \dots, J$. However, since we have to evaluate two boundary value problems in order to calculate $g^j(m)$, it is desirable to reduce the number of iterations until the stopping criterion is fulfilled. The (regularizing) sequential subspace optimization method ((RE)SESOP) has been developed for this purpose [21]. In the subsequent sections, we first give an overview of the applied reconstruction techniques, before we present some numerical reconstructions of m from synthetic noisy data $y^\delta \in \mathbb{C}^{N \times J}$.

3.1 Reconstruction Techniques

The RESESOP method, which is used to solve the inverse problem of THz tomography, is a slight variation of the RESESOP algorithm with two search directions as presented in [21]. Generally, the methods discussed in [21] are suited to solve nonlinear inverse problems in real Hilbert spaces. In the case of THz tomography, we are dealing with an inverse problem in complex Hilbert spaces. We give a short introduction to sequential subspace optimization and derive a method that meets the requirements of complex Hilbert spaces.

The basic idea of sequential subspace optimization is to reduce the number of iterations by projecting sequentially onto suitable subsets of the source space in order to find an approximate solution of a nonlinear operator equation

$$F(x) = y.$$

In each step $n \in \mathbb{N}$, we choose a finite index set $I_n^\delta \subseteq \mathbb{N}$. The subsets are intersections of stripes

$$H_{n,l}^\delta := H\left(u_{n,l}^\delta, \alpha_{n,l}^\delta, \xi_{n,l}^\delta\right), \quad l \in I_n^\delta,$$

where $u_{n,l}^\delta$ is the normal vector of the bounding affine hyperplanes, $\alpha_{n,l}^\delta$ is the offset, and $\xi_{n,l}^\delta \geq 0$ determines the width of the stripe. A *stripe* in a real Hilbert space X is defined as

$$H(u, \alpha, \xi) := \left\{x \in X : \left| \langle u, x \rangle - \alpha \right| \leq \xi \right\}.$$

It is essential that each of these stripes contains the solution set of the unperturbed operator equation. The shape, i.e., the width and the normal vector, are chosen such that the nonlinear character and the noise level δ are taken into account, guaranteeing a descent property of the form

$$\left\| z - x_{n+1}^\delta \right\|^2 \leq \left\| z - x_n^\delta \right\|^2 - C_n,$$

where $\{x_n^\delta\}$ is the sequence of iterates and $C_n > 0$. This is realized by an iteration of the form

$$x_{n+1}^\delta = x_n^\delta - \sum_{l \in I_n^\delta} t_{n,l}^\delta \cdot u_{n,l}^\delta, \tag{36}$$

and the optimization parameters $t_{n,i}^\delta$ are calculated such that the current iterate x_n^δ is projected onto the intersection of the respective stripes $H_{n,l}^\delta$, $l \in I_n^\delta$.

For our reconstruction, we specify the index set as $I_n^\delta := \{n-1, n\}$ and the search directions $u_{n,l}^\delta := g_l^\delta$ for all $n \in \mathbb{N}$. The result is a *fast regularizing method with two search directions* per iteration. For the stripes $H_{n,l}^\delta$, where $l = n-1, n$, we choose the offset

$$\alpha_{n,l}^\delta := \langle u_{n,l}^\delta, x_l^\delta \rangle$$

and the width

$$\xi_{n,l}^\delta := \|R_l^\delta\| \cdot \big(\delta + c_{\text{tc}}\big(\|R_l^\delta\| + \delta\big)\big),$$

where δ is the noise level and c_{tc} is the constant from the tangential cone condition (28) with S replaced by F. The norm of the residual

$$R_n^\delta := F(x_n^\delta) - y^\delta$$

not only occurs as a factor in the width of the stripes, but is also needed for the discrepancy principle.

In the following, we will take a closer look at our inverse problem of THz tomography,

$$F^j(m) = y^{j,\delta}, \quad j = 1, \dots, J.$$

Note, that $F^j : L^2(\Omega) \to \mathbb{C}^N$ is a nonlinear operator between complex Hilbert spaces.

Remark 6 The search directions we use in our RESESOP algorithm (36) are *averaged gradients*

$$g_n^\delta := \frac{1}{J} \sum_{j=1}^{J} g_n^{j,\delta} = \frac{1}{J} \sum_{j=1}^{J} \left(F^j\right)'\left(m_n^\delta\right)^* \left(F^j\left(m_n^\delta\right) - y^{j,\delta}\right). \tag{37}$$

Before we are able to apply the RESESOP algorithm, a small and straightforward adaption is necessary: The source space is a complex Hilbert space, whereas the algorithm requires it to be a real Hilbert space. By splitting up the gradient into real and imaginary part, we obtain separate real-valued search directions. The new iterate is then obtained by a separate subspace optimization for real and imaginary part. This means that the full forward model is used to calculate the search direction $F'(m_n)^*(F(m_n)-y^\delta)$, whereas the step width is determined individually for real and imaginary part:

$$\begin{aligned} m_{n+1}^\delta &= \mathrm{Re}(m_{n+1}^\delta) + i\cdot \mathrm{Im}(m_{n+1}^\delta) \\ &= \left(\mathrm{Re}(m_n^\delta) - \sum_{l\in I_n^\delta} t_{n,l}^{\delta,\mathrm{r}}\cdot \mathrm{Re}(g_l^\delta)\right) + i\cdot \left(\mathrm{Im}(m_n^\delta) - \sum_{l\in I_n^\delta} t_{n,l}^{\delta,\mathrm{i}}\cdot \mathrm{Im}(g_l^\delta)\right). \end{aligned}$$

The optimization parameter $t_{n,l}^{\delta,\mathrm{r}}$ is calculated such that $\mathrm{Re}(m_n^\delta)$ is projected onto the intersection of the stripes $H_{n,l}^{\mathrm{r},\delta}$, $l = n-1, n$, with width

$$\alpha_{n,l}^{\delta,\mathrm{r}} = \left\langle \mathrm{Re}(g_l^\delta), \mathrm{Re}(m_n^\delta)\right\rangle$$

and offset

$$\xi_l^\delta = \|R_l^\delta\|\cdot\big(\delta + c_{\mathrm{tc}}\big(\|R_l^\delta\| + \delta\big)\big).$$

The stripes $H_{n,l}^{\mathrm{i},\delta}$ are defined analogously. Note that the width of the stripes must not be adapted, as it is influenced by the norm of the residual R_l^δ, which is determined by the full forward problem. We set

$$\|R_n^\delta\| := \max_{j=1,\ldots,J} \left\|F^j(m_n^\delta) - y^{j,\delta}\right\|.$$

The iteration is stopped by the discrepancy principle after step n_* satisfying

$$\max_{j=1,\ldots,J} \left\|F^j(m_{n_*}^\delta) - y^{j,\delta}\right\| \leq \tau\delta < \max_{j=1,\ldots,J} \left\|F^j(m_n^\delta) - y^{j,\delta}\right\|$$

for all $n < n_*$.

3.2 Numerical Experiments

We start with an example of a reconstruction from simulated data that were generated in a test with radiation of the frequency $f = 0.1$ THz. The second example serves as a performance test for the two techniques.

3.2.1 An Example in the THz Regime

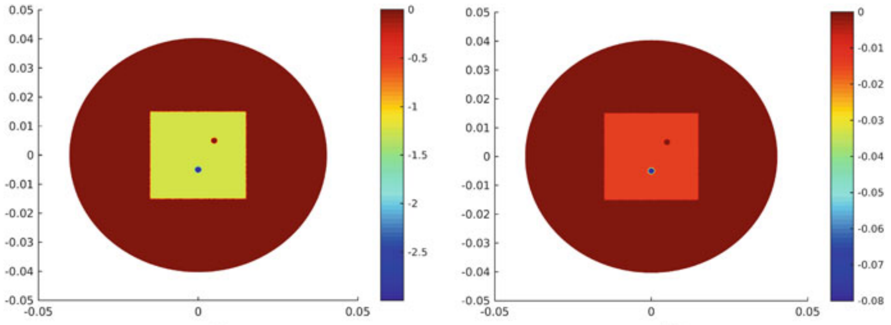
For our first numerical experiment, we choose a plastic block with a complex refractive index $m_1 := 1 - (1.5 + i \cdot 0.005)^2 = -1.249975 - i \cdot 0.015$ as a test object and assume the outer interfaces to be known. Inside the object, there are two inclusions. One is a hole filled with air, which thus has the complex refractive index $m_2 = 0$, the other one consists of an optically denser material with complex refractive index $m_3 = 1 - (1.8 + i \cdot 0.02)^2 = -2.2396 - i \cdot 0.072$. There is no a priori information about the inclusion or any material parameter. Note that the size of the inclusions is of the order of the wave length of the THz beam. The domain in which we reconstruct m is set to

$$\Omega = \left\{ \mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2^2 \leq (0.04\mathrm{m})^2 \right\} .$$

The exact real and imaginary part are displayed in Fig. 2.

For the generation of (noisy) data we use the parameters in Table 1.

We set the starting value to $m_0(\mathbf{x}) = (-1 - i \cdot 0.001) \cdot \chi_D(\mathbf{x})$, where the support of the (intact) block is denoted by $D = \{\mathbf{x} \in \Omega : -0.0015\mathrm{m} \leq x, y \leq 0.0015\mathrm{m}\}$ and χ_D is the characteristic function of D.



(a) (b)

Fig. 2 Real (**a**) and imaginary (**b**) part of m

Table 1 Parameters of the numerical RESESOP experiment

Parameter	Value
Frequency f	$0.1 \cdot 10^{12}$ Hz
Wave length λ	$2.998 \cdot 10^{-3}$ m
Beam waist W_0	0.015 m
Rayleigh zone y_0	0.02 m
Number of receivers N	40
Number of positions J	180
c_{tc}	0.9
τ	20

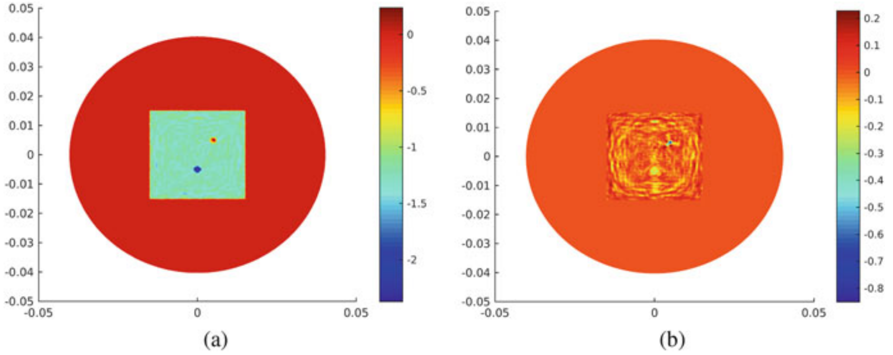


Fig. 3 Reconstruction $m_{n_*}^{\delta}$ of real (**a**) and imaginary (**b**) part of m after $n_* = 30$ iterations with the RESESOP method (using two search directions). The object was scanned using a Gaussian beam of frequency $f = 0.1$ THz

The occurring boundary value problems are solved numerically with a Matlab solver, which uses the finite element method with linear basis functions. The maximal size of the finite elements should not exceed a tenth of the wave length. Since the numerical effort to solve one boundary value problem grows quadratically with the wave number, these evaluations are expensive.

The implementation of the observation operator, however, leads to an unavoidable error in the data, which we take into account in an increased noise level. Additionally, we add 2% uniformly distributed noise to the data.

As the plots in Fig. 3 indicate, the real part of the complex refractive index is reconstructed satisfactorily (with a relative error of 7.98%), allowing quantitative and qualitative conclusions, also on location, shape, and size of the inclusions. However, the reconstructed imaginary part does not yield any information on the value of $\mathrm{Im}(m)$, but admits a localization of the inclusions. A possible reason might be the model itself, which is based on several assumptions that hold only approximately (such as the idealized Robin boundary conditions).

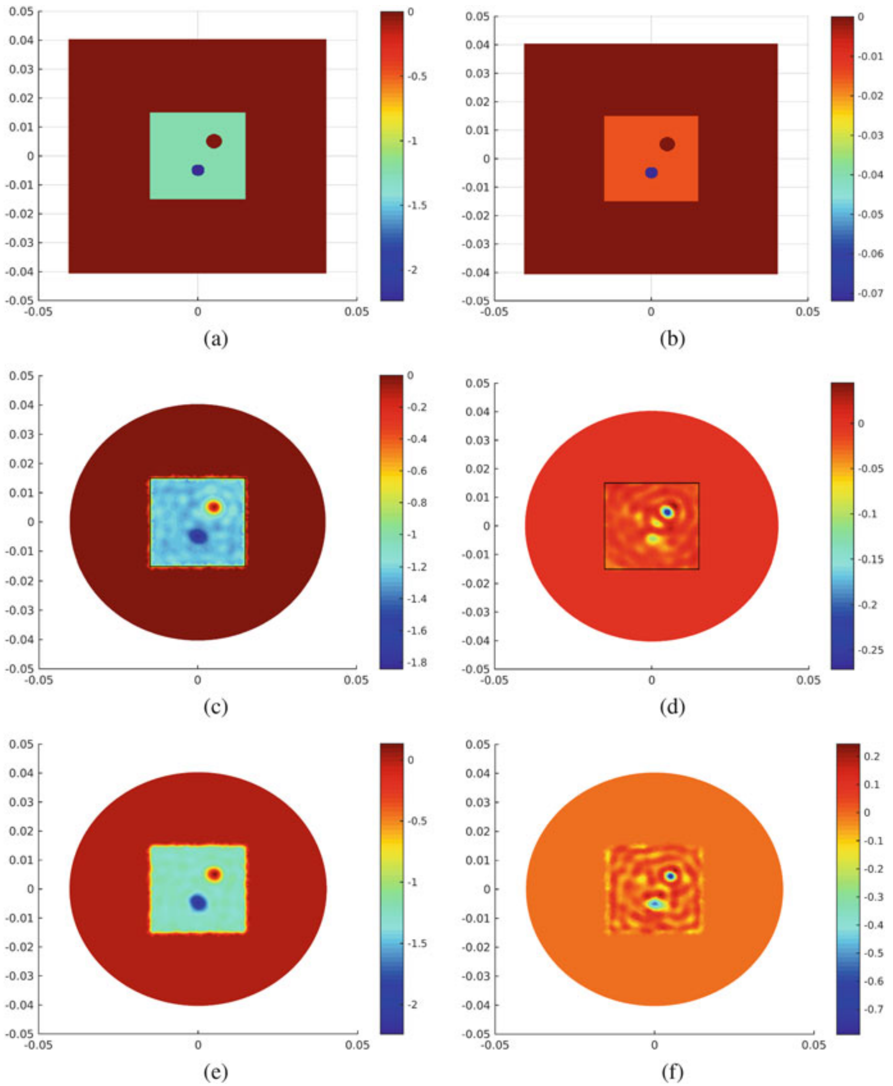
Remark 7 There are a lot of parameters occurring in the model, and also in the RESESOP algorithm, that need to be chosen by trial and error. This leaves some room for further improvement. Due to the numerically expensive evaluations of the boundary value problems, a thorough testing of the parameters is time-consuming and not very economical. This underlines the necessity of faster reconstruction methods such as the RESESOP method.

3.2.2 A Comparison of Landweber and RESESOP Method

In an additional experiment, we have compared the performance of our RESESOP method with the standard Landweber method

$$m_{n+1}^{\delta} = m_n^{\delta} - \omega g_n^{\delta}(m_n^{\delta}), \quad n \in \mathbb{N},$$

with a fixed relaxation parameter ω (see, e.g., [13]). Both iterations were stopped by the discrepancy principle. We used microwave radiation with a frequency of $2.5 \cdot 10^{10}$ Hz in order to reduce the computational effort. We tested an object as in the previous experiment, but with larger inclusions (due to the lower frequency). The object's complex refractive index is plotted in Fig. 4, along with the respective reconstructions.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4 Real (**a**) and imaginary (**b**) part of *m* and the respective reconstructions with the Landweber method ((**c**) and (**d**)) and with the RESESOP method ((**e**) and (**f**)), respectively

Table 2 Performance of Landweber and RESESOP method at the reconstruction of the complex refractive index

	Landweber	RESESOP
Number of iterations n_*	155	20
Execution time	6 h 38 min 49 s	1 h 12 min 40 s

The synthetic noisy data was generated for $N = 20$ receivers in $J = 180$ positions. Table 2 shows the performance of the two methods, clearly stating the reduced run-time of the RESESOP algorithm. We also see that the execution of a single step in the RESESOP method is a little slower than in the Landweber iteration, such that the gain is due to the reduction of iterations steps. The RESESOP method is thus especially interesting, when the calculation of the gradient is particularly expensive.

Remark 8 We want to state some concluding remarks.

- We have to make a good guess for the constant c_{tc} from the tangential cone condition (28) in the RESESOP method since (28) has not been proven for F^j. Experiments indicate that the choice $c_{\mathrm{tc}} = 0.6$ is fine for $f = 2.5 \cdot 10^{10}$ Hz and $c_{\mathrm{tc}} = 0.9$ for $f = 0.1$ THz.
- The value of c_{tc} and the consequences for the reconstruction represent the nonlinearity of the forward operator: The RESESOP algorithm requires $\tau > (1 + c_{\mathrm{tc}}) \cdot (1 - c_{\mathrm{tc}})^{-1}$ for the parameter τ from the discrepancy principle, such that a large value of c_{tc} involves a large value of τ, and the stopping criterion tends to be fulfilled earlier than for smaller values of τ. This is not surprising, since c_{tc} can be interpreted as an indicator for the nonlinearity of the forward operator F.

4 Conclusion and Future Work

We outlined a physical description of the forward model of THz tomography and analyzed in particular the properties of the underlying scattering map S. We have shown that there is a unique (weak) solution to the variational scattering problem. Additionally, we have proved the Fréchet differentiability of S, analyzed the Fréchet derivative $S'(m)$ of S in m, and calculated the adjoint of the composition $\gamma S'(m)$. Finally, we included a linear observation operator and obtained a mathematical description of the forward operator, on which the inverse problem of THz tomography relies in practical applications, that means, if the measurement geometry is taken into account. In view of numerical experiments, we derived all necessary tools to calculate the current gradient of the least squares functional $\frac{1}{2}\left\|F(m) - y^{\delta}\right\|^{2}$, which is essential for the RESESOP algorithm and also for the Landweber iteration.

Further research includes a more general framework for inverse problems based on the Helmholtz equation, i.e., a generalization of the results presented in this paper towards a more general parameter m. Additionally, an extension of the subspace optimization methods to solving nonlinear inverse problems in Banach spaces is a natural continuation of our work, with benefits also in the field of THz tomography. Finally, a combination of our methods and the adapted ART presented in [20] may yield a hybrid reconstruction algorithm which produces more precise results, especially for the imaginary part of the complex refractive index. Our numerical evaluations demonstrate that from the refractive index the singular support of the object can be well detected. This a priori information is essential for the ART method developed in [20]. To this end, the observation operator has to be adapted to the actual measuring process.

Acknowledgements The authors are indebted to Heiko Hoffmann for valuable discussions about the proof of Theorem 2.

References

1. W. Arendt, K. Urban, *Partielle Differenzialgleichungen* (Spektrum Akademischer Verlag, Berlin, 2010)
2. K. Atkinson, W. Han, *Theoretical Numerical Analysis* (Springer, New York, 2001)
3. G. Bao, P. Li, Inverse medium scattering for the Helmholtz equation at fixed frequency. Inverse Probl. **21**(5), 16–21 (2005)
4. G. Bao, P. Li, Inverse medium scattering problems in near-field optics. J. Comput. Math. **25**(3), 252–265 (2007)
5. H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations* (Springer Science+Business Media, New York, 2011)
6. W.L. Chan, J. Deibel, D.M. Mittleman, Imaging with terahertz radiation. Rep. Prog. Phys. **70**(8), 1325–1379 (2007)
7. D. Colton, R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory* (Springer, New York, 2013)
8. L.C. Evans, *Partial Differential Equations*. Graduate Studies in Mathematics (American Mathematical Society, Providence, 1998)
9. B. Ferguson, X.-C. Zhang, Materials for terahertz science and technology. Nat. Mater. **1**(1), 26–33 (2002)
10. D. Gilbarg, N. Trudinger, *Elliptic Partial Differential Equations of Second Order* (Springer, Berlin, 2001)
11. J.P. Guillet, B. Recur, L. Frederique, B. Bousquet, L. Canioni, I. Manek-Hönninger, P. Desbarats, P. Mounaix, Review of terahertz tomography techniques. J. Infrared Millim. Terahertz Waves **35**(4), 382–411 (2014)
12. P.R. Halmos, *Measure Theory* (Springer, Berlin, 2013)
13. M. Hanke, A. Neubauer, O. Scherzer, A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. Numer. Math. **72**(1), 21–37 (1995)
14. G. Narkiss, M. Zibulevsky, Sequential subspace optimization method for large-scale unconstrained optimization. Technical report, Technion - The Israel Institute of Technology, Department of Electrical Engineering, 2005
15. F. Natterer, *The Mathematics of Computerized Tomography* (Vieweg+Teubner Verlag, Berlin, 1986)
16. S. Sauter, C. Schwab, *Boundary Element Methods* (Springer, Berlin, 2011)
17. F. Scheck. *Theoretische Physik 3* (Springer, Berlin, 2010)

18. F. Schöpfer, T. Schuster, Fast regularizing sequential subspace optimization in Banach spaces. Inverse Probl. **25**(1), 015013 (2009)
19. F. Schöpfer, A.K. Louis, T. Schuster, Metric and Bregman projections onto affine subspaces and their computation via sequential subspace optimization methods. J. Inverse Ill-Posed Probl. **16**(5), 479–206 (2008)
20. J. Tepe, T. Schuster, B. Littau, A modified algebraic reconstruction technique taking refraction into account with an application in terahertz tomography. Inverse Probl. Sci. Eng. **25**, 1448–1473 (2016)
21. A. Wald, T. Schuster, Sequential subspace optimization for nonlinear inverse problems. J. Inverse Ill-posed Probl. **25**(4), 99–117 (2016)
22. D. Werner, *Funktionalanalysis* (Springer, New York, 2011)

Adaptivity and Oracle Inequalities in Linear Statistical Inverse Problems: A (Numerical) Survey

Frank Werner

Abstract We investigate a posteriori parameter choice methods for filter based regularizations $\hat{f}_\alpha = q_\alpha (T^*T) T^*Y$ in statistical inverse problems $Y = Tf + \sigma\xi$. Here we assume that T is a bounded linear operator between Hilbert spaces, and ξ is Gaussian white noise.

We discuss order optimality of a posteriori parameter choice rules by means of oracle inequalities and review known results for the discrepancy principle, the unbiased risk estimation procedure, the Lepskiĭ-type balancing principle, and the quasi-optimality principle.

The main emphasis of this paper is on numerical comparisons of the mentioned parameter choice rules. We investigate estimation of the second derivative as a mildly ill-posed example, and furthermore satellite gradiometry and the backwards heat equation as severely ill-posed examples. The performance is illustrated by means of empirical convergence rates and inefficiency compared to the best possible (oracle) choice.

1 Introduction

Many practical problems ranging from astrophysics to cell biology can be described by models of the form

$$Y = Tf + \sigma\xi. \tag{1}$$

Here $f \in \mathscr{X}$ is the unknown quantity of interest, $T : \mathscr{X} \to \mathscr{Y}$ is a bounded linear operator between Hilbert spaces $\mathscr{X}$ and $\mathscr{Y}$, $\sigma > 0$ denotes the noise level and ξ is the observation noise. The aim is to estimate f from the measurements Y. The difficulty of this problem crucially depends on the behavior of T, and unfortunately

F. Werner (✉)
Statistical Inverse Problems in Biophysics Group, Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, Göttingen, Germany
e-mail: Frank.Werner@mpibpc.mpg.de

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_15

in most applications T is not continuously invertible (e.g. if T is compact). In this case, no direct inversion is possible and regularization is needed.

Here we focus on purely random noise ξ, more specifically we will assume that ξ is a Gaussian white noise on $\mathscr{Y}$. In this setup, the model (1) has to be understood in a weak sense as $\xi \notin \mathscr{Y}$ with probability 1, this is for each $y \in \mathscr{Y}$ we have access to observations of the form

$$\langle Y, y \rangle = \langle Tf, y \rangle_{\mathscr{Y}} + \sigma \langle \xi, y \rangle$$

where $\langle \xi, y \rangle \sim \mathscr{N}\left(0, \|y\|_{\mathscr{Y}}^2\right)$ and $\mathbb{E}\left[\langle \xi, y_1 \rangle \langle \xi, y_2 \rangle\right] = \langle y_1, y_2 \rangle$ for all $y_1, y_2 \in \mathscr{Y}$. Gaussian white noise can be considered as a prototype in many applications, and hence such models have been studied extensively in the literature, see e.g. [9, 16, 32, 33, 43, 51, 64].

In this paper we focus on *filter-based* regularization methods, i.e. given a family $q_\alpha : [0, \|T^*T\|] \to \mathbb{R}$, $\alpha \in \mathscr{A} \subset \mathbb{R}_+$ of functions we consider estimators of the form

$$\hat{f}_\alpha = q_\alpha\left(T^*T\right) T^*Y. \tag{2}$$

To ensure that $\hat{f}_\alpha$ is well-defined, we will always assume that T is a Hilbert–Schmidt operator, i.e. the squares of its singular values are summable. In this case we can identify T^*Y with a Hilbert space valued random variable in $\mathscr{X}$. Estimators of the form (2) cover many popular methods frequently applied in practice, i.e. spectral cut-off and Tikhonov regularization, and are well-known from the literature (see [22] and the references therein).

Given a family $\left\{\hat{f}_\alpha\right\}_{\alpha>0}$ of estimators as in (2), the remaining question is how to choose the parameter $\alpha > 0$. A common measure for the performance of some estimator $\hat{f}_\alpha$ (or in our situation for the parameter α) is the mean square error (MSE) $\mathbb{E}\left[\|\hat{f}_\alpha - f\|_{\mathscr{X}}^2\right]$. Typically, the optimal value

$$\alpha_{\mathrm{or}} := \underset{\alpha>0}{\operatorname{argmin}}\, \mathbb{E}\left[\|\hat{f}_\alpha - f\|_{\mathscr{X}}^2\right]$$

will depend not only on Y and σ, but also on the unknown truth $f \in \mathscr{X}$. Here we focus on *a posteriori* parameter choice methods, this is choices $\bar{\alpha}$ depending only on the data Y and the noise level σ which automatically *adapt* to all possible $f \in \mathscr{W}$ where $\mathscr{W} \subset \mathscr{X}$ is some smoothness class of solutions, e.g. a Sobolev or Besov ball. The ultimate aim is to find adaptive parameter choice methods $\bar{\alpha}$ for which the MSE evaluated at $\alpha = \bar{\alpha}(Y, \sigma)$ decays of the same order (or at least almost the same order up to some logarithmic factors) as $\mathbb{E}\left[\|\hat{f}_{\alpha_{\mathrm{or}}} - f\|_{\mathscr{X}}^2\right]$ as $\sigma \searrow 0$. One important tool to obtain such *order optimality* results are so-called *oracle inequalities*, which relate the MSE evaluated at $\bar{\alpha}$ with the MSE evaluated at α_{or}.

In this study we will try to provide a (limited) survey over some common a posteriori parameter choice methods, related oracle inequalities, and theoretical results on order optimality. Furthermore, we will compare the investigated methods in numerical simulations with a focus on severely ill-posed operators, i.e. situations when the singular values of T decay exponentially fast.

The rest of the paper is organized as follows: In Sect. 2 we introduce the notion of a filter, give some examples, and present main facts about the error analysis of estimators of the form (2). Furthermore we also give a brief introduction to oracle inequalities. Section 3 is then devoted to a posteriori parameter choice rules, focusing on four prominent examples. In Sect. 4 we present a numerical comparison of these strategies for different filter-based regularization methods. Afterwards, some conclusions are drawn in Sect. 5.

2 Filter-Based Regularization

In this section we will recall basic facts about filter-based regularization. In what follows, suppose that $\mathscr{A} \subset \mathbb{R}_+$ is a bounded set with accumulation point 0. A family of functions $q_\alpha : [0, \|T^*T\|] \to \mathbb{R}$ is called a *filter* if there exist constants $C_q', C_q{\prime\prime} > 0$ such that for every $\alpha \in \mathscr{A}$ and every $\lambda \in [0, \|T^*T\|]$ it holds

$$\alpha \, |q_\alpha(\lambda)| \le C_q', \tag{3a}$$

$$\lambda \, |q_\alpha(\lambda)| \le C_q''. \tag{3b}$$

Let $\{\sigma_k, u_k, v_k\}_{k\in\mathbb{N}}$ be a singular value decomposition (SVD) of T. Then the model (1) is equivalent to the Gaussian sequence model

$$Y_k = \sigma_k f_k + \sigma \xi_k, \qquad k \in \mathbb{N}$$

where $Y_k := \langle Y, v_k\rangle, f_k := \langle f, u_k\rangle$ and $\xi_k := \langle \xi, v_k\rangle \overset{\text{i.i.d.}}{\sim} \mathscr{N}(0,1)$. Therefore, the estimation of f from Y is equivalent to estimate the coefficients f_k from Y_k, and the most simple (and unbiased) estimator therefore is given by

$$\hat{f}_k = \sigma_k^{-1} \hat{Y}_k. \tag{4}$$

Unfortunately, in the ill-posed situation one has $\sigma_k \searrow 0$ as $k \to \infty$, and hence this estimator is highly sensitive to the noise ξ_k. The sensitivity is decoded in the rate of decay of σ_k, and the operator T is called *mildly ill-posed*, if σ_k decays polynomially, and *severely ill-posed*, if the decay is at least exponential.

With the help of the SVD, the estimator $\hat{f}_\alpha$ in (2) can be written as

$$\hat{f}_\alpha = \sum_{k=1}^{\infty} \sigma_k q_\alpha \left(\sigma_k^2\right) Y_k u_k.$$

Compared to (4), the unstable inversion σ_k^{-1} is replaced by a function $\sigma_k q_\alpha\left(\sigma_k^2\right)$. More precisely, if $q_\alpha(\lambda) = \lambda^{-1}$, then (2) and (4) coincide. Condition (3a) ensures boundedness of q_α (i.e. stability of $\hat{f}_\alpha$ w.r.t. noise), whereas (3b) restricts q_α to functions which are sufficiently close to $(\cdot)^{-1}$.

There exists a vast variety of filters satisfying (3), and many common regularization methods can be written as (2) with suitable q_α. Here we will focus on the following three:

- *Spectral cut-off regularization:* If $q_\alpha(\lambda) = \frac{1}{\lambda}\mathbf{1}_{[\alpha,\infty)}(\lambda)$, then (2) is known as cut-off estimator. The condition (3) is satisfied with $C_q' = C_q'' = 1$.
- *Tikhonov regularization:* If $q_\alpha(\lambda) = \frac{1}{\lambda+\alpha}$, then (2) is known as Tikhonov regularization or ridge regression. The condition (3) is satisfied with $C_q' = C_q'' = 1$.
- *Showalter's method:* If $q_\alpha(\lambda) = \frac{1-\exp\left(-\frac{\lambda}{\alpha}\right)}{\lambda}$, the estimator (2) arises from Showalter's method. Again, (3) is satisfied with $C_q' = C_q'' = 1$.

For a further discussion of these and other filter-based methods we refer to the monograph [22]. Let us briefly comment on the reasons to focus on the above three methods: From an analytical point of view, spectral cut-off regularization can be seen as an optimal regularization method, as in many situations the corresponding estimators turn out to be (order) optimal over a wide range of smoothness classes $\mathscr{W} \subset \mathscr{X}$. However, the implementation of (2) in spectral cut-off regularization requires the SVD of T, which might be unknown analytically and difficult to compute in practice. Therefore we also consider Tikhonov regularization, which can be implemented directly by $\hat{f}_\alpha = (T^*T + \alpha I)^{-1} T^*Y$, i.e. the SVD does not have to be known. Unfortunately, Tikhonov regularization suffers from so-called *saturation*, which means that the corresponding estimator can only be minimax if the class $\mathscr{W} \subset \mathscr{X}$ is not smoother than the range of T^*T (see the discussion after (7) below). Consequently, we also consider Showalter's method, which does not suffer from saturation, but can still be implemented avoiding an SVD by employing Runge–Kutta schemes, cf. [22, Ex. 4.7] or [57].

2.1 Error Analysis

It can be seen by straight-forward computations, that the risk of $\hat{f}_\alpha$ satisfies the *error decomposition*

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f}_\alpha - f\right\|_{\mathscr{X}}^2\right] &= \left\|\mathbb{E}\left[\hat{f}_\alpha\right] - f\right\|_{\mathscr{X}}^2 + \mathbb{E}\left[\left\|\hat{f}_\alpha - \mathbb{E}\left[\hat{f}_\alpha\right]\right\|_{\mathscr{X}}^2\right] \\ &= \left\|r_\alpha\left(T^*T\right)f\right\|_{\mathscr{X}}^2 + \sigma^2\mathbb{E}\left[\left\|q_\alpha\left(T^*T\right)T^*\xi\right\|_{\mathscr{X}}^2\right] \qquad (5)\end{aligned}$$

where we abbreviated $r_\alpha(\lambda) := 1-\lambda q_\alpha(\lambda)$. The first term $\|r_\alpha(T^*T)f\|_{\mathscr{X}}$ is purely deterministic and called the *bias* or *approximation error*, as it should tend to 0 as $\alpha \searrow 0$. It is in fact caused by approximating σ_k^{-1} by $\sigma_k q_\alpha\left(\sigma_k^2\right)$. The second term $\sigma^2\mathbb{E}\left[\|q_\alpha(T^*T)T^*\xi\|_{\mathscr{X}}^2\right]$ is called the *variance* or *propagated data noise error*, and this term will typically diverge as $\alpha \searrow 0$. Therefore, the optimal α should perform a trade-off between both error contributions.

Due to the classical result by Schock [58], the rate of convergence of $\hat{f}_\alpha$ towards the true solution f will be arbitrarily slow in general. Here we will follow the common paradigm to assume that f satisfies a *spectral source condition* of the form

$$f = \varphi\left(T^*T\right)\omega, \qquad \|\omega\|_{\mathscr{X}} \le \rho, \tag{6}$$

where $\varphi : [0, \|T^*T\|] \to \mathbb{R}$ is a so-called *index function*, i.e. $\varphi(0) = 0$, φ is continuous and strictly increasing. For any underlying truth f there exists a function φ such that (6) is fulfilled (cf. [50]).

To ensure that the regularization scheme (2) can take advantage of (6) we have to assume that the function φ in (6) is a *qualification* of the filter q_α, this is

$$\sup_{\lambda\in[0,\|T^*T\|]} \varphi(\lambda)\,|r_\alpha(\lambda)| \le C_\varphi\varphi(\alpha) \qquad \text{for all } \alpha \in \mathscr{A} \tag{7}$$

for some constant $C_\varphi > 0$. We refer to [52] for further details on qualification conditions. In case of spectral cut-off regularization, any index function φ is a qualification of q_α, whereas for Tikhonov regularization this is only true for functions which increase slower than $\varphi(\lambda) = \lambda$ close to 0. For Showalter regularization, it can be said that at least all functions which increase slower than **some** polynomial around 0 are qualifications of the corresponding filter q_α.

Under (6) and (7), the bias in (5) can obviously be bounded by

$$\left\|r_\alpha\left(T^*T\right)f\right\|_{\mathscr{X}} \le C_\varphi\varphi(\alpha)\,\rho.$$

The estimation of the variance term in (5) requires more complicated techniques. If the noise ξ was deterministic and an element of $\mathscr{Y}$, the straight-forward estimate

$$\left\|q_\alpha\left(T^*T\right)T^*\xi\right\|_{\mathscr{X}} \le \left\|q_\alpha\left(T^*T\right)T^*\right\|\,\|\xi\|_{\mathscr{Y}} \le \frac{1}{\alpha}\,\|\xi\|_{\mathscr{Y}} \tag{8}$$

already leads to order optimal results. However, in the stochastic case, $\xi \notin \mathscr{Y}$ with probability 1 and hence one has to proceed more carefully. Explicit computations yield

$$\mathbb{E}\left[\left\|q_\alpha\left(T^*T\right)T^*\xi\right\|_{\mathscr{X}}^2\right] = \sum_{k=1}^{\infty}\sigma_k^2 q_\alpha\left(\sigma_k^2\right)^2 = \operatorname{Trace}\left(q_\alpha\left(T^*T\right)^2 T^*T\right). \tag{9}$$

In this situation it has been shown by Bissantz et al. [9] that under suitable assumptions on the decay of the singular values one has

$$\mathbb{E}\left[\left\|q_\alpha\left(T^*T\right)T^*\xi\right\|_{\mathscr{X}}^2\right] \leq C\frac{\#\left\{k \in \mathbb{N} \mid \sigma_k^2 \geq \alpha\right\}}{\alpha} \qquad \text{as} \qquad \alpha \searrow 0. \tag{10}$$

Note that this bound can be substantially larger than the deterministic one in (8), as it depends on the decay rate of the singular values σ_k.

2.2 Oracle Inequalities

An important tool to analyze the performance of a given parameter choice method are oracle inequalities. Suppose we are given a family of estimators $\hat{f}_\alpha$, $\alpha \in \mathscr{A}$, and define the *weak* and *strong risks*

$$r_{\mathrm{w}}\left(\alpha,f\right) := \mathbb{E}\left[\left\|T\hat{f}_\alpha - Tf\right\|_{\mathscr{Y}}^2\right], \qquad r_{\mathrm{s}}\left(\alpha,f\right) := \mathbb{E}\left[\left\|f_\alpha - f\right\|_{\mathscr{X}}^2\right],$$

which are deterministic functions of the parameter α. (Strong) oracle inequalities do now relate the MSE of $\hat{f}_{\bar{\alpha}}$ with some parameter choice $\bar{\alpha} = \bar{\alpha}\left(Y,\sigma\right)$ with the *oracle risks*, this is inequalities of the form

$$\mathbb{E}\left[\left\|\hat{f}_{\bar{\alpha}} - f\right\|_{\mathscr{X}}^2\right] \leq \Theta\left(\inf_\alpha r_{\mathrm{s}}\left(\alpha,f\right), \inf_\alpha r_{\mathrm{w}}\left(\alpha,f\right)\right) \qquad \text{as} \qquad \sigma \searrow 0 \tag{11}$$

with some function $\Theta : \mathbb{R}_{\geq 0}^2 \to \mathbb{R}_{\geq 0}$ which hold true uniformly for $f \in \mathscr{W}$. Note that the expectation on the left-hand side is taken w.r.t. Y and hence affects also $\bar{\alpha}$, which implies especially $\mathbb{E}\left[\left\|\hat{f}_{\bar{\alpha}} - f\right\|_{\mathscr{X}}^2\right] \neq \mathbb{E}\left[r_{\mathrm{s}}\left(\bar{\alpha},f\right)\right]$. The infimum on the right-hand side of (11) can either be taken over all possible parameters $\alpha \in \mathscr{A}$, or over a finite subset depending on the underlying regularization scheme. Equation (11) might seem strange on first glance, as the classical understanding of an oracle inequality is of the form

$$\mathbb{E}\left[\left\|\hat{f}_{\bar{\alpha}} - f\right\|_{\mathscr{X}}^2\right] \leq C\inf_\alpha r_{\mathrm{s}}\left(\alpha,f\right) + R\left(\sigma\right) \qquad \text{as} \qquad \sigma \searrow 0 \tag{12}$$

with some constant $C \geq 1$ and a remainder term $R\left(\sigma\right) = o\left(1\right)$ as $\sigma \searrow 0$. Obviously, this is a special case of (11) with $\Theta\left(a,b\right) = Ca + R\left(\sigma\right)$. If R decays at least as fast as the strong oracle risk, then the oracle inequality (12) ensures that $\bar{\alpha}$ performs up to a constant as good as choosing the optimal α on the smoothness class $\mathscr{W}$. Nevertheless, oracle inequalities of the form (12) are hard to obtain in practice, see e.g. the discussion in [18], where consequently an oracle inequality of the form (11) with Θ depending only on the second argument is proven. Furthermore, as the weak and strong oracle risks are of different order, it is to be expected that Θ is non-linear w.r.t. the second argument.

Similar to strong oracle inequalities, weak oracle inequalities relate the weak risk with the oracle risks, but due to the ill-posedness the weak risk carries only little information about the performance of $\bar{\alpha}$.

Oracle inequalities in statistical inverse problems have been studied intensively in the literature over the last two decades. We refer to [13, Sec. 3.2] for a slightly different introduction to oracle inequalities. Furthermore we mention [12, 20, 34, 35] for oracle inequalities in wavelet shrinkage approaches and [14–18, 24–26] for oracle inequalities in weighted projection methods (partially including some of the filter based methods discussed here). More recently, in [43] an oracle inequality for general filter based regularization with a specific parameter choice rule also discussed here (see Sect. 3.3) has been obtained. Lepskiĭ [40] and Blanchard et al. [11] discuss the usage of oracle inequalities in inverse problems from a more general perspective.

3 A Posteriori Parameter Choice Rules

In this section we will now discuss different a posteriori parameter choice rules for estimators of the form (2). Over the last decades, a vast variety of such methods have been proposed, developed, analyzed and compared. For a recent overview with numerical comparison we refer to [5]. We also refer to a series of papers by Hämarik et al. [27–29], where whole families of parameter choice rules are compared also for the situation of incomplete information about the noise level. In the following, we will always assume that the noise level is known exactly, and focus on four popular methods, namely

- the *discrepancy principle*,
- *Empirical risk minimization*, also known as *Mallow's C_L* or Stein's *unbiased risk estimator* (URE),
- the *Lepskiĭ-type balancing principle*, and
- the *quasi-optimality principle*.

There are many other parameter choice methods, which we do not consider here for various reasons. We emphasize, that a method's omission does not mean that it is too difficult to implement or that it performs poorly. Nevertheless, here we focus on methods for which oracle inequalities (11) are known, and which are meaningful in the infinite dimensional statistical setup chosen here (or can readily be turned meaningful by discretization as the discrepancy principle). Note that the discrepancy principle and the quasi-optimality principle have their mathematical origin in the deterministic setting, whereas URE and the Lepskiĭ-type balancing principle have been developed in statistics initially (cf. [5, Table 3]).

Below we will describe the four parameter choice methods investigated here, highlight their development, and briefly recall known results about their performance and analysis. As far as possible, all the results will be given for general $\mathscr{A}$, but some of the methods require a discretization of the parameter space by definition. We furthermore emphasize that all methods are defined by minimizing or maximizing some score function, and for discretized $\mathscr{A}$ they can consequently be implemented straight forward by evaluating the score function at all values and taking the optimal one. The ordering of the candidate values in $\mathscr{A}$ does not matter in this case. Nevertheless, the discrepancy principle can be implemented faster by checking the candidate values in a decreasing order, whereas the Lepskiĭ-type balancing principle can be implemented faster by checking in an increasing order.

3.1 The Bakushinskiĭ veto

In deterministic inverse problems, the famous Bakushinskiĭ veto [1] tells that any parameter choice method independent of the noise level cannot lead to a convergent regularization scheme for an ill-posed problem. More precisely, given deterministic data $Y \in \mathscr{Y}$, a filter q_α and a parameter choice $\bar{\alpha} = \bar{\alpha}(Y)$, then it states that worst case convergence

$$\lim_{\sigma \searrow 0} \sup \left\{ \left\| R_{\bar{\alpha}(Y)} - f \right\|_{\mathscr{X}} \;\middle|\; Y \in \mathscr{Y}, \|Y - Tf\|_{\mathscr{Y}} \leq \sigma \right\} = 0$$

for all $f \in \mathscr{X}$ implies boundedness of T^{-1} on its range. Consequently, if the underlying problem is ill-posed, any convergent a posteriori parameter choice rule must depend not only on Y, but also on the noise level σ. However, this result is not directly transferable to statistical inverse problems as considered here. The question of transferability was initially raised in [7] and extensively answered in [8]. In the specific case we consider here, i.e. that the probability distribution of the noise does not change if the noise level changes, it follows from the results there that the Bakushinskiĭ veto does not hold true. As an immediate consequence, there are convergent a posteriori parameter choice rules independent of the noise level σ.

3.2 Discrepancy Principle

The discrepancy principle dates back to Phillips and Morozov [54, 56] and is based on the simple idea that the chosen reconstruction $\hat{f}_\alpha$ should not try to explain the observed data better than the accuracy of the data actually is. If $\mathbb{E}\left[\|\xi\|_{\mathscr{Y}}\right] < \infty$, this means that α should be chosen such that $\left\| T\hat{f}_\alpha - Y \right\|_{\mathscr{Y}} \approx \sigma \|\xi\|_{\mathscr{Y}}$. Typically, $\alpha \mapsto \left\| T\hat{f}_\alpha - Y \right\|_{\mathscr{Y}}$ is increasing as α increases, and hence a reasonable parameter choice in this spirit is

$$\alpha_{\mathrm{DP}} = \max\left\{\alpha \in \mathscr{A} \;\middle|\; \left\|T\hat{f}_\alpha - Y\right\|_{\mathscr{Y}} \leq \tau\sigma\mathbb{E}\left[\|\xi\|_{\mathscr{Y}}\right]\right\} \tag{13}$$

with a tuning parameter $\tau \geq 1$. However, in the situation of Gaussian white noise ξ this principle is not applicable as $\xi \notin \mathscr{Y}$. This difficulty can be overcome for example by first applying T^* to the data, which then yields an operator equation in $\mathscr{X}$ of the form

$$Z = T^*Tf + \sigma T^*\xi,$$

i.e. with noise $T^*\xi$ with finite $\mathscr{X}$-norm [10, 44]. Here we proceed differently and apply the original formulation (13) to the discretized equation where T is an $n \times n$ matrix and $\xi \sim \mathscr{N}(0, I_n)$. If $\|\cdot\|_{\mathscr{Y}}$ is the Euclidean norm, then $\mathbb{E}\left[\|\xi\|_{\mathscr{Y}}\right] = \sqrt{n}$, and hence (13) can be used. This choice of α has e.g. been analyzed in [19, 46], and in [11] also an oracle inequality for some iterative regularization methods has been obtained. One essential drawback of the discrepancy principle is that for order optimality a higher qualification condition is required. More precisely, not φ as in (6), but $\lambda \mapsto \sqrt{\lambda}\varphi(\lambda)$ has to be a qualification of the filter q_α, this is [cf. (7)]

$$\sup_{\lambda \in [0, \|T^*T\|]} \sqrt{\lambda}\varphi(\lambda)\left|r_\alpha(\lambda)\right| \leq C'_\varphi\sqrt{\alpha}\varphi(\alpha) \qquad \text{for all } \alpha \in \mathscr{A}. \tag{14}$$

It should also be noted that the actual performance can be quite sensitive w.r.t. the tuning parameter τ, and there is no clear roadmap how to choose τ in a specific example. In our simulations we will use $\tau = 1.5$. In case of a discretized parameter set it is immediately clear that the discrepancy principle is computationally very cheap, as only the residuals $\left\|T\hat{f}_{\alpha_k} - Y\right\|_{\mathscr{Y}}$ have to be evaluated.

3.3 *Empirical Risk Minimization*

As the ideal α should minimize the weak risk $r_{\mathrm{w}}(\alpha, f)$, one possible idea is to mimic this behavior by minimizing an (up to a constant) unbiased estimator of $r_{\mathrm{w}}(\alpha, f)$. This idea dates back to Mallows [48] and Stein [59], and therefore it is also known as *Mallow's C_L* or Stein's *unbiased risk estimator* (URE). Straight forward computations show that with

$$\hat{r}_{\mathrm{w}}(\alpha, Y) := \left\|T\hat{f}_\alpha\right\|^2_{\mathscr{Y}} - 2\left\langle T\hat{f}_\alpha, Y\right\rangle + 2\sigma^2\mathrm{Trace}\left(T^*Tq_\alpha\left(T^*T\right)\right) \tag{15}$$

it holds

$$\mathbb{E}\left[\hat{r}_{\mathrm{w}}(\alpha, Y)\right] = r_{\mathrm{w}}(\alpha, f) - \sum_{k=1}^{\infty}\sigma_k^2 f_k^2, \qquad \alpha \in \mathscr{A}.$$

Therefore, we define

$$\alpha_{\mathrm{URE}} \in \operatorname*{argmin}_{\alpha\in\mathscr{A}} \hat{r}_{\mathrm{w}}(\alpha, Y). \tag{16}$$

For a more detailed derivation of α_{URE} we refer to [62, Sec. 7.1] and [47]. Note that there is some commonality in the definition of α_{URE} and generalized cross-validation (GCV) as discussed in [21, 41, 63], but we emphasize that α_{URE} is meaningful also in the continuous setting (1), in which GCV cannot be applied. It is known that choosing $\alpha = \alpha_{\mathrm{URE}}$ in combination with certain regularization schemes leads to an order optimal method w.r.t. the weak MSE $\mathbb{E}\left[\left\|T\hat{f}_{\alpha_{\mathrm{URE}}} - Tf\right\|_{\mathscr{Y}}^2\right]$, see e.g. [42, 45, 61]. Using the seminal results by Kneip [37], it has recently been shown in [43] that α_{URE} yields an oracle inequality of the form (11) with Θ depending only on the second argument, which then leads to order-optimality w.r.t. the MSE $\mathbb{E}\left[\left\|\hat{f}_{\alpha_{\mathrm{URE}}} - f\right\|_{\mathscr{X}}^2\right]$ for mildly ill-posed operators if the stronger qualification condition (14) is satisfied. Concerning the computational cost, it must be said that (16) is more expensive than the discrepancy principle due to the evaluation of the trace operator in (15).

Note that it is also possible to estimate the strong risk $r_{\mathrm{s}}(\alpha, f)$ in an unbiased way, as discussed in [16]. Unfortunately, the corresponding parameter choice is only order-optimal for mildly ill-posed operators, and its practical performance deteriorates as the degree of ill-posedness grows (see e.g. the simulations in [18]).

3.4 *The Lepskiĭ-Type Balancing Principle*

The Lepskiĭ-type balancing principle was originally introduced by Lepskiĭ [39], and was further developed for usage in (statistical) inverse problems in [2, 49, 52, 53, 65]. Suppose m possible values $\alpha_1 < \ldots < \alpha_m$ for the regularization parameter α are given. Then the Lepskiĭ-type balancing principle consists in choosing

$$j^* = \max\left\{1 \le j \le m \;\middle|\; \left\|\hat{f}_{\alpha_j} - \hat{f}_{\alpha_k}\right\|_{\mathscr{X}} \le 4\kappa\sigma\sqrt{\operatorname{Trace}\left(q_{\alpha_k}\left(T^*T\right)^2 T^*T\right)} \text{ for all } 1 \le k \le j\right\}. \tag{17}$$

and $\alpha_{\mathrm{LEP}} := \alpha_{j^*}$. Here, $\kappa \ge 1$ is again a tuning parameter. For an explanatory derivation of this choice we refer to [49]. It is worth noting that $\sigma\sqrt{\operatorname{Trace}(q_{\alpha_k}(T^*T)^2 T^*T)}$ is in fact an upper bound for the standard deviation of the estimator $\hat{f}_{\alpha_k}$ as seen in (5). Under suitable assumptions on the filter, the noise behavior and the definition of the regularization parameters $\alpha_1, \ldots, \alpha_m$ it has been shown in [6, 53] that one obtains an oracle inequality of the form (11) with $\Theta(a, b) = C_1 a + C_2\sqrt{m}\exp\left(-C_3\kappa^2\right)$. From this one can deduce that the MSE decays of optimal order up to a logarithmic factor, if κ is chosen appropriately. To be more precise, if T is mildly ill-posed one should use $\kappa \sim \sqrt{-\log(\sigma)}$ and one obtains the optimal rate up to $\log(-\sigma)$ [53], and if T is severely ill-posed, the choice $\kappa \sim \sqrt{\log(-\log(\sigma))}$ leads to the optimal rate up to $\log(\log(-\sigma))$ [6]. In our simulations we always set $\kappa = 1$. Especially for severely ill-posed T, the growth of $\sqrt{\log(-\log(\sigma))}$ is too slow that any difference between this choice and $\kappa = 1$ could be seen in simulations.

Finally, we mention that the computation of α_{LEP} is even more expensive than the one of α_{URE}, as the reconstructions have to be compared among each other.

3.5 The Quasi-Optimality Criterion

The quasi-optimality criterion was originally introduced by Tikhonov and Glasko [60], and was further developed for usage in (statistical) inverse problems by [3, 4, 7, 36, 55]. In the literature, there are different definitions of the quasi-optimality criterion, depending on the considered setting. In the setting of general filters, Neubauer [55] defines

$$\alpha_{\mathrm{QO}} := \operatorname*{argmin}_{\alpha \in \mathscr{A}} \left\| r_\alpha\left(T^*T\right)\hat{f}_\alpha \right\|_{\mathscr{X}}.$$

Apparently, this is not a meaningful choice for spectral cut-off regularization, where $r_\alpha\,(T^*T)\hat{f}_\alpha = 0$ for all $\alpha \in \mathscr{A}$. The more common and initial definition—which we will use here—works for a discrete set $\alpha_1 < \ldots. < \alpha_m$ of possible regularization parameters and is given by

$$n_{\mathrm{QO}} := \operatorname*{argmin}_{1 \le n \le m-1} \left\| \hat{f}_{\alpha_n} - \hat{f}_{\alpha_{n+1}} \right\|_{\mathscr{X}} \qquad \text{and} \qquad \alpha_{\mathrm{QO}} := \alpha_{n_{\mathrm{QO}}}. \tag{18}$$

If $\alpha_n = \alpha_0 q^n$, it can readily be seen that both definitions are consistent for Tikhonov regularization. Note that the computation of α_{QO} is substantially more simple than the ones discussed above, as neither an estimate for the variances of $\hat{f}_{\alpha_j}$ nor the noise level σ is required. The latter is also very helpful in practice, as no or only rough knowledge of the noise level makes the methods discussed above not applicable or unstable. As mentioned before, the famous Bakushinskiĭ veto [1] does not hold true in the situation considered in this paper. In fact, it has been shown that the quasi-optimality criterion leads to a convergent regularization scheme under suitable assumptions. For Tikhonov regularization, [3, 4] show an oracle inequality of the form (11) if the unknown solution satisfies a source condition (6) of Hölder-type (this is $\varphi(\lambda) = \lambda^\nu$) with $0 < \nu \le 1$ and if T is mildly ill-posed. For spectral cut-off regularization, order optimality has been shown in a Bayesian setting in [7]. We also mention [36] for results in the purely deterministic case.

4 Numerical Comparison

In this section we will compare the aforementioned parameter choice rules in several numerical examples. Note that all of them consist in minimizing some functional w.r.t. α, which can be challenging in general. We therefore discretize the set $\mathscr{A}$ of possible regularization parameters in a logarithmically equispaced way. Furthermore we restrict ourselves to regularization parameters in $\left[\sigma^2, \|T^*T\|\right]$, as the optimal parameter will (asymptotically) be larger that σ^2 and smaller than $\|T^*T\|$. Consequently,

$$\mathscr{A}_{\mathrm{d}} = \left\{\alpha_k = \sigma^2 \cdot r^k \mid k = 0, 1, \ldots, \left\lfloor (\log(r))^{-1} \log\left(\sigma^{-2} \|T^*T\|\right) \right\rfloor\right\} \tag{19}$$

with $r > 1$ should be an appropriate approximation of the continuous parameter set $\mathscr{A}$. Clearly, a larger value of r will result in a worse practical performance, whereas $r \approx 1$ typically makes the computations unfeasible. Note that in many examples one finds from the error decomposition (5) that the discrete parameter set $\mathscr{A}_{\mathrm{d}}$ is able to resemble the optimal behavior of a continuous parameter set $\mathscr{A}$ up to a constant depending on r (see e.g. [65]). In our simulations we will always use $r = 1.2$. We also tried different values of r which did not influence the results significantly.

In practice, some of the investigated parameter choice methods are very sensitive w.r.t. too small or too large values of α. This is especially the case for the quasi-optimality criterion, which is known to oversmooth the solution if $\mathscr{A}_{\mathrm{d}}$ contains too large α's, and to undersmooth if it contains too small values (cf. [5]). To avoid both, we furthermore consider only a subset $\mathscr{A}_{\mathrm{d}}' \subset \mathscr{A}_{\mathrm{d}}$ such that $\max_{\alpha \in \mathscr{A}_{\mathrm{d}}'} \alpha \leq 1$ and $\Phi(\alpha) \leq 2^{-1} \max_{\alpha \in \mathscr{A}_{\mathrm{d}}} \Phi(\alpha)$ for all $\alpha \in \mathscr{A}_{\mathrm{d}}'$ with the variance function

$$\Phi^2(\alpha) = \sigma^2 \mathrm{Trace}\left(q_\alpha\left(T^*T\right)^2 T^*T\right).$$

The rationale behind the second restriction is as follows. As Φ is monotonically decreasing one has $\max_{\alpha \in \mathscr{A}_{\mathrm{d}}} \Phi(\alpha) = \Phi(\alpha_0)$. Furthermore there exists some $\bar{\alpha}$ such that $\Phi(\alpha) \leq 2^{-1}\Phi(\alpha_0)$ for all $\alpha \geq \bar{\alpha}$ and $\Phi(\alpha) > 2^{-1}\Phi(\alpha_0)$ for all $\alpha < \bar{\alpha}$. Using that the first term in (5) is monotonically increasing we can now compute

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f}_{\alpha_0} - f\right\|_{\mathscr{X}}^2\right] &= \left\|r_{\alpha_0}\left(T^*T\right)f\right\|_{\mathscr{X}}^2 + \sigma^2\Phi(\alpha_0)^2 \leq 4\left\|r_\alpha\left(T^*T\right)f\right\|_{\mathscr{X}}^2 + 4\sigma^2\Phi(\alpha)^2 \\ &= 4\mathbb{E}\left[\left\|\hat{f}_\alpha - f\right\|_{\mathscr{X}}^2\right]\end{aligned}$$

for all $\alpha \in \mathscr{A}_{\mathrm{d}}, \alpha < \bar{\alpha}$. But for the optimal $\alpha \in \mathscr{A}_{\mathrm{d}}$, a lower bound as proven above is not to be expected, and hence it is reasonable that the optimal $\alpha \in \mathscr{A}_{\mathrm{d}}$ should satisfy $\alpha \geq \bar{\alpha}$ or equivalently $\Phi(\alpha) \leq 2^{-1} \max_{\alpha \in \mathscr{A}_{\mathrm{d}}} \Phi(\alpha)$, i.e. it should be contained in $\mathscr{A}_{\mathrm{d}}'$.

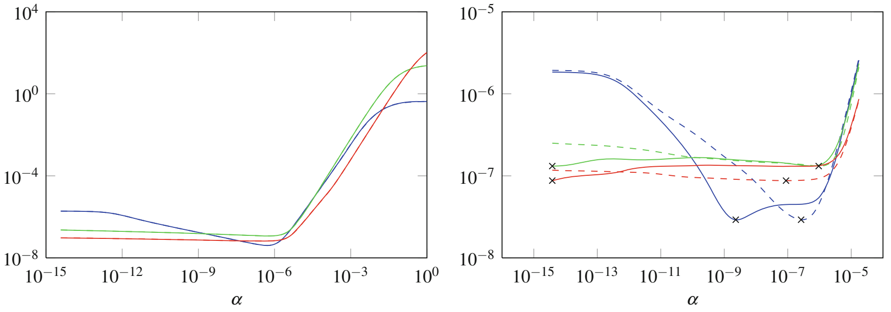


Fig. 1 Empirical means of the unbiased risk estimation score function $\alpha \mapsto \hat{r}_{\mathrm{w}}(\alpha, Y)$ and the risk $r_{\mathrm{w}}(\alpha, f)$ (simulated from 10^4 runs; left) and problematic example cases of both (right) for Tikhonov regularization with $\sigma = 6.1035 \cdot 10^{-5}$. The corresponding minima are always marked by a cross. Shown are the antiderivative problem (score function: ——, true risk: - - -), the satellite gradiometry problem (score function: ——, true risk: - - -), and the backwards heat problem (score function: ——, true risk: - - -)

It will turn out that the score function $\hat{r}_{\mathrm{w}}(\alpha, Y)$ in the unbiased risk estimation principle is to some extend sensitive w.r.t. the noise in Y, which causes some instabilities in the corresponding parameter choice strategy. Even though $\mathbb{E}[\hat{r}_{\mathrm{w}}(\alpha, Y)] = r_{\mathrm{w}}(\alpha, f)$ up to a constant independent of α, both quantities can vary significantly for single experiments. This is shown in Fig. 1, where on the left empirical means of both quantities are shown (simulated from 10^4 runs), and on the right some problematic cases of both are depicted. In all cases, $\hat{r}_{\mathrm{w}}(\cdot, Y)$ has been shifted such that $\min_{\alpha \in \mathscr{A}_{\mathrm{d}}'} \hat{r}_{\mathrm{w}}(\alpha, Y) = \min_{\alpha \in \mathscr{A}_{\mathrm{d}}'} r_{\mathrm{w}}(\alpha, f)$. In the plot it can be seen that even though in expectation both quantities agree quite well, in some cases the variation causes a big difference in the minimizers. We will see that this leads to instabilities in the parameter choice strategy based on empirical risk minimization. Up to some extend this is caused by the design of the method, as it is based on some quantity ($\hat{r}_{\mathrm{w}}(\alpha, Y)$) which behaves correctly *in expectation*, whereas all the other parameter choice strategies are designed based on the available (single) instance of the problem.

Furthermore we emphasize that $r_{\mathrm{w}}(\cdot, f)$ as well as $\hat{r}_{\mathrm{w}}(\cdot, Y)$ vary over several orders of magnitude, and that the more ill-posed the problem, the more flat is $\hat{r}_{\mathrm{w}}(\cdot, Y)$ around its minimum.

In the following we will compare the empirical MSE and its variance under the investigated parameter choice methods from Sect. 3 in three different problems with three different regularization methods. The three problems and the corresponding results will be given below. As regularization methods we consider spectral cut-off regularization, Tikhonov regularization, and Showalter regularization, which are all three described in Sect. 2. The empirical MSE and its variance will be computed by Monte Carlo simulations with 10^4 experiments per noise level $\sigma \in \{2^{-15}, \ldots, 2^{-25}\}$. For comparison, we will also depict results for the optimal but practically unavailable parameter choice rule

$$\alpha_{\mathrm{or}} := \operatorname*{argmin}_{\alpha \in \mathscr{A}_{\mathrm{d}}'} r_{\mathrm{s}}(\alpha, f) \tag{20}$$

and (in case of the MSE) the optimal rate of convergence known from the theory. For the empirical variance, the optimal rate of convergence is not discussed in this study. Nevertheless, asymptotically we expect that

$$\mathbb{E}\left[\left\|\hat{f}_\alpha - f\right\|_{\mathscr{X}}^2\right] \sim \mathcal{N}\left(m(\sigma), v(\sigma)\right)$$

with functions $m, v : \mathbb{R}^+ \to \mathbb{R}^+$ corresponding to the analytical mean and variance of the MSE of the estimator. The decay of m is typically considered as the rate of convergence of the estimator, but the decay of v is interesting as well, as it can be interpreted as the rate of concentration. If v does not tend to 0, then the estimator might have a good MSE, but still its practical performance is questionable. Consequently, in our plots we depict estimates of m and v.

4.1 A Mildly Ill-Posed Problem: The Second Antiderivative

At first we investigate the empirical rate of convergence in a mildly ill-posed situation borrowed from [31]. Consider the following Fredholm integral operator $T : \mathbf{L}^2([0,1]) \to \mathbf{L}^2([0,1])$ of the first kind

$$(Tf)(x) = \int_0^1 k(x,y) f(y)\, \mathrm{d}y, \qquad x \in [0,1]$$

with kernel $k(x,y) = \min\{x \cdot (1-y), y \cdot (1-x)\}, x, y \in [0,1]$. This implies that $(Tf)'' = -f$ for all $f \in \mathbf{L}^2([0,1])$. Explicit computations show that the singular values σ_k of T satisfy $\sigma_k \sim k^{-2}$.

For the discretization of this operator we choose the composite midpoint rule, i.e. with the equidistant points $x_1 = \frac{1}{2n}, x_2 = \frac{3}{2n}, \ldots, x_n = \frac{2n-1}{2n}$ we approximate

$$(Tf)(x_i) = \int_0^1 k(x_i, y) f(y)\, \mathrm{d}y \approx \frac{1}{n}\sum_{j=1}^n k\left(x_i, x_j\right) f\left(x_j\right), \qquad 1 \le i \le n.$$

As exact solution we consider the continuous function

$$f(x) = \begin{cases} x & \text{if } 0 \le x \le \frac{1}{2}, \\ 1 - x & \text{if } \frac{1}{2} \le x \le 1. \end{cases} \tag{21}$$

As argued in [43], the optimal rate of convergence in this situation is $\mathcal{O}\left(\sigma^{\frac{3}{4}-\varepsilon}\right)$ for any $\varepsilon > 0$. To avoid an inverse crime, the exact data $g = Tf$ is implemented analytically:

$$g\left(x\right) = \begin{cases} -\frac{x\left(4x^2-3\right)}{24} & \text{if } 0 \leq x \leq \frac{1}{2}, \\ \frac{(x-1)\left(4x^2-8x+1\right)}{24} & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

The results are shown in Figs. 2, 3, and 4. We find that all parameter choice rules under investigation yield the optimal rate of convergence. The quasi-optimality criterion seems to be most sensitive w.r.t. the possible regularization parameters, which is mostly observed in combination with spectral cut-off regularization. This

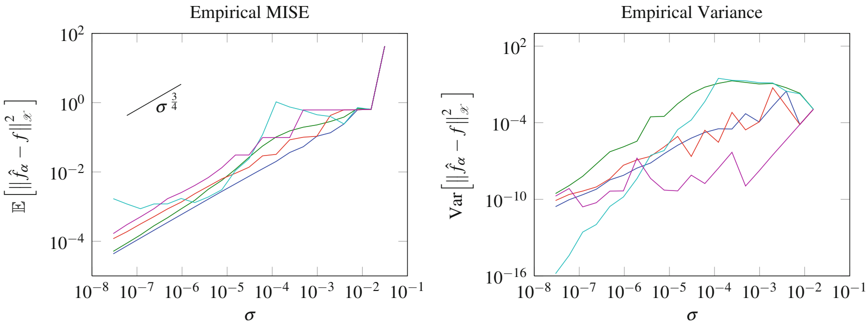


Fig. 2 Simulation results for spectral cut-off regularization in the antiderivative problem from Sect. 4.1 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

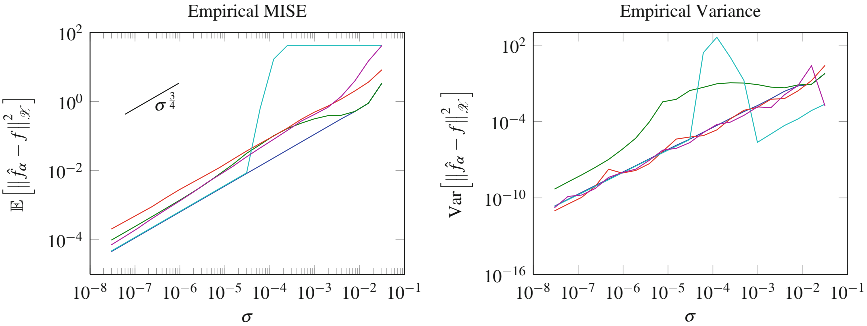


Fig. 3 Simulation results for Tikhonov regularization in the antiderivative problem from Sect. 4.1 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

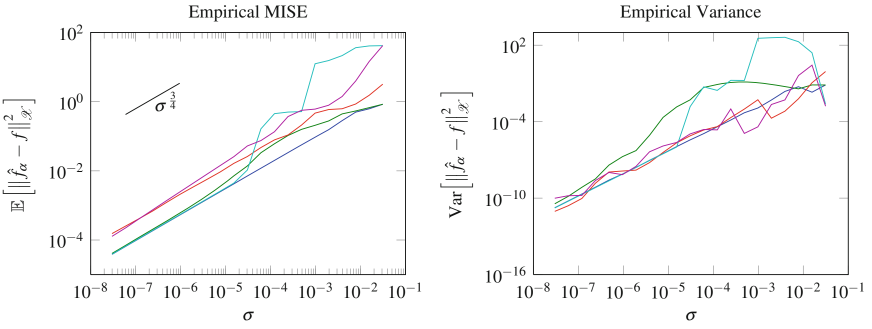


Fig. 4 Simulation results for Showalter regularization in the antiderivative problem from Sect. 4.1 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

effect has already been observed in [5] and will also be seen in the other two examples below. Note that the conditions mentioned in Sect. 3.5 to ensure order optimality are satisfied here. Concerning the other parameter choice strategies, it seems that empirical risk minimization has a slightly higher variance than the discrepancy principle and the Lepskiĭ-type balancing principle. It is very likely that this is caused by problems in minimizing $\alpha \mapsto \hat{r}_{\mathrm{w}}(\alpha, Y)$ as show in Fig. 1.

4.2 A Severely Ill-Posed Operator: Satellite Gradiometry

As a second example we consider an inverse problem in satellite gradiometry [23]. Consider the unit ball $B = \left\{x \in \mathbb{R}^d \;\middle|\; \|x\|_2 = 1\right\}$ and the unit sphere $S := \partial B$. Given measurements of $g = \frac{\partial^2 u}{\partial r^2}$ on RS with $R > 1$ we want to find f in

$$\begin{cases} \Delta u = 0 & \text{in } \mathbb{R}^d \setminus B, \\ u = f & \text{on } S, \\ |u(x)| = \mathscr{O}\left(\|x\|_2^{-1}\right) & \text{as } \|x\|_2 \to \infty. \end{cases}$$

If $d = 3$ and B is considered as an approximation of the earth, then u describes the gravitational potential of the earth, and we want to determine this potential on the earth's surface from satellite measurements of the potentials second derivative of the gravitational potential in radial direction. For computational simplicity, we consider $d = 2$ in the following. Let us define the forward operator $T : f \mapsto g$ as $T : \mathbf{L}^2(S, \mu) \to \mathbf{L}^2(RS, \mu)$ with the surface measure μ on S. Note that u can be computed explicitly using the Poisson formula. If the corresponding integral kernel is furthermore expanded as a power series in R, this gives an explicit representation of T in form of a Fourier series (cf. [38]), i.e.

$$(Tf)(x) = \sum_{k=-\infty}^{\infty} |k|\left(|k|+1\right) R^{-|k|-2} \exp(ikx) \hat{f}(k) .$$

Consequently, T is severely ill-posed with singular values $\sigma_k = |k|\left(|k|+1\right) R^{-|k|-2}$. Furthermore it follows that $\hat{f}(0)$ cannot be determined from Tf. Hence we choose

$$f(x) = \frac{\pi}{2} - |x| , \qquad x \in [-\pi, \pi]$$

as exact solution. This ensures that $\|f + c\|_{\mathbf{L}^2}^2$ is minimal for $c = 0$, i.e. all regularization schemes considered here will produce reconstructions which converge towards f. It follows from [30, Prop.16] and similar computations as in [43] that the optimal rate of convergence in this situation is $\mathscr{O}\left(\left(-\log(\sigma)\right)^{-3+\varepsilon}\right)$ for any $\varepsilon > 0$. Furthermore, $g = Tf$ can be computed analytically from

$$g(x) = \frac{4}{\pi} \sum_{m \in \mathbb{N}} \left(1 + \frac{1}{2m+1}\right) R^{-2m-3} \cos\left((2m+1)x\right) , \qquad x \in [-\pi, \pi] .$$

To discretize T we choose again equidistant points $x_1, \ldots, x_n$ in $[-\pi, \pi]$ and replace f by the piecewise constant approximation $\tilde{f} = \sum_{j=1}^{n} f\left(x_j\right) \mathbf{1}_{\left[x_j - \pi/n, x_j + \pi/n\right]}$. This yields

$$(Tf)(x_i) \approx \sum_{j=1}^{n} \left(\frac{2}{\pi} \sum_{m \in \mathbb{N}} \sin\left(\frac{\pi m}{n}\right)(m+1) R^{-m-2} \cos\left(m\left(x_i - x_j\right)\right)\right) f\left(x_j\right) , \quad 1 \le i \le n.$$

The inner sum is truncated at $m = 64$, and to avoid an inverse crime, the summation in the definition of g is truncated at $m = 128$. In our simulations, we set $R = 2$.

The results are shown in Figs. 5, 6, and 7. For spectral cut-off regularization we observe a different behavior than for Tikhonov and Showalter regularization. This is due to the fact that the ‘interesting’ regularization parameters for spectral cut-off regularization are exactly $\alpha = \sigma_k$ with the singular values σ_k of T, but those are not well covered by our set $\mathscr{A}'_{\mathrm{d}}$. More precisely, there are some $\alpha_k \in \mathscr{A}'_{\mathrm{d}}$ which yield the same spectral cut-off reconstruction, and some pairs with several singular values in between. Still it can be seen that all parameter choice strategies but the quasi-optimality principle yield the order optimal convergence rate for spectral cut-off, and all strategies yield the order optimal convergence rate in Tikhonov and Showalter regularization. Furthermore we observe that empirical risk minimization behaves slightly worse and its variance decreases only slowly. This is due to the fact that $\alpha \mapsto \hat{r}_{\mathrm{w}}(\alpha, Y)$ is nearly constant around its minimum, and hence the minimizer has a high variance itself, cf. Fig. 1. Consequently, it is questionable if URE performs optimal in this situation.

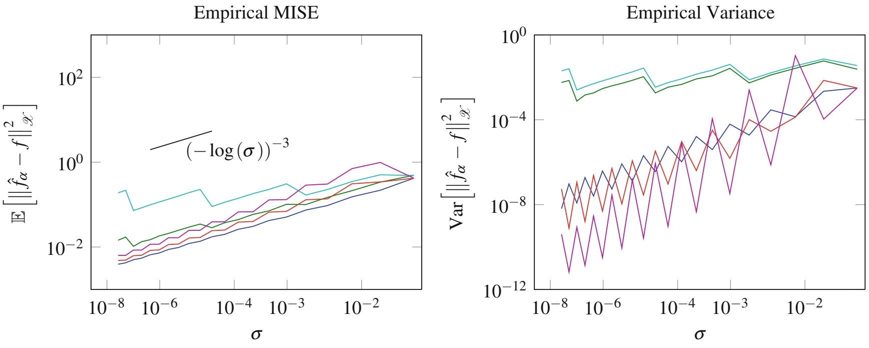


Fig. 5 Simulation results for spectral cut-off regularization in the satellite gradiometry problem from Sect. 4.2 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

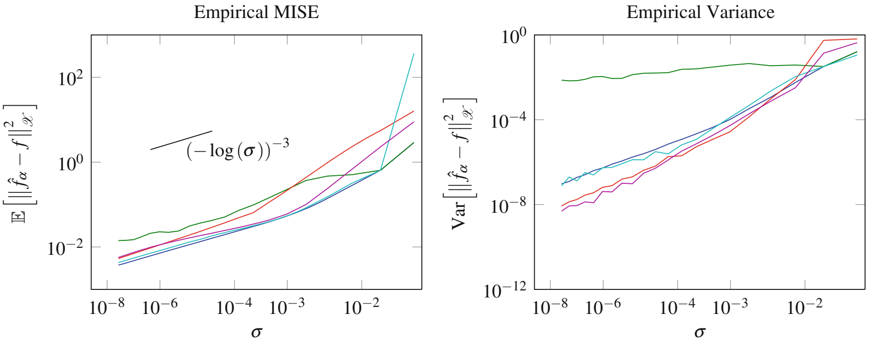


Fig. 6 Simulation results for Tikhonov regularization in the satellite gradiometry problem from Sect. 4.2 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

4.3 Another Severely Ill-Posed Operator: The Backwards Heat Equation

As a third example we consider the so-called backwards heat equation. Given measurements of $g = u(\cdot, \bar{t})$ with $\bar{t} > 0$ we want to find f in the periodic heat equation

$$\begin{cases} \frac{\partial u}{\partial t}(x,t) = \frac{\partial^2 u}{\partial t^2}(x,t) & \text{in } (-\pi, \pi] \times (0, \bar{t}), \\ u(x,0) = f(x) & \text{on } [-\pi, \pi], \\ u(-\pi, t) = u(\pi, t) & \text{on } t \in (0, \bar{t}]. \end{cases}$$

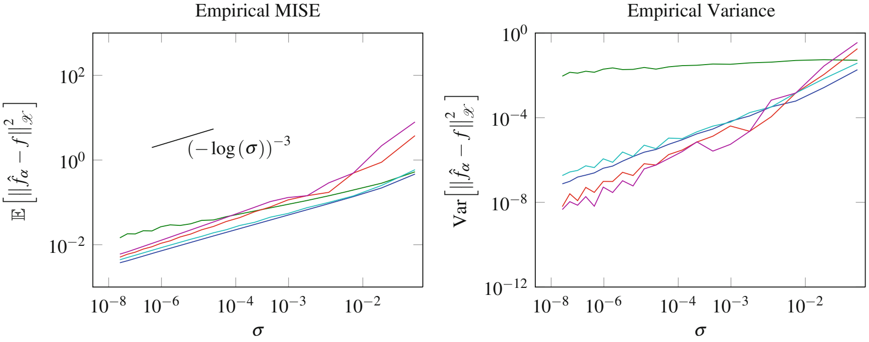


Fig. 7 Simulation results for Showalter regularization in the satellite gradiometry problem from Sect. 4.2 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

Let us define the forward operator $T : f \mapsto g$ as $T : \mathbf{L}^2([-\pi,\pi]) \to \mathbf{L}^2([-\pi,\pi])$. Separation of variables gives an explicit representation of T in form of a Fourier series, i.e.

$$(Tf)(x) = \sum_{k=-\infty}^{\infty} \exp\left(-k^2\bar{t}\right) \exp(\mathrm{i}kx) \hat{f}(k).$$

Consequently, T is severely ill-posed with singular values $\sigma_k = \exp\left(-k^2\bar{t}\right)$. As exact solution we choose again

$$f(x) = \frac{\pi}{2} - |x|, \qquad x \in [-\pi,\pi].$$

Similarly as in [30, Rem.15] it can be seen that the optimal rate of convergence is $\mathscr{O}\left((-\log(\sigma))^{-3/2+\varepsilon}\right)$ for any $\varepsilon > 0$. Furthermore, $g = Tf$ can be computed analytically from

$$g(x) = \frac{4}{\pi} \sum_{m\in\mathbb{N}} \frac{\exp\left(-(2m+1)^2\,\bar{t}\right)}{(2m+1)^2} \cos((2m+1)x), \qquad x \in [-\pi,\pi].$$

To discretize T we proceed as in Sect. 4.2, which yields

$$(Tf)(x_i) \approx \sum_{j=1}^{n} \left(\frac{2}{\pi} \sum_{m\in\mathbb{N}} \frac{\exp\left(-m^2\bar{t}\right)}{m} \sin\left(\frac{\pi m}{n}\right) \cos\left(m\left(x_i - x_j\right)\right) + \frac{1}{n}\right) f\left(x_j\right), \quad 1 \le i \le n.$$

The inner sum is truncated at $m = 64$, and to avoid an inverse crime, the summation in the definition of g is truncated at $m = 128$. In our simulations, we set $\bar{t} = 0.1$.

The results are shown in Figs. 8, 9, and 10. Again for spectral cut-off regularization the empirical MSE behaves less regular, which is due to the extremely fast decay of the singular values. A difference is only to be expected once σ falls below the next singular value, which explains the step-like behavior. Besides this, it seems that all parameter choice strategies yield the order optimal convergence rate for spectral cut-off, Tikhonov and Showalter regularization. Even though the severely ill-posed case is not covered by the assumptions from Sect. 3.5 to ensure this for the quasi-optimality criterion, this result suggests that something similar should hold true for severely ill-posed operators. For all regularization methods, the variances behave comparably irregular, even though they are small compared to

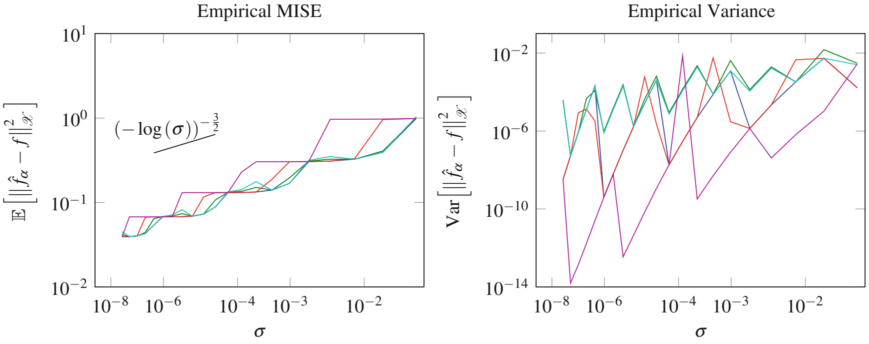


Fig. 8 Simulation results for spectral cut-off regularization in the backwards heat problem from Sect. 4.3 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

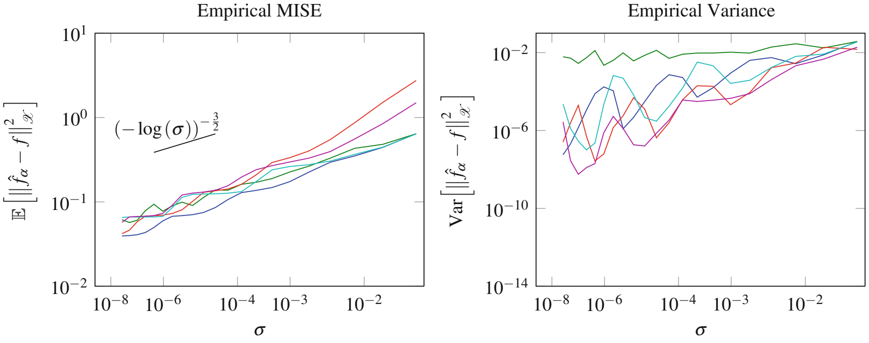


Fig. 9 Simulation results for Tikhonov regularization in the backwards heat problem from Sect. 4.3 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

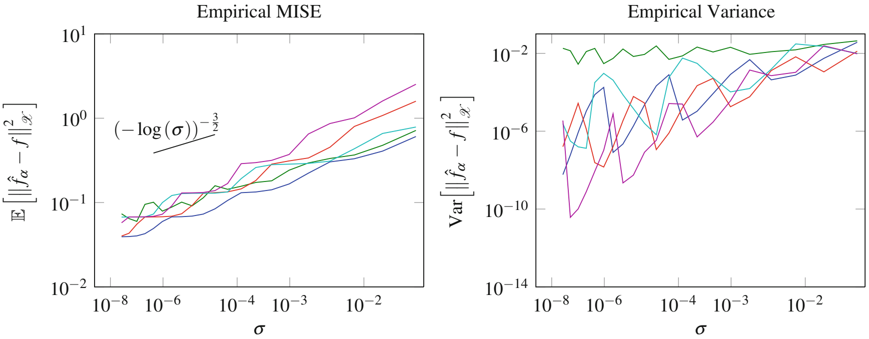


Fig. 10 Simulation results for Showalter regularization in the backwards heat problem from Sect. 4.3 with different parameter choice methods: oracle choice (——), empirical risk minimization (——), Lepskiĭ-type balancing principle (——), quasi-optimality criterion (——), and discrepancy principle (——)

the MSE and decay as $\sigma \searrow 0$. The only exception is empirical risk minimization where the variance roughly stays constant. This is again due to the fact that $\alpha \mapsto \hat{r}_{\mathrm{w}}(\alpha, Y)$ is nearly constant around its minimum, and hence the minimizer has a high variance itself, cf. Fig. 1. Again, it is questionable if URE performs optimal in this situation.

4.4 Inefficiency Simulations

As we are not only interested in convergence rates simulations, but also in oracle inequalities, we will now try to infer numerically if an oracle inequality of the form (12) holds, more precisely we want to know if

$$\mathbb{E}\left[\left\|\widehat{f_{\bar{\alpha}}} - f\right\|_{\mathscr{X}}^2\right] \leq c \min_{\alpha \in \mathscr{A}_{\mathbf{d}}'} \mathbb{E}\left[\left\|\hat{f}_\alpha - f\right\|_{\mathscr{X}}^2\right] \tag{22}$$

is satisfied and if so, what is the best possible value of $c \geq 1$. Inspired by [5, 18] we consider the following setup. The forward operator is a 300×300 diagonal matrix with singular values $\lambda(k) = \exp(-ak)$ with fixed parameter $a > 0$. Consequently, the ill-posedness is comparable to the satellite gradiometry problem. Then we repeat the following experiment 10^4 times : Given a parameter ν we generate a random ground truth $f \in \mathbb{R}^{300}$ by $f(k) = \pm k^{-\nu} \cdot \left(1 + \mathscr{N}\left(0, 0.1^2\right)\right)$ where the sign is independent and uniformly distributed for each component. From this ground truth, data is generated according to $Y(k) = \lambda(k) \cdot f(k) + \mathscr{N}\left(0, \sigma^2\right)$ where the noise is again independent in each component. Based on the data compute empirical

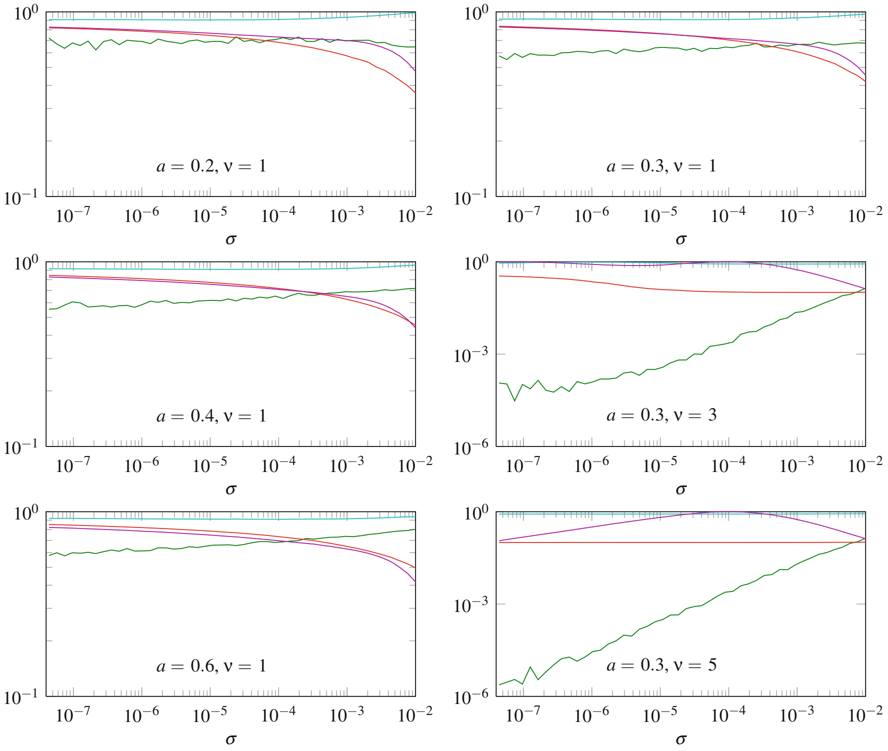


Fig. 11 Efficiency simulations for Tikhonov regularization with different smoothness parameters a and ν. Shown are $R_{\mathrm{OR}}/R_{\mathrm{URE}}$ (——), $R_{\mathrm{OR}}/R_{\mathrm{LEP}}$ (——), $R_{\mathrm{OR}}/R_{\mathrm{QO}}$ (——), and $R_{\mathrm{OR}}/R_{\mathrm{DP}}$ (——)

versions of the MSEs

$$R_*(\sigma) := \mathbb{E}\left[\left\|\widehat{f}_{\alpha_*} - f\right\|_{\mathscr{X}}^2\right], \qquad \sigma = 10^{-2}1.3^{-k}, \quad k = 0, \dots, 47$$

with $* \in \{\mathrm{OR}, \mathrm{URE}, \mathrm{LEP}, \mathrm{QO}, \mathrm{DP}\}$. In Fig. 11 we depict the fractions of the oracle risk with the different MSEs for several parameters ν and a to compare the average behavior of these parameter choice methods. The closer the value of such a fraction is to the (optimal) value of 1, the better performs this parameter choice strategy and the smaller c in (22).

In conclusion we empirically find that the quasi-optimality principle performs most stable and is in all investigated situations nearly as good as the oracle choice. This is not clear from the analytical results in Sect. 3.5, as those are limited to mildly ill-posed problems. The Lepskiĭ-type balancing principle also performs well in our simulations, but with a larger constant c in (22). The discrepancy principle behaves comparable. These results are in agreement with the theoretical facts that an oracle inequality is satisfied, and that we do not expect the loss of a logarithmic factor for α_{LEP} to be visible here.

The choice α_{URE} shows a behavior which is harder to interpret. For α_{URE} we observe in most situations that the performance decreases as σ becomes smaller, and this effect is stronger for smoother solutions, whereas the ill-posedness seems to have smaller effect. Especially for smooth solutions f it is questionable if (22) can be satisfied. On the other hand, the theoretical results on empirical risk minimization show that (22) is too ambitious anyway, and that a weaker oracle inequality could still be satisfied.

5 Conclusion

In this study we have investigated four different parameter choice methods in filter based regularization of statistical inverse problems. For the discrepancy principle, unbiased risk estimation, the Lepskiĭ-type balancing principle, and the quasi-optimality principle we have recalled the most important theoretical facts on order optimality and oracle inequalities, and afterwards compared all of them in a simulation study with focus on severely ill-posed operators. It turned out all four seem to perform order optimal in the situations we investigated, with unbiased risk estimation having a higher variance than the others. We also investigated the efficiency in terms of the constant c in an oracle inequality of the form (22). In this simulation, the quasi-optimality principle turned out to be best, followed by the Lepskiĭ-type balancing principle. For unbiased risk estimation it is questionable from our simulations if an oracle inequality of the form (22) is satisfied.

In conclusion, the quasi-optimality principle seems to be the most favorable the parameter choice strategy, as it outperforms the other investigated strategies and is most simple to implement. Nevertheless, the set of regularization parameters needs to be chosen carefully. The second favorable choice seems to be the Lepskiĭ-type balancing principle, which also performs well and very stable in all investigated situations, but at the price of a substantially higher computational effort.

Acknowledgements Financial support by the German Research Foundation DFG through CRC 755, project A07 is gratefully acknowledged.

References

1. A.B. Bakushinskiĭ, Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. USSR Comput. Math. Math. Phys. **24**(4), 181–182 (1984)
2. F. Bauer, T. Hohage, A Lepskij-type stopping rule for regularized Newton methods. Inverse Probl. **21**(6), 1975 (2005)
3. F. Bauer, S. Kindermann, The quasi-optimality criterion for classical inverse problems. Inverse Probl. **24**(3), 035002, 20 pp. (2008)

4. F. Bauer, S. Kindermann, Recent results on the quasi-optimality principle. J. Inverse Ill Posed Probl. **17**(1), 5–18 (2009)
5. F. Bauer, M.A. Lukas, Comparing parameter choice methods for regularization of ill-posed problems. Math. Comput. Simul. **81**(9), 1795–1841 (2011)
6. F. Bauer, S. Pereverzev, Regularization without preliminary knowledge of smoothness and error behaviour. Eur. J. Appl. Math. **16**(3), 303–317 (2005)
7. F. Bauer, M. Reiß, Regularization independent of the noise level: an analysis of quasi-optimality. Inverse Probl. **24**(5), 055009, 16 pp. (2008)
8. S.M.A. Becker, Regularization of statistical inverse problems and the Bakushinskiĭ veto. Inverse Probl. **27**(11), 115010, 22 pp. (2011)
9. N. Bissantz, T. Hohage, A. Munk, F. Ruymgaart, Convergence rates of general regularization methods for statistical inverse problems and applications. SIAM J. Numer. Anal. **45**(6), 2610–2636 (2007)
10. G. Blanchard, P. Mathé, Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. Inverse Probl. **28**(11), 115011, 23 pp. (2012)
11. G. Blanchard, M. Hoffmann, M. Reiß, Optimal adaptation for early stopping in statistical inverse problems, arXiv:1606.07702 (2016)
12. E.J. Candès, Modern statistical estimation via oracle inequalities. Acta Numer. **15**, 257–325 (2006)
13. L. Cavalier, Nonparametric statistical inverse problems. Inverse Probl. **24**(3), 034004, 19 pp. (2008)
14. L. Cavalier, Y. Golubev, Risk hull method and regularization by projections of ill-posed inverse problems. Ann. Stat. **34**(4), 1653–1677 (2006)
15. L. Cavalier, A.B. Tsybakov, Sharp adaptation for inverse problems with random noise. Probab. Theory Relat. Fields **123**(3), 323–354 (2002)
16. L. Cavalier, G.K. Golubev, D. Picard, A.B. Tsybakov, Oracle inequalities for inverse problems. Ann. Stat. **30**(3), 843–874 (2002). Dedicated to the memory of Lucien Le Cam
17. L. Cavalier, Y. Golubev, O. Lepski, A. Tsybakov, Block thresholding and sharp adaptive estimation in severely ill-posed inverse problems. Teor. Veroyatnost. i Primenen. **48**(3), 534–556 (2003)
18. E. Chernousova, Y. Golubev, Spectral cut-off regularizations for ill-posed linear models. Math. Methods Stat. **23**(2), 116–131 (2014)
19. A.R. Davies, R.S. Anderssen, Improved estimates of statistical regularization parameters in Fourier differentiation and smoothing. Numer. Math. **48**(6), 671–697 (1986)
20. D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)
21. B. Efron, Selection criteria for scatterplot smoothers. Ann. Stat. **29**(2), 470–504 (2001)
22. H. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems* (Springer, New York, 1996)
23. W. Freeden, F. Schneider, Regularization wavelets and multiresolution. Inverse Probl. **14**(2), 225–243 (1998)
24. Y. Golubev, The principle of penalized empirical risk in severely ill-posed problems. Probab. Theory Relat. Fields **130**(1), 18–38 (2004)
25. Y. Golubev, On universal oracle inequalities related to high-dimensional linear models. Ann. Stat. **38**(5), 2751–2780 (2010)
26. G.K. Golubev, Exponential weighting and oracle inequalities for projection estimates. Probl. Inf. Transm. **48**(3), 271–282 (2012). Translation of Problemy Peredachi Informatsii **48**(3), 83–95 (2012)
27. U. Hämarik, R. Palm, T. Raus, Comparison of parameter choices in regularization algorithms in case of different information about noise level. Calcolo **48**(1), 47–59 (2011)
28. U. Hämarik, R. Palm, T. Raus, A family of rules for parameter choice in Tikhonov regularization of ill-posed problems with inexact noise level. J. Comput. Appl. Math. **236**(8), 2146–2157 (2012)

29. U. Hämarik, R. Palm, T. Raus, A family of rules for the choice of the regularization parameter in the Lavrentiev method in the case of rough estimate of the noise level of the data. J. Inverse Ill Posed Probl. **20**(5–6), 831–854 (2012)
30. T. Hohage, Regularization of exponentially ill-posed problems. Numer. Funct. Anal. Optim. **21**, 439–464 (2000)
31. T. Hohage, F. Werner, Convergence rates for inverse problems with impulsive noise. SIAM J. Numer. Anal. **52**(3), 1203–1221 (2014)
32. Y. Ingster, T Sapatinas, I.A. Suslina, Minimax signal detection in ill-posed inverse problems. Ann. Stat. **40**(3), 1524–1549 (2012)
33. Y. Ingster, B. Laurent, C. Marteau, Signal detection for inverse problems in a multidimensional framework. Math. Methods Stat. **23**(4), 279–305 (2014)
34. I.M. Johnstone, Wavelet shrinkage for correlated data and inverse problems: adaptivity results. Stat. Sin. **9**(1), 51–83 (1999)
35. I.M. Johnstone, B.W. Silverman, Wavelet threshold estimators for data with correlated noise. J. R. Stat. Soc. Ser. B **59**(2), 319–351 (1997)
36. S. Kindermann, A. Neubauer, On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. Inverse Probl. Imaging **2**(2), 291–299 (2008)
37. A. Kneip, Ordered linear smoothers. Ann. Stat. **22**(2), 835–866 (1994)
38. C. König, F. Werner, T. Hohage, Convergence rates for exponentially ill-posed inverse problems with impulsive noise. SIAM J. Numer. Anal. **54**(1), 341–360 (2016)
39. O.V. Lepskiĭ, On a problem of adaptive estimation in Gaussian white noise. Theory Probab. Appl. **35**(3), 454–466 (1991)
40. O.V. Lepskiĭ, Adaptive estimation over anisotropic functional classes via oracle approach. Ann. Stat. **43**(3), 1178–1242 (2015)
41. K.-C. Li, Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. Ann. Stat. **14**(3), 1101–1112 (1986)
42. K.-C. Li, Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann. Stat. **15**(3), 958–975 (1987)
43. H. Li, F. Werner, Empirical risk minimization as parameter choice rule for general linear regularization methods, arXiv:1703.07809 (2017)
44. S. Lu, P. Mathé, Discrepancy based model selection in statistical inverse problems. J. Complex. **30**(3), 290–308 (2014)
45. M.A. Lukas, Asymptotic optimality of generalized cross-validation for choosing the regularization parameter. Numer. Math. **66**(1), 41–66 (1993)
46. M.A. Lukas, On the discrepancy principle and generalised maximum likelihood for regularisation. Bull. Aust. Math. Soc. **52**(3), 399–424 (1995)
47. M.A. Lukas, Comparisons of parameter choice methods for regularization with discrete noisy data. Inverse Probl. **14**(1), 161–184 (1998)
48. C.L. Mallows, Some comments on C_p. Technometrics **15**(4), 661–675 (1973)
49. P. Mathé, The Lepskiĭ principle revisited. Inverse Probl. **22**(3), L11–L15 (2006)
50. P. Mathé, B. Hofmann, How general are general source conditions? Inverse Probl. **24**(1), 015009, 5 pp. (2008)
51. P. Mathé, S.V. Pereverzev, Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. SIAM J. Numer. Anal. **38**(6), 1999–2021 (2001)
52. P. Mathé, S.V. Pereverzev, Geometry of linear ill-posed problems in variable Hilbert scales. Inverse Probl. **19**(3), 789–803 (2003)
53. P. Mathé, S.V. Pereverzev, Regularization of some linear ill-posed problems with discretized random noisy data. Math. Comput. **75**(256), 1913–1929 (electronic) (2006)
54. V.A. Morozov, On the solution of functional equations by the method of regularization. Sov. Math. Dokl. **7**, 414–417 (1966)
55. A. Neubauer, The convergence of a new heuristic parameter selection criterion for general regularization methods. Inverse Probl. **24**(5), 055005, 10 pp. (2008)

56. D.L. Phillips, A technique for the numerical solution of certain integral equations of the first kind. J. Assoc. Comput. Mach. **9**, 84–97 (1962)
57. A. Rieder, Runge-Kutta integrators yield optimal regularization schemes. Inverse Probl. **21**(2), 453–471 (2005)
58. E. Schock, Approximate solution of ill-posed equations: arbitrarily slow convergence vs. superconvergence. In: *Constructive Methods for the Practical Treatment of Integral Equations (Oberwolfach, 1984)*. Internat. Schriftenreihe Numer. Math., vol. 73 (Birkhäuser, Basel, 1985), pp. 234–243
59. C.M. Stein, Estimation of the mean of a multivariate normal distribution. Ann. Stat. **9**(6), 1135–1151 (1981)
60. A.N. Tikhonov, V.B. Glasko, Use of the regularization method in non-linear problems. USSR Comput. Math. Math. Phys. **5**(3), 93–107 (1965)
61. C.R. Vogel, Optimal choice of a truncation level for the truncated SVD solution of linear first kind integral equations when data are noisy. SIAM J. Numer. Anal. **23**(1), 109–117 (1986)
62. C.R. Vogel, *Computational Methods for Inverse Problems*. Frontiers in Applied Mathematics, vol. 23 (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002)
63. G. Wahba, *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59 (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990)
64. F. Werner, On convergence rates for iteratively regularized Newton-type methods under a Lipschitz-type nonlinearity condition. J. Inverse Ill Posed Probl. **23**(1), 75–84 (2015)
65. F. Werner, T. Hohage, Convergence rates in expectation for Tikhonov-type regularization of inverse problems with Poisson data. Inverse Probl. **28**(10), 104004, 15 pp. (2012)

Relaxing Alternating Direction Method of Multipliers (ADMM) for Linear Inverse Problems

Zehui Wu and Shuai Lu

Abstract We investigate the Alternating Direction Method of Multipliers (ADMM) for solving linear inverse problems. In particular, a relaxing factor is introduced to the standard algorithm allowing more flexible updating of the Lagrange multiplier. The convergence result is established for the Relaxing ADMM for the noise free data under appropriate assumptions. We also calibrate the convergence of the algorithm for the noisy data when noise vanishes by a modified discrepancy principle.

1 Introduction and Preliminaries

We consider a linear inverse problem

$$Ax = b, \quad x \in \mathscr{X}, \tag{1}$$

where A is a bounded linear operator acting from $\mathscr{X}$ to $\mathscr{H}$, b is the observation data and x is the unknown variable. We call b consistent if there exists x such that (1) holds true. Such an inverse problem is usually ill-posed in the sense that the recovery of x does not depend continuously on the data b. In particular, a small perturbation in the observation may lead to huge deviation of the solution, see for example [5, 8, 12].

To solve the ill-posed problem (1) stably, we usually minimize a functional $f(Wx)$ with respect to the constraint such that the following minimization problem is considered

$$\min_x \quad f(Wx) \qquad s.t. \quad Ax = b, \quad x \in \mathscr{D}(W)$$

Z. Wu · S. Lu (✉)
School of Mathematical Sciences, Fudan University, Shanghai 200433, China
e-mail: zehuiwu13@fudan.edu.cn; slu@fudan.edu.cn

B. Hofmann et al. (eds.), *New Trends in Parameter Identification for Mathematical Models*, Trends in Mathematics, https://doi.org/10.1007/978-3-319-70824-9_16

where f is a specific functional described below. By defining an auxiliary variable y, we transfer the above problem into a general form

$$\begin{aligned} &\min_{x,y} \quad f(y) \\ &s.t. \quad Ax = b, \quad Wx = y, \quad x \in \mathscr{D}(W). \end{aligned} \tag{2}$$

We collect following definitions and assumptions before we proceed further.

Definition 1 Let $\mathscr{Y}$ be a Hilbert space. The functional $f : \mathscr{Y} \to (-\infty, +\infty]$ is called proper, if

$$\mathscr{D}(f) := \{y \in \mathscr{Y} : f(y) < \infty\}$$

is not empty.

The functional $f : \mathscr{Y} \to (-\infty, +\infty]$ is called strongly convex, if there exists a constant $c_0 > 0$ such that for any $y_1, y_2 \in \mathscr{Y}$ and $0 \le t \le 1$ the following inequality holds true

$$f(ty_1 + (1-t)y_2) + c_0 t(1-t)\|y_1 - y_2\|^2 \le tf(y_1) + (1-t)f(y_2).$$

□

The main assumption is presented below.

Assumption 1 *Let $\mathscr{X}$, $\mathscr{Y}$, $\mathscr{H}$ be Hilbert spaces. Following items are assumed to be true.*

(I) $A : \mathscr{X} \to \mathscr{H}$ is a linear bounded (or compact) operator, $A^* : \mathscr{H} \to \mathscr{X}$ is its adjoint operator;
(II) $f : \mathscr{Y} \to (-\infty, +\infty]$ is a proper, lower semi-continuous, strongly convex functional;
(III) $\mathscr{D}(W)$ is dense in $\mathscr{X}$ and $W : \mathscr{X} \to \mathscr{Y}$ is a closed linear operator;
(IV) There exists a constant $c_1 > 0$ such that for any $x \in \mathscr{D}(W)$ there holds

$$\|Ax\|^2 + \|Wx\|^2 \ge c_1\|x\|^2.$$

Items (*III*) and (*IV*) in Assumption 1 are standard in classic regularization theory where W is usually considered as a differential operator, c.f. [5, 12]. Different choices of f can be found in [11, 13, 16]. In particular the monograph [15] collects a systematic investigation of different regularization schemes in Banach spaces, where f is not quadratic.

The following theorem shows that under above Assumption 1, the variational problem (2) has a unique solution referring to [10].

Theorem 1 ([10]) *Let the observation data b be consistent and Assumption 1 (I)–(IV) hold true, then the optimization problem (2) has a unique solution $x^* \in \mathscr{D}(W)$ and $Wx^* \in \mathscr{D}(f)$.*

We denote $\partial f(y)$ be the subgradient of the functional f at y defined by

$$\partial f(y) = \{\mu \in \mathscr{Y} : f(\hat{y}) - f(y) - \langle \mu, \hat{y} - y\rangle \geq 0, \quad \forall \hat{y} \in \mathscr{Y}\}.$$

In particular, we need the Bregman distance in [2] between $\hat{y}$ and y which is defined by

$$D_\mu f(\hat{y}, y) = f(\hat{y}) - f(y) - \langle \mu, \hat{y} - y\rangle, \;\; \mu \in \partial f(y).$$

We note that the Bregman distance plays important roles in convergence analysis of regularization schemes for linear inverse problems [3, 9, 14] and the Bregman distance of any strongly convex functional has the following property.

Proposition 1 ([10]) *Let f be a strongly convex functional, μ and $\hat{\mu}$ be the subgradient of f at point y and $\hat{y}$ respectively. Then the following estimates hold true*

$$D_\mu f(\hat{y}, y) \geq c_0 \|\hat{y} - y\|^2,$$

$$\langle \mu - \hat{\mu}, y - \hat{y}\rangle \geq 2c_0 \|y - \hat{y}\|^2.$$

We consider the augmented Lagrangian functional of the minimization problem (2) below

$$\begin{aligned} L_{\rho_1,\rho_2}(x, y; \lambda, \mu) &= f(y) + \langle \lambda, Ax - b\rangle + \langle \mu, Wx - y\rangle \\ &\quad + \frac{\rho_1}{2}\|Ax - b\|^2 + \frac{\rho_2}{2}\|Wx - y\|^2, \end{aligned}$$

where weighted constants ρ_1 and ρ_2 are positive. Referring to the optimality conditions, we investigate the Relaxing ADMM such that

$$\lambda_{k+1} = \lambda_k + \gamma\rho_1(Ax_k - b), \tag{3}$$

$$\mu_{k+1} = \mu_k + \rho_2(Wx_k - y_k), \tag{4}$$

$$x_{k+1} = \arg\min_x L_{\rho_1,\rho_2}(x, y_k; \lambda_{k+1}, \mu_{k+1}), \tag{5}$$

$$y_{k+1} = \arg\min_y L_{\rho_1,\rho_2}(x_{k+1}, y; \lambda_{k+1}, \mu_{k+1}). \tag{6}$$

Compared with the standard ADMM [1], we have introduced a relaxing factor γ in (3) to allow more flexible updating of the Lagrange multiplier λ. In the Relaxing ADMM (3)–(6), the sub-optimization problem (5) is quadratic, whereas the sub-optimization problem (6) can be solved analytically with some prior information of a specific choice of f. We shall mention that another form of ADMM updates firstly the iterates (x_{k+1}, y_{k+1}) then the Lagrange multipliers $(\lambda_{k+1}, \mu_{k+1})$. Its performance in linear inverse problems has been well discussed in [10]. A relaxing ADMM has been investigated in [4] where a simplified case with $W = I$ is considered. Convergence analysis for the Relaxing ADMM (3)–(6) in current work is not trivial compared with the existing ones.

Rest of the paper is organized as follows. In Sect. 2, we collect several lemmas verifying that the sub-optimization problems (5) and (6) are well-posed. In Sect. 3, we prove the convergence of the Relaxing ADMM (3)–(6) with exact data under Assumption 1. In Sect. 4, we calibrate the convergence analysis of the same algorithm for noisy data when noise vanishes. Finally a conclusion Sect. 5 ends the manuscript.

2 Basic Lemmas

In this section, we collect several lemmas which are important to carry out the convergence analysis. We first show that the sub-optimization problems (5)–(6) can be simplified below.

$$\begin{aligned} x_{k+1} &= \arg\min_x \left\{ \frac{\rho_1}{2} \left\| Ax - b + \frac{\lambda_{k+1}}{\rho_1} \right\|^2 + \frac{\rho_2}{2} \left\| Wx - y_k + \frac{\mu_{k+1}}{\rho_2} \right\|^2 \right\}, \\ y_{k+1} &= \arg\min_y \left\{ f(y) + \frac{\rho_2}{2} \left\| y - Wx_{k+1} - \frac{\mu_{k+1}}{\rho_2} \right\|^2 \right\}. \end{aligned} \tag{7}$$

In particular, the above sub-optimization problems are well-defined as shown in [10].

Lemma 1 ([10]) *Let Assumption 1 (I)–(IV) hold true. For any $h \in \mathscr{H}$ and $v \in \mathscr{Y}$ the optimization problem*

$$\min_{z \in \mathscr{D}(W)} \frac{\rho_1}{2} \|Az - h\|^2 + \frac{\rho_2}{2} \|Wz - v\|^2$$

has a unique solution and the variables z and Wz depend continuously on h and v.

Lemma 2 ([10]) *Let Assumption 1 (I)–(IV) hold true. The following optimization problem*

$$\min_{y \in \mathscr{Y}} L_2(y) = f(y) + \frac{\rho_2}{2} \|y - v\|^2$$

has a unique solution y and the variables y and $f(y)$ depend continuously on v.

By taking the Euler equations of (7), we obtain

$$\begin{aligned} &A^*[\lambda_{k+1} + \rho_1(Ax_{k+1} - b)] + W^*[\mu_{k+1} + \rho_2(Wx_{k+1} - y_k)] = 0, \\ &\mu_{k+2} = \mu_{k+1} + \rho_2(Wx_{k+1} - y_{k+1}) \in \partial f(y_{k+1}). \end{aligned} \tag{8}$$

Notice that the functional *f* is strongly convex and we have the following inequality,

$$\langle \mu_{k+2} - \mu_{k+1}, y_{k+1} - y_k \rangle \geq 2c_0 \| y_{k+1} - y_k \|^2. \tag{9}$$

Moreover, we define the following residuals

$$r_k = Ax_k - b, \quad s_k = Wx_k - y_k. \tag{10}$$

The following lemma is essential to carry out the convergence analysis of the Relaxing ADMM.

Lemma 3 *Let k be the iteration of the Relaxing ADMM. If $k = 1$, there holds*

$$A^*\lambda_2 = (\gamma - 1)\rho_1 A^* r_1 + \rho_2 W^*(y_0 - y_1) - W^*\mu_2.$$

For any $k \geq 2$, there holds

$$\rho_1 A^* r_k + (\gamma - 1)\rho_1 A^* r_{k-1} = \rho_2 W^*[(y_{k-1} - y_k) - (y_{k-2} - y_{k-1})] - \rho_2 W^* s_k.$$

Proof The proof follows directly after the sub-optimization problems (3)–(4) and the optimality condition (8) such that

$$\begin{aligned} A^*\lambda_{k+1} =& A^*\lambda_k + \gamma\rho_1 A^* r_k + W^*(\rho_2 s_k + \mu_k - \mu_{k+1}) \\ =& -\rho_1 A^* r_k - W^*[\mu_k + \rho_2(Wx_k - y_{k-1})] + \gamma\rho_1 A^* r_k \\ &+ W^*(\rho_2 s_k + \mu_k - \mu_{k+1}) \\ =& (\gamma - 1)\rho_1 A^* r_k + \rho_2 W^*(y_{k-1} - y_k) - W^*\mu_{k+1}. \end{aligned}$$

Let $k = 1$, we obtain the first equality. By implementing (3) and the above equality, we derive

$$\begin{aligned} \gamma\rho_1 A^* r_k &= A^*\lambda_{k+1} - A^*\lambda_k \\ &= (\gamma - 1)\rho_1 A^*(r_k - r_{k-1}) \\ &+ \rho_2 W^*[(y_{k-1} - y_k) - (y_{k-2} - y_{k-1})] - W^*(\mu_{k+1} - \mu_k), \end{aligned}$$

which yields the second equality referring to (8) and (10). □

Lemma 4 *Denote $E_k = \rho_1\gamma\|r_k\|^2 + \rho_2\|s_k\|^2 + \rho_2\|y_k - y_{k-1}\|^2$. For any $\gamma \in (0, 2)$, E_k is monotonically non-increasing such that*

$$E_{k+1} - E_k \leq -\rho_1(2 - \gamma)\|r_{k+1} - r_k\|^2 - 4c_0\|y_{k+1} - y_k\|^2 \leq 0,$$

and

$$\sum_{k=1}^{\infty} \| y_k - y_{k+1} \|^2 < \infty.$$

Proof By the definition in (10), we have

$$r_{k+1} - r_k = A(x_{k+1} - x_k), \qquad s_{k+1} - s_k = Wx_{k+1} - Wx_k + y_k - y_{k+1}.$$

Implement Lemma 3, we have the following equality,

$$\begin{aligned}
&\rho_1 \langle r_{k+1} - r_k, r_{k+1} \rangle + \rho_2 \langle s_{k+1} - s_k, s_{k+1} \rangle \\
=&\rho_1 \langle A(x_{k+1} - x_k), r_{k+1} \rangle + \rho_2 \langle Wx_{k+1} - Wx_k + (y_k - y_{k+1}), s_{k+1} \rangle \\
=&\langle x_{k+1} - x_k, (1-\gamma)\rho_1 A^* r_k \rangle + \langle x_{k+1} - x_k, \rho_2 W^* [(y_k - y_{k+1}) - (y_{k-1} - y_k)] \rangle \\
&+ \rho_2 \langle y_k - y_{k+1}, s_{k+1} \rangle.
\end{aligned}$$

From the inequality (9) we derive,

$$\begin{aligned}
&\rho_1 \langle r_{k+1} - r_k, r_{k+1} \rangle + \rho_2 \langle s_{k+1} - s_k, s_{k+1} \rangle + (\gamma - 1)\rho_1 \langle r_{k+1} - r_k, r_k \rangle \\
\leq&\rho_2 \langle s_{k+1} - s_k + y_{k+1} - y_k, (y_k - y_{k+1}) - (y_{k-1} - y_k) \rangle - 2c_0 \| y_{k+1} - y_k \|^2.
\end{aligned}$$

Following three equalities are straight-forward

$$\begin{aligned}
\rho_1 \langle r_{k+1} - r_k, r_{k+1} \rangle &= \frac{\rho_1}{2} (\|r_{k+1}\|^2 - \|r_k\|^2 + \|r_{k+1} - r_k\|^2), \\
\rho_2 \langle s_{k+1} - s_k, s_{k+1} \rangle &= \frac{\rho_2}{2} (\|s_{k+1}\|^2 - \|s_k\|^2 + \|s_{k+1} - s_k\|^2), \\
\rho_1 \langle r_{k+1} - r_k, r_k \rangle &= \frac{\rho_1}{2} (-\|r_k\|^2 - \|r_k - r_{k+1}\|^2 + \|r_{k+1}\|^2).
\end{aligned}$$

The Cauchy–Schwarz inequality further yields

$$\begin{aligned}
&\rho_2 \langle (s_{k+1} - s_k) + (y_{k+1} - y_k), (y_k - y_{k+1}) - (y_{k-1} - y_k) \rangle \\
\leq&\frac{\rho_2}{2} \|s_{k+1} - s_k\|^2 + \frac{\rho_2}{2} \|(y_k - y_{k+1}) - (y_{k-1} - y_k)\|^2 \\
&- \rho_2 \| y_{k+1} - y_k \|^2 + \rho_2 \langle y_k - y_{k+1}, y_{k-1} - y_k \rangle \\
=&\frac{\rho_2}{2} \|s_{k+1} - s_k\|^2 - \frac{\rho_2}{2} \| y_{k+1} - y_k \|^2 + \frac{\rho_2}{2} \| y_k - y_{k-1} \|^2.
\end{aligned}$$

The lemma is then proven by considering above equalities and inequalities. □

Meanwhile, the following technical lemma is helpful and will be recalled occasionally in the context.

Lemma 5 *Let* $s_k = Wx_k - y_k$ *and integers* $m < n$. *For any* $\xi > 0$, *there holds*

$$-\langle y_{n-1} - y_n, W(\hat{x} - x_n)\rangle \le \frac{1}{4}\|W(\hat{x} - x_m)\|^2 + \xi(\|s_m\|^2 + \|s_n\|^2) + \frac{1}{2\xi}\sum_{k=m}^{n-1}\| y_k - y_{k+1}\|^2 + \frac{\xi}{2}(n-m-1)\| y_{n-1} - y_n\|^2. \tag{11}$$

Proof The proof is a direct consequence of the Cauchy–Schwarz inequality such that

$$\begin{aligned}
&- \langle y_{n-1} - y_n, W(\hat{x} - x_n)\rangle \\
=& - \langle y_{n-1} - y_n, W(\hat{x} - x_m)\rangle - \sum_{k=m}^{n-1}\langle y_{n-1} - y_n, W(x_k - x_{k+1})\rangle \\
\le& \frac{1}{4}\|W(\hat{x} - x_m)\|^2 + \| y_{n-1} - y_n\|^2 - \sum_{k=m}^{n-1}\langle y_{n-1} - y_n, s_k + y_k - y_{k+1} - s_{k+1}\rangle \\
=& \frac{1}{4}\|W(\hat{x} - x_m)\|^2 - \langle y_{n-1} - y_n, s_m - s_n\rangle - \sum_{k=m}^{n-2}\langle y_{n-1} - y_n, y_k - y_{k+1}\rangle \\
\le& \frac{1}{4}\|W(\hat{x} - x_m)\|^2 + \xi(\|s_m\|^2 + \|s_n\|^2) + \frac{1}{2\xi}\sum_{k=m}^{n-1}\|y_k - y_{k+1}\|^2 + \frac{\xi}{2}(n-m-1)\|y_{n-1} - y_n\|^2.
\end{aligned}$$

□

3 Convergence Analysis with Exact Data

We proceed to the convergence analysis of the Relaxing ADMM with exact data. First lemma considers the asymptotic behavior of the Bregman distance between neighbouring iterates and E_k.

Lemma 6 *Let* $\gamma \in (0, 2)$, *the minimizing sequences* $\{x_k\}$ *and* $\{ y_k\}$ *are bounded and satisfies*

$$\sum_{k=1}^{\infty}\left(D_{\mu_{k+1}} f(y_{k+1}, y_k) + E_k\right) < \infty.$$

Proof Denote $(\hat{x}, \hat{y})$ be a feasible point of the optimization problem (2). By the Bregman distance and (4), we have

$$\begin{aligned}
&D_{\mu_{k+2}} f(\hat{y}, y_{k+1}) - D_{\mu_{k+1}} f(\hat{y}, y_k) + D_{\mu_{k+1}} f(y_{k+1}, y_k) \\
&= \langle \mu_{k+1} - \mu_{k+2}, \hat{y} - y_{k+1} \rangle = -\rho_2 \langle s_{k+1}, W\hat{x} - y_{k+1} \rangle \\
&= -\rho_2 \langle s_{k+1}, W(\hat{x} - x_{k+1}) + s_{k+1} \rangle = -\rho_2 \|s_{k+1}\|^2 - \rho_2 \langle s_{k+1}, W(\hat{x} - x_{k+1}) \rangle.
\end{aligned}$$

Implement Lemma 3, we derive

$$\begin{aligned}
&D_{\mu_{k+2}} f(\hat{y}, y_{k+1}) - D_{\mu_{k+1}} f(\hat{y}, y_k) + D_{\mu_{k+1}} f(y_{k+1}, y_k) \\
= & - \rho_2 \|s_{k+1}\|^2 - \langle \hat{x} - x_{k+1}, \rho_2 W^* [(y_k - y_{k+1}) - (y_{k-1} - y_k)] \rangle \\
& + \langle \hat{x} - x_{k+1}, \rho_1 A^* r_{k+1} + (\gamma - 1) \rho_1 A^* r_k \rangle \\
= & - \rho_2 \|s_{k+1}\|^2 - \rho_1 \|r_{k+1}\|^2 - (\gamma - 1) \rho_1 \langle r_k, r_{k+1} \rangle \\
& + \rho_2 \langle y_{k-1} - y_k, W(\hat{x} - x_{k+1}) \rangle - \rho_2 \langle y_k - y_{k+1}, W(\hat{x} - x_{k+1}) \rangle.
\end{aligned} \tag{12}$$

We sum up both sides of (12) and obtain

$$\begin{aligned}
&D_{\mu_{n+1}} f(\hat{y}, y_n) - D_{\mu_{m+1}} f(\hat{y}, y_m) + \sum_{k=m}^{n-1} D_{\mu_{k+1}} f(y_{k+1}, y_k) \\
= & - \sum_{k=m+1}^{n} \left(\rho_2 \|s_k\|^2 + \rho_1 \|r_k\|^2 \right) - (\gamma - 1) \rho_1 \sum_{k=m}^{n-1} \langle r_k, r_{k+1} \rangle \\
& + \rho_2 \sum_{k=m-1}^{n-2} \langle y_k - y_{k+1}, W(\hat{x} - x_{k+2}) \rangle - \rho_2 \sum_{k=m}^{n-1} \langle y_k - y_{k+1}, W(\hat{x} - x_{k+1}) \rangle \\
= & - \sum_{k=m+1}^{n} \left(\rho_2 \|s_k\|^2 + \rho_1 \|r_k\|^2 \right) - (\gamma - 1) \rho_1 \sum_{k=m}^{n-1} \langle r_k, r_{k+1} \rangle \\
& + \rho_2 \sum_{k=m}^{n-2} \langle y_k - y_{k+1}, W(x_{k+1} - x_{k+2}) \rangle + \rho_2 \langle y_{m-1} - y_m, W(\hat{x} - x_{m+1}) \rangle \\
& - \rho_2 \langle y_{n-1} - y_n, W(\hat{x} - x_n) \rangle.
\end{aligned} \tag{13}$$

Let $m = 1$, we thus have

$$D_{\mu_{n+1}} f(\hat{y}, y_n) + \sum_{k=1}^{n-1} D_{\mu_{k+1}} f(y_{k+1}, y_k)$$

$$
\begin{aligned}
=& D_{\mu_2} f(\hat{y}, y_1) - \sum_{k=2}^{n} (\rho_2 \|s_k\|^2 + \rho_1 \|r_k\|^2) - (\gamma - 1)\rho_1 \sum_{k=1}^{n-1} \langle r_k, r_{k+1} \rangle \\
& + \rho_2 \langle y_0 - y_1, W(\hat{x} - x_2) \rangle + \rho_2 \sum_{k=1}^{n-2} \langle y_k - y_{k+1}, W(x_{k+1} - x_{k+2}) \rangle \\
& - \rho_2 \langle y_{n-1} - y_n, W(\hat{x} - x_n) \rangle . \qquad (14)
\end{aligned}
$$

We provide the error estimates for three items appearing in the right-hand side of (14) below

$$
\sum_{k=1}^{n-1} |\langle r_k, r_{k+1} \rangle| \le \sum_{k=1}^{n-1} \frac{\|r_k\|^2 + \|r_{k+1}\|^2}{2} \le \sum_{k=1}^{n} \|r_k\|^2, \qquad (15)
$$

$$
\begin{aligned}
& \sum_{k=1}^{n-2} \langle y_k - y_{k+1}, W(x_{k+1} - x_{k+2}) \rangle \\
& \quad = \sum_{k=1}^{n-2} \langle y_k - y_{k+1}, s_{k+1} + (y_{k+1} - y_{k+2}) - s_{k+2} \rangle \\
& \quad \le \sum_{k=1}^{n-2} \| y_k - y_{k+1} \| (\|s_{k+1}\| + \| y_{k+1} - y_{k+2} \| + \|s_{k+2}\|) \\
& \quad \le \sum_{k=1}^{n-2} \left(\frac{1}{8} \|s_{k+1}\|^2 + \frac{1}{8} \|s_{k+2}\|^2 + \frac{9}{2} \| y_k - y_{k+1} \|^2 + \frac{1}{2} \| y_{k+1} - y_{k+2} \|^2 \right) \\
& \quad \le \frac{1}{4} \sum_{k=2}^{n} \|s_k\|^2 + 5 \sum_{k=1}^{n-1} \| y_k - y_{k+1} \|^2, \qquad (16)
\end{aligned}
$$

and

$$
\begin{aligned}
-\langle y_{n-1} - y_n, W(\hat{x} - x_n) \rangle \le & \frac{1}{4} \|W(\hat{x} - x_1)\|^2 + \frac{1}{4} \left(\|s_1\|^2 + \|s_n\|^2 \right) \\
& + 2 \sum_{k=1}^{n-1} \| y_k - y_{k+1} \|^2 + \frac{n}{8} \| y_{n-1} - y_n \|^2 \qquad (17)
\end{aligned}
$$

by letting $m = 1, \xi = \frac{1}{4}$ in Lemma 5.

With the aid of above (in)equalities (14)–(17), we have the following estimate

$$\begin{aligned}
D_{\mu_{n+1}}f(\hat{y}, y_n) + \sum_{k=1}^{n-1} D_{\mu_{k+1}}f(y_{k+1}, y_k) \le{}& C - \sum_{k=2}^{n}(\rho_1\|r_k\|^2 + \rho_2\|s_k\|^2)\\
&+ |\gamma - 1|\rho_1 \sum_{k=1}^{n}\|r_k\|^2\\
&+ \frac{\rho_2}{4}\sum_{k=2}^{n}\|s_k\|^2 + 5\rho_2\sum_{k=1}^{n-1}\|y_k - y_{k+1}\|^2 + \frac{\rho_2}{4}\|s_n\|^2\\
&+ 2\rho_2\sum_{k=1}^{n-1}\|y_k - y_{k-1}\|^2 + \frac{n\rho_2}{8}\|y_{n-1} - y_n\|^2\\
\le{}& C - \rho_1(1 - |\gamma - 1|)\sum_{k=2}^{n}\|r_k\|^2 - \frac{\rho_2}{2}\sum_{k=2}^{n}\|s_k\|^2\\
&+ 7\rho_2\sum_{k=1}^{n-1}\|y_k - y_{k+1}\|^2 + \frac{n\rho_2}{8}\|y_{n-1} - y_n\|^2 \qquad (18)
\end{aligned}$$

where the constant C does not depend on n.

At the same time, Lemma 4 has shown that

$$\sum_{k=1}^{\infty}\|y_k - y_{k+1}\|^2 < \infty.$$

We can find a subsequence $n_j \to \infty$ such that

$$n_j\|y_{n_j} - y_{n_j+1}\|^2 \to 0 \quad (j \to \infty).$$

Therefore for arbitrary $0 < \gamma < 2$, we have

$$\sum_{k=1}^{n_j-1} D_{\mu_{k+1}}f(y_{k+1}, y_k) + \rho_1(1 - |1 - \gamma|)\sum_{k=2}^{n_j}\|r_k\|^2 + \frac{\rho_2}{2}\sum_{k=2}^{n_j}\|s_k\|^2 \le C.$$

Let $j \to \infty$,

$$\sum_{k=1}^{\infty}\left(D_{\mu_{k+1}}f(y_{k+1}, y_k) + E_k\right) < \infty.$$

By Lemma 4, we know that $\{E_k\}$ is monotonically non-increasing, thus

$$n\rho_2\| y_n - y_{n-1}\|^2 \leq nE_n \leq \sum_{k=1}^{n} E_k < C,$$

and

$$\lim_{n\to\infty} \| y_n - y_{n-1}\| = 0, \quad \lim_{n\to\infty} E_n = 0.$$

By (18) and the strong convexity of f, we derive

$$C \geq D_{\mu_{n+1}}f(\hat{y}, y_n) \geq c_0\|\hat{y} - y_n\|^2.$$

Therefore $\{ y_n\}$ is a bounded sequence. Moreover, because $\lim_{n\to\infty} E_n = 0$, we further have

$$\lim_{n\to\infty} Ax_n = b, \quad \lim_{n\to\infty} (Wx_n - y_n) = 0.$$

Since $\{Ax_n\}$ and $\{Wx_n\}$ are bounded sequences, by Assumption 1 (*IV*),

$$c_1\|x_n\|^2 \leq \|Ax_n\|^2 + \|Wx_n\|^2 < \infty,$$

we prove that $\{x_n\}$ is also a bounded sequence. □

Lemma 7 *Denote* $(\hat{x}, \hat{y})$ *be a feasible point of (2) and let* $\gamma \in (0, 2)$. *Then the Bregman distance sequence* $\{D_{\mu_{k+1}}f(\hat{y}, y_k)\}$ *converges.*

Proof We consider (13) and obtain

$$\begin{aligned}
&\left|D_{\mu_{n+1}}f(\hat{y}, y_n) - D_{\mu_{m+1}}f(\hat{y}, y_m)\right| \\
\leq &\sum_{k=m}^{n-1} D_{\mu_{k+1}}f(y_{k+1}, y_k) + \sum_{k=m+1}^{n} (\rho_1\|r_k\|^2 + \rho_2\|s_k\|^2) \\
&+ \rho_2\left|\sum_{k=m}^{n-2}\langle y_k - y_{k+1}, W(x_{k+1} - x_{k+2})\rangle\right| + \rho_2|\langle y_{m-1} - y_m, W(\hat{x} - x_{m+1})\rangle| \\
&+ \rho_2|\langle y_{n-1} - y_n, W(\hat{x} - x_n)\rangle| + |1 - \gamma|\rho_1\sum_{k=m}^{n-1}|\langle r_k, r_{k+1}\rangle|.
\end{aligned} \tag{19}$$

Notice that

$$\left|\sum_{k=m}^{n-2}\langle y_k - y_{k+1}, W(x_{k+1} - x_{k+2})\rangle\right| \leq \frac{1}{4}\sum_{k=m+1}^{n} \|s_k\|^2 + 5\sum_{k=m}^{n-1} \| y_k - y_{k+1}\|^2.$$

Substitute it into (19), we obtain

$$\begin{aligned}&\left|D_{\mu_{n+1}}f(\hat{y},y_n)-D_{\mu_{m+1}}f(\hat{y},y_m)\right|\\ \le&\sum_{k=m}^{n-1}D_{\mu_{k+1}}f(y_{k+1},y_k)+\sum_{k=m+1}^{n}(\rho_1\|r_k\|^2+\rho_2\|s_k\|^2)+|1-\gamma|\rho_1\sum_{k=m}^{n}\|r_k\|^2\\ &+\frac{\rho_2}{4}\sum_{k=m+1}^{n}\|s_k\|^2+5\rho_2\sum_{k=m}^{n-1}\|y_k-y_{k+1}\|^2\\ &+\rho_2\|y_{m-1}-y_m\|\|W(\hat{x}-x_{m+1})\|+\rho_2\|y_{n-1}-y_n\|\|W(\hat{x}-x_n)\|.\end{aligned}\tag{20}$$

By Lemma 6, we verify that the right hand side of (20) tends to 0 when $m,n\to\infty$ such that

$$\lim_{m,n\to\infty}|D_{\mu_{n+1}}f(\hat{y},y_n)-D_{\mu_{m+1}}f(\hat{y},y_m)|=0.$$

□

We thus prove convergence of the Relaxing ADMM.

Theorem 2 *Let Assumption 1 (I)–(IV) hold true, $\gamma\in\left[\frac{2}{3},2\right)$ and the observation data b be consistent. Denote $x^*, y^*=Wx^*$ be the exact solution. Then the Relaxing ADMM (3)–(6) converges such that*

$$\begin{gathered}x_k\to x^*,\quad y_k\to y^*,\quad Wx_k\to y^*,\\ f(y_k)\to f(y^*),\quad D_{\mu_{k+1}}f(y^*,y_k)\to 0,\quad (k\to\infty).\end{gathered}$$

Proof We first prove $\{y_k\}$ is a Cauchy sequence. Assume $(\hat{x},\hat{y})$ is a feasible point, we have the following equality

$$D_{\mu_{m+1}}f(y_n,y_m)-D_{\mu_{m+1}}f(\hat{y},y_m)+D_{\mu_{n+1}}f(\hat{y},y_n)=\langle\mu_{n+1}-\mu_{m+1},y_n-\hat{y}\rangle.\tag{21}$$

From (4) we have,

$$\begin{aligned}\langle\mu_{n+1}-\mu_{m+1},y_n-\hat{y}\rangle&=\sum_{k=m+1}^{n}\langle\mu_{k+1}-\mu_k,y_n-\hat{y}\rangle\\ &=\rho_2\sum_{k=m+1}^{n}\langle s_k,y_n-\hat{y}\rangle\\ &=-\rho_2\sum_{k=m+1}^{n}\langle s_k,s_n\rangle+\rho_2\sum_{k=m+1}^{n}\langle s_k,W(x_n-\hat{x})\rangle.\end{aligned}$$

By using Lemma 3, we have

$$\begin{aligned}
&|\langle \mu_{n+1} - \mu_{m+1}, y_n - \hat{y}\rangle| \\
=&\left| -\rho_2 \sum_{k=m+1}^{n} \langle s_k, s_n\rangle + \sum_{k=m+1}^{n} \langle \rho_2 W^*[(y_{k-1} - y_k) - (y_{k-2} - y_{k-1})], x_n - \hat{x}\rangle \right. \\
&\left. - \sum_{k=m+1}^{n} \langle \rho_1 A^* r_k + (\gamma - 1)\rho_1 A^* r_{k-1}, x_n - \hat{x}\rangle \right| \\
=&\left| -\rho_2 \sum_{k=m+1}^{n} \langle s_k, s_n\rangle + \rho_2 \langle y_m - y_n - y_{m-1} + y_{n-1}, W(x_n - \hat{x})\rangle \right. \\
&\left. -\rho_1 \sum_{k=m+1}^{n} \langle r_k, r_n\rangle - (\gamma - 1)\rho_1 \sum_{k=m+1}^{n} \langle r_{k-1}, r_n\rangle \right| \\
\leq&\frac{\rho_2}{2} \sum_{k=m+1}^{n} (\|s_k\|^2 + \|s_n\|^2) + \rho_2 |\langle y_{n-1} - y_n, W(x_n - \hat{x})\rangle| \\
&+ \rho_2 |\langle y_{m-1} - y_m, W(x_n - \hat{x})\rangle| + \frac{\rho_1}{2} \sum_{k=m+1}^{n} (\|r_k\|^2 + \|r_n\|^2) \\
&+ |1 - \gamma| \frac{\rho_1}{2} \sum_{k=m+1}^{n} (\|r_{k-1}\|^2 + \|r_n\|^2) \\
\leq&\frac{\rho_2}{2} \sum_{k=m+1}^{n} \|s_k\|^2 + \frac{\rho_1(1 + |1 - \gamma|)}{2} \sum_{k=m}^{n} \|r_k\|^2 + \frac{\rho_2(n - m)}{2} \|s_n\|^2 \\
&+ \frac{\rho_1}{2}(1 + |1 - \gamma|)(n - m)\|r_n\|^2 + \rho_2 |\langle y_{n-1} - y_n, W(x_n - \hat{x})\rangle| \\
&+ \rho_2 |\langle y_{m-1} - y_m, W(x_n - \hat{x})\rangle|. \qquad (22)
\end{aligned}$$

Notice that $\gamma \in [2/3, 2)$ allows

$$\begin{aligned}
\left|\langle \mu_{n+1} - \mu_{m+1}, y_n - \hat{y}\rangle\right| \leq& 2 \sum_{k=m+1}^{n} E_k + \rho_2 \left|\langle y_{n-1} - y_n, W(x_n - \hat{x})\rangle\right| \\
&+ \rho_2 |\langle y_{m-1} - y_m, W(x_n - \hat{x})\rangle|. \qquad (23)
\end{aligned}$$

Implement Lemma 6 and (23), we have

$$\langle \mu_{n+1} - \mu_{m+1}, y_n - \hat{y}\rangle \to 0 \quad (m, n \to \infty). \qquad (24)$$

Lemma 7 and (21) further yield

$$|D_{\mu_{m+1}} f(y_n, y_m)| \to 0 \quad (m, n \to \infty)$$

and the property of Bregman distance shows $\lim_{m,n\to\infty} \| y_n - y_m\| = 0$. Thus $\{ y_n\}$ is a Cauchy sequence and $y_n \to \tilde{y}(n \to \infty)$.

Secondly we prove that there exists $\tilde{x} \in \mathscr{D}(W)$ such that

$$x_k \to \tilde{x} \quad (k \to \infty),$$
$$A\tilde{x} = b, \quad W\tilde{x} = \tilde{y}.$$

From Lemma 6 and $y_k \to \tilde{y}$, we have

$$Wx_k \to \tilde{y}, \quad Ax_k \to b.$$

Furthermore, by using Assumption 1 (IV) we can prove that $\{x_k\}$ is a Cauchy sequence and

$$x_n \to \tilde{x} \quad (n \to \infty).$$

Thus we have $b = \lim_{n\to\infty} Ax_n = A\tilde{x}$. Notice that W is a closed operator in Assumption 1 (III), there holds

$$\tilde{x} \in \mathscr{D}(W), \; W\tilde{x} = \tilde{y}.$$

Next, we prove that

$$\tilde{y} \in \mathscr{D}(f), \quad \lim_{k\to\infty} f(y_k) = f(\tilde{y}), \quad \lim_{k\to\infty} D_{\mu_{k+1}} f(\tilde{y}, y_k) = 0.$$

For any feasible point $(\hat{x}, \hat{y})$ we have,

$$f(\hat{y}) + \langle \mu_{k+1}, y_k - \hat{y}\rangle \geq f(y_k). \tag{25}$$

Since f is a proper function and y_k is a solution of optimization problem (7), $f(y_k)$ is finite. Because of the lower semi-continuity, we further obtain

$$f(\tilde{y}) \leq \liminf_{k\to\infty} f(y_k) < \infty.$$

Therefore $\tilde{y} \in \mathscr{D}(f)$.

Since $(\tilde{x}, \tilde{y})$ is a feasible point, we replace $(\hat{x}, \hat{y})$ by $(\tilde{x}, \tilde{y})$ and re-consider (23),

$$\limsup_{k\to\infty} |\langle \mu_{k+1}, y_k - \tilde{y}\rangle| \leq \sum_{i=m+1}^{\infty} E_i.$$

Let $m \to \infty$, we have $|\langle \mu_{k+1}, y_k - \tilde{y}\rangle| \to 0\ (k \to \infty)$, hence, $\limsup_{k\to\infty} f(y_k) \leq f(\tilde{y})$. By the lower semi-continuity, we have $\lim_{k\to\infty} f(y_k) = f(\tilde{y})$ which yields

$$\lim_{k\to\infty} D_{\mu_{k+1}} f(\tilde{y}, y_k) = 0.$$

Finally, we prove that $(\tilde{x}, \tilde{y})$ is the exact solution (x^*, y^*). For any $\epsilon > 0$, referring to (24), there exists $k_0 > 0$ which satisfies

$$\begin{aligned} |\langle \mu_{k+1} - \mu_{k_0+1}, y_k - \hat{y}\rangle| &\leq \epsilon, \\ \rho_2 |\langle y_{k_0-1} - y_{k_0}, W(x_k - \hat{x})\rangle| &\leq \epsilon,\ \forall k > k_0. \end{aligned} \tag{26}$$

Then (25) yields

$$f(y_k) \leq f(\hat{y}) + \epsilon + \langle \mu_{k_0+1}, y_k - \hat{y}\rangle.$$

However,

$$\begin{aligned} \langle \mu_{k_0+1}, y_k - \hat{y}\rangle &= -\langle \mu_{k_0+1}, s_k\rangle + \langle \mu_{k_0+1}, W(x_k - \hat{x})\rangle \\ &= -\langle \mu_{k_0+1}, s_k\rangle + \langle \mu_2, W(x_k - \hat{x})\rangle + \rho_2 \sum_{i=2}^{k_0} \langle s_i, W(x_k - \hat{x})\rangle. \end{aligned}$$

From Lemma 3, we have

$$\begin{aligned} \langle \mu_{k_0+1}, y_k - \hat{y}\rangle =& - \langle \mu_{k_0+1}, s_k\rangle - \langle \lambda_2, A(x_k - \hat{x})\rangle + (\gamma - 1)\rho_1 \langle r_1, r_k\rangle \\ & + \rho_2 \langle y_0 - y_1, W(x_k - \hat{x})\rangle - \rho_1 \sum_{i=2}^{k_0} \langle r_i, r_k\rangle - (\gamma - 1)\rho_1 \sum_{i=2}^{k_0} \langle r_{i-1}, r_k\rangle \\ & + \rho_2 \sum_{i=2}^{k_0} \langle (y_{i-1} - y_i) - (y_{i-2} - y_{i-1}), W(x_k - \hat{x})\rangle \\ =& - \langle \mu_{k_0+1}, s_k\rangle - \langle \lambda_2, r_k\rangle + (\gamma - 1)\rho_1 \langle r_1, r_k\rangle \\ & - \rho_1 \sum_{i=2}^{k_0} \langle r_i, r_k\rangle - (\gamma - 1)\rho_1 \sum_{i=2}^{k_0} \langle r_{i-1}, r_k\rangle \\ & + \rho_2 \langle y_{k_0-1} - y_{k_0}, W(x_k - \hat{x})\rangle. \end{aligned} \tag{27}$$

Combine (26) and (27), we obtain

$$\begin{aligned} |\langle \mu_{k_0+1}, y_k - \hat{y}\rangle| \leq& \|\mu_{k_0+1}\| \|s_k\| + \|\lambda_2\| \|r_k\| + |1 - \gamma| \rho_1 \|r_1\| \|r_k\| \\ & + (1 + |\gamma - 1|)\rho_1 (\sum_{i=1}^{k_0} \|r_i\|) \|r_k\| + \epsilon. \end{aligned}$$

From Lemma 6, we derive

$$\limsup_{k\to\infty} f(y_k) \le f(\hat{y}) + 2\epsilon.$$

Because f is lower semi-continuous, we thus have

$$f(\tilde{y}) \le \liminf_{k\to\infty} f(y_k) \le f(\hat{y}) + 2\epsilon.$$

Let $\epsilon \to 0$, then $f(\tilde{y}) \le f(\hat{y})$. Because of the uniqueness in Theorem 1, we thus conclude

$$\tilde{x} = x^*, \quad \tilde{y} = y^*.$$

□

We shall mention that the interval $\gamma \in \left[\frac{2}{3}, 2\right)$ in Theorem 2 can be released to $\gamma \in [\zeta, 2)$ for any positive constant $\zeta < 2/3$ such that the right-hand of (22) can be bounded by some constant, depending on ζ, multiplying the right-hand of (23). We consider a simple interval $\gamma \in \left[\frac{2}{3}, 2\right)$ here just illustrating the role of the relaxing factor.

4 Convergence Analysis with Noisy Data

Notice that the observation data b is usually contaminated by some noise, careful calibration of the Relaxing ADMM (3)–(6) shall be considered when noisy data b^δ is known instead of the exact one. Moreover, we shall assume that there is a noisy level δ satisfying

$$\|b^\delta - b\| \le \delta.$$

We introduce the residual

$$r_k^\delta = Ax_k^\delta - b^\delta, \quad s_k^\delta = Wx_k^\delta - y_k^\delta$$

and present the algorithm for noisy data as below.

Algorithm 1 (Relaxing ADMM for Noisy Data)

Given the forward operator A and the noisy data b^δ.

(a) Set initial guesses $x_0 \in \mathscr{D}(W)$, $y_0 \in \mathscr{Y}$, $\lambda_0 \in \mathscr{H}$, $\mu_0 \in \mathscr{Y}$, constants $\rho_1 > 0$, $\rho_2 > 0$ and the relaxing parameter $\gamma > 0$.

(b) *Let* $k = 0$, $x_0^\delta = x_0$, $y_0^\delta = y_0$, $\lambda_0^\delta = \lambda_0$, $\mu_0^\delta = \mu_0$.

(c) *If the stopping criterion is satisfied, with a constant* $\tau > 1$,

$$\gamma\rho_1^2\|r_k^\delta\|^2 + \rho_2^2\|s_k^\delta\|^2 \le \max(\rho_1^2, \rho_2^2)\tau^2\delta^2, \tag{28}$$

then the algorithm terminates.

(d) *Update the Lagrange multipliers* λ, μ *and solutions* x, y *by*

$$\begin{aligned}
\lambda_{k+1}^\delta &= \lambda_k^\delta + \gamma\rho_1(Ax_k^\delta - b^\delta);\\
\mu_{k+1}^\delta &= \mu_k^\delta + \rho_2(Wx_k^\delta - y_k^\delta);\\
x_{k+1}^\delta &= \underset{x\in\mathscr{D}(W)}{\arg\min}\langle\lambda_{k+1}^\delta, Ax\rangle + \langle\mu_{k+1}^\delta, Wx\rangle + \frac{\rho_1}{2}\|Ax - b^\delta\|^2 + \frac{\rho_2}{2}\|Wx - y_k^\delta\|^2;\\
y_{k+1}^\delta &= \underset{y\in\mathscr{Y}}{\arg\min} f(y) - \langle\mu_{k+1}^\delta, y\rangle + \frac{\rho_2}{2}\|Wx_{k+1}^\delta - y\|^2.
\end{aligned}$$

Let $k = k + 1$ *and return to (c).*

We shall emphasize that (28) is a modified discrepancy principle which allows a stable recovery of the unknown variable x. These type of stopping criterion is necessary in solving ill-posed problems, see for example [6, 7].

Theorem 3 *Let* $(x_k^\delta, y_k^\delta, \lambda_k^\delta, \mu_k^\delta)$ *be the kth iterate of the Relaxing ADMM for noisy data and* $(x_k, y_k, \lambda_k, \mu_k)$ *be the kth iterate of the Relaxing ADMM for exact data. Then for any* $k > 0$ *and* $\delta \to 0$ *there holds*

$$\begin{aligned}
&x_k^\delta \to x_k, \quad y_k^\delta \to y_k, \quad Wx_k^\delta \to Wx_k,\\
&\lambda_k^\delta \to \lambda_k, \quad \mu_k^\delta \to \mu_k, \quad f(y_k^\delta) \to f(y_k).
\end{aligned}$$

Proof Assume that the argument holds true when $k = n$, we consider the case of $k = n + 1$. If $\delta \to 0$, then there holds

$$\lambda_{n+1}^\delta = \lambda_n^\delta + \gamma\rho_1(Ax_n^\delta - b^\delta) \to \lambda_n + \gamma\rho_1(Ax_n - b) = \lambda_{n+1},$$

and similarly $\mu_{n+1}^\delta \to \mu_{n+1}$. By Lemma 1 we further derive

$$x_{n+1}^\delta \to x_{n+1}, \quad Wx_{n+1}^\delta \to Wx_{n+1} \quad \text{if} \quad \delta \to 0.$$

Then by implementing Lemma 2 we obtain

$$y_{n+1}^\delta \to y_{n+1}, \quad f(y_{n+1}^\delta) \to f(y_{n+1}) \quad \text{if} \quad \delta \to 0.$$

□

Recall Algorithm 1, we have the explicit updating form

$$\begin{aligned}
\lambda_{k+1}^{\delta}-\lambda_{k}^{\delta} &= \gamma\rho_1 r_k^{\delta};\\
\mu_{k+1}^{\delta}-\mu_{k}^{\delta} &= \rho_2 s_k^{\delta};\\
A^*\lambda_{k+1}^{\delta}+\rho_1 A^* r_{k+1}^{\delta} &= -W^*[\mu_{k+1}^{\delta}+\rho_2(Wx_{k+1}^{\delta}-y_k^{\delta})];\\
\mu_{k+2}^{\delta} &\in \partial f(y_{k+1}^{\delta}).
\end{aligned}$$

By adjusting the proof of Lemmas 3 and 4, we have the following

Lemma 8 *Let k be the iteration of the Relaxing ADMM for noisy data. If* $k=1$, *there holds*

$$A^*\lambda_2^{\delta}=(\gamma-1)\rho_1 A^* r_1^{\delta}+\rho_2 W^*(y_0^{\delta}-y_1^{\delta})-W^*\mu_2^{\delta}.$$

For any $k\geq 2$, *there holds*

$$\rho_1 A^* r_k^{\delta}+(\gamma-1)\rho_1 A^* r_{k-1}^{\delta}=\rho_2 W^*\left[(y_{k-1}^{\delta}-y_k^{\delta})-(y_{k-2}^{\delta}-y_{k-1}^{\delta})\right]-\rho_2 W^* s_k^{\delta}.$$

Lemma 9 *Denote* $E_k^{\delta}=\rho_1\gamma\|r_k^{\delta}\|^2+\rho_2\|s_k^{\delta}\|^2+\rho_2\|y_k^{\delta}-y_{k-1}^{\delta}\|^2$. *For any* $\gamma\in(0,2)$, $\{E_k^{\delta}\}$ *is monotonically non-increasing such that*

$$E_{k+1}^{\delta}-E_k^{\delta}\leq-\rho_1(2-\gamma)\|r_{k+1}^{\delta}-r_k^{\delta}\|^2-4c_0\|y_{k+1}^{\delta}-y_k^{\delta}\|^2\leq 0$$

and

$$\sum_{k=m}^{n-1}\|y_{k+1}^{\delta}-y_k^{\delta}\|^2\leq\frac{1}{4c_0}E_m^{\delta}, \tag{29}$$

$$(n-m)\rho_2\|y_n^{\delta}-y_{n-1}^{\delta}\|^2\leq\sum_{k=m+1}^{n}E_k^{\delta}.$$

If we have the noisy data, the Bregman distance between the iterates and the exact unknown variable shall be re-estimated more carefully.

Lemma 10 *Denote* k_δ *be the first k which satisfies the stopping criterion* (28). *Then* k_δ *is finite for any* $\delta>0$.

Let $\tilde{\gamma}(\tau)$ *be the solution of* $\gamma+\frac{\gamma^{3/2}}{\tau}=2$. *If* γ *satisfies* $0<\frac{1}{\tau^2}<\gamma<\tilde{\gamma}(\tau)<2$, *for* $1\leq m<n\leq k_\delta-1$, *the following estimate holds*

$$D_{\mu_{n+1}^{\delta}}f(\hat{y},y_n^{\delta})+\frac{c_2}{2}\sum_{k=m+1}^{n}E_k^{\delta}\leq D_{\mu_{m+1}^{\delta}}f(\hat{y},y_m^{\delta})+|\gamma-1|\rho_1\|r_m^{\delta}\|(\|r_m^{\delta}\|/2+\delta)$$

$$+ \rho_2 \langle y_{m-1}^\delta - y_m^\delta, W(\hat{x} - x_{m+1}^\delta)\rangle + \frac{\rho_2}{4}\|W(\hat{x} - x_m^\delta)\|^2$$
$$+ \frac{c_2}{6}\rho_2\|s_m^\delta\|^2 + CE_m^\delta, \tag{30}$$

where constants c_2 and C only depend on c_0, ρ_2, τ and γ.

Proof Similar to (21), we have the estimate

$$\begin{aligned} &D_{\mu_{k+2}^\delta} f(\hat{y}, y_{k+1}^\delta) - D_{\mu_{k+1}^\delta} f(\hat{y}, y_k^\delta) + D_{\mu_{k+1}^\delta} f(y_{k+1}^\delta, y_k^\delta) \\ =& - \langle \mu_{k+2}^\delta - \mu_{k+1}^\delta, \hat{y} - y_{k+1}^\delta\rangle. \end{aligned} \tag{31}$$

By using Lemma 8, we have

$$\begin{aligned} &- \langle \mu_{k+2}^\delta - \mu_{k+1}^\delta, \hat{y} - y_{k+1}^\delta\rangle = -\rho_2\langle s_{k+1}^\delta, \hat{y} - y_{k+1}^\delta\rangle \\ =& - \rho_2\langle s_{k+1}^\delta, W(\hat{x} - x_{k+1}^\delta) + s_{k+1}^\delta\rangle = -\rho_2\|s_{k+1}^\delta\|^2 - \langle \rho_2 W^* s_{k+1}^\delta, \hat{x} - x_{k+1}^\delta\rangle \\ =& - \rho_2\|s_{k+1}^\delta\|^2 - \langle \rho_2 W^*[(y_k^\delta - y_{k+1}^\delta) - (y_{k-1}^\delta - y_k^\delta)], \hat{x} - x_{k+1}^\delta\rangle \\ &+ \langle \rho_1 A^* r_{k+1}^\delta + (\gamma - 1)\rho_1 A^* r_k^\delta, \hat{x} - x_{k+1}^\delta\rangle \\ =& - \rho_2\|s_{k+1}^\delta\|^2 - \rho_1\|r_{k+1}^\delta\|^2 - (\gamma - 1)\rho_1\langle r_{k+1}^\delta, r_k^\delta\rangle \\ &+ (\gamma - 1)\rho_1\langle r_k^\delta, b - b^\delta\rangle + \rho_1\langle r_{k+1}^\delta, b - b^\delta\rangle \\ &+ \rho_2\langle y_{k-1}^\delta - y_k^\delta, W(\hat{x} - x_{k+1}^\delta)\rangle - \rho_2\langle y_k^\delta - y_{k+1}^\delta, W(\hat{x} - x_{k+1}^\delta)\rangle. \end{aligned} \tag{32}$$

Notice that by (31) and (32), the following items are equivalent

$$\begin{aligned} &(I) = \rho_2\langle y_{k-1}^\delta - y_k^\delta, W(\hat{x} - x_{k+1}^\delta)\rangle - \rho_2\langle y_k^\delta - y_{k+1}^\delta, W(\hat{x} - x_{k+1}^\delta)\rangle \\ =& D_{\mu_{k+2}^\delta} f(\hat{y}, y_{k+1}^\delta) - D_{\mu_{k+1}^\delta} f(\hat{y}, y_k^\delta) + D_{\mu_{k+1}^\delta} f(y_{k+1}^\delta, y_k^\delta) + \rho_2\|s_{k+1}^\delta\|^2 + \rho_1\|r_{k+1}^\delta\|^2 \\ &+ (\gamma - 1)\rho_1\langle r_{k+1}^\delta, r_k^\delta\rangle - (\gamma - 1)\rho_1\langle r_k^\delta, b - b^\delta\rangle - \rho_1\langle r_{k+1}^\delta, b - b^\delta\rangle = (II). \end{aligned} \tag{33}$$

Choose m, n such that $1 \le m < n < k_\delta$, we add the left-hand side of (33) from m to $n - 1$ and obtain

$$\begin{aligned} \sum_{k=m}^{n-1}(I) =& \rho_2 \sum_{k=m-1}^{n-2}\langle y_k^\delta - y_{k+1}^\delta, W(\hat{x} - x_{k+2}^\delta)\rangle - \rho_2\sum_{k=m}^{n-1}\langle y_k^\delta - y_{k+1}^\delta, W(\hat{x} - x_{k+1}^\delta)\rangle \\ =& \rho_2\langle y_{m-1}^\delta - y_m^\delta, W(\hat{x} - x_{m+1}^\delta)\rangle - \rho_2\langle y_{n-1}^\delta - y_n^\delta, W(\hat{x} - x_n^\delta)\rangle \\ &+ \rho_2\sum_{k=m}^{n-2}\langle y_k^\delta - y_{k+1}^\delta, W(x_{k+1}^\delta - x_{k+2}^\delta)\rangle. \end{aligned} \tag{34}$$

Two items in the right-hand side of (34) are estimated below.

$$\begin{aligned}
&\sum_{k=m}^{n-2}\langle y_k^\delta - y_{k+1}^\delta, W(x_{k+1}^\delta - x_{k+2}^\delta)\rangle = \sum_{k=m}^{n-2}\langle y_k^\delta - y_{k+1}^\delta, s_{k+1}^\delta + (y_{k+1}^\delta - y_{k+2}^\delta) - s_{k+2}^\delta\rangle \\
&\leq \sum_{k=m}^{n-2}\| y_k^\delta - y_{k+1}^\delta\|(\|s_{k+1}^\delta\| + \| y_{k+1}^\delta - y_{k+2}^\delta\| + \|s_{k+2}^\delta\|) \\
&\leq \sum_{k=m}^{n-2}\left\{\frac{\epsilon}{2}\|s_{k+1}^\delta\|^2 + \frac{\epsilon}{2}\|s_{k+2}^\delta\|^2 + \left(\frac{1}{2}+\frac{1}{\epsilon}\right)\| y_k^\delta - y_{k+1}^\delta\|^2 + \frac{1}{2}\| y_{k+1}^\delta - y_{k+2}^\delta\|^2\right\} \\
&\leq \epsilon\sum_{k=m+1}^{n}\|s_k^\delta\|^2 + \left(1+\frac{1}{\epsilon}\right)\sum_{k=m+1}^{n}\| y_k^\delta - y_{k-1}^\delta\|^2, \qquad (35)
\end{aligned}$$

Let $\xi = \epsilon$ in Lemma 5, we obtain

$$\begin{aligned}
-\langle y_{n-1}^\delta - y_n^\delta, W(\hat{x} - x_n^\delta)\rangle \leq& \frac{1}{4}\|W(\hat{x} - x_m^\delta)\|^2 + \epsilon(\|s_m^\delta\|^2 + \|s_n^\delta\|^2) \\
&+ \frac{1}{2\epsilon}\sum_{k=m+1}^{n}\| y_k^\delta - y_{k-1}^\delta\|^2 + \epsilon(n-m)\| y_{n-1}^\delta - y_n^\delta\|^2.
\end{aligned} \qquad (36)$$

On the other hand, we add the right-hand side of (33) from m to $n-1$. By the strongly convexity of f, we have

$$\begin{aligned}
&\sum_{k=m}^{n-1}(II) = D_{\mu_{n+1}^\delta}f(\hat{y}, y_n^\delta) - D_{\mu_{m+1}^\delta}f(\hat{y}, y_m^\delta) + \sum_{k=m}^{n-1}D_{\mu_{k+1}^\delta}f(y_{k+1}^\delta, y_k^\delta) \\
&\quad + \sum_{k=m+1}^{n}(\rho_2\|s_k^\delta\|^2 + \rho_1\|r_k^\delta\|^2) + (\gamma-1)\rho_1\sum_{k=m}^{n-1}\langle r_{k+1}^\delta, r_k^\delta\rangle \\
&\quad - (\gamma-1)\rho_1\sum_{k=m}^{n-1}\langle r_k^\delta, b - b^\delta\rangle - \rho_1\sum_{k=m+1}^{n}\langle r_k^\delta, b - b^\delta\rangle \\
&\geq D_{\mu_{n+1}^\delta}f(\hat{y}, y_n^\delta) - D_{\mu_{m+1}^\delta}f(\hat{y}, y_m^\delta) + c_0\sum_{k=m}^{n-1}\| y_{k+1} - y_k^\delta\|^2 \\
&\quad + \sum_{k=m+1}^{n}(\rho_2\|s_k^\delta\|^2 + \rho_1\|r_k^\delta\|^2) - |\gamma-1|\rho_1\sum_{k=m}^{n-1}\frac{\|r_{k+1}^\delta\|^2 + \|r_k^\delta\|^2}{2}
\end{aligned}$$

$$
\begin{aligned}
&-\gamma\rho_1\sum_{k=m+1}^{n-1}\|r_k^\delta\|\delta-|\gamma-1|\rho_1\|r_m^\delta\|\delta-\rho_1\|r_n^\delta\|\delta\\
\geq &D_{\mu_{n+1}^\delta}f(\hat{y},y_n^\delta)-D_{\mu_{m+1}^\delta}f(\hat{y},y_m^\delta)+\sum_{k=m+1}^{n}(\rho_2\|s_k^\delta\|^2+\rho_1\|r_k^\delta\|^2+c_0\|\,y_k^\delta-y_{k-1}^\delta\|^2)\\
&-|\gamma-1|\rho_1\sum_{k=m+1}^{n}\|r_k^\delta\|^2-|\gamma-1|\rho_1\frac{\|r_m^\delta\|^2}{2}-\gamma\sum_{k=m+1}^{n-1}\|r_k^\delta\|\delta\rho_1\\
&-|\gamma-1|\|r_m^\delta\|\delta\rho_1-\|r_n^\delta\|\delta\rho_1. \qquad (37)
\end{aligned}
$$

By (34)–(37), we derive

$$
\begin{aligned}
&D_{\mu_{n+1}^\delta}f(\hat{y},y_n^\delta)+(E1)\leq D_{\mu_{m+1}^\delta}f(\hat{y},y_m^\delta)+|\gamma-1|\rho_1\|r_m^\delta\|(\frac{\|r_m^\delta\|}{2}+\delta)+\\
&\rho_2\langle y_{m-1}^\delta-y_m^\delta,W(\hat{x}-x_{m+1}^\delta)\rangle+\epsilon\rho_2\|s_m^\delta\|^2+\frac{\rho_2}{4}\|W(\hat{x}-x_m^\delta)\|^2+(E2);
\end{aligned}
$$

with

$$
\begin{aligned}
(E1)=&\sum_{k=m+1}^{n}\left\{\rho_1(1-|\gamma-1|)\|r_k^\delta\|^2+\rho_2\|s_k^\delta\|^2+c_0\|\,y_k^\delta-y_{k-1}^\delta\|^2\right\}\\
&-\rho_1\delta\|r_n^\delta\|-\gamma\rho_1\sum_{k=m+1}^{n-1}\|r_k^\delta\|\delta; \qquad (38)\\
(E2)=&2\epsilon\rho_2\sum_{k=m+1}^{n}\|s_k^\delta\|^2+\rho_2\left(1+\frac{3}{2\epsilon}\right)\sum_{k=m+1}^{n}\|\,y_k^\delta-y_{k-1}^\delta\|^2\\
&+\epsilon(n-m)\rho_2\|\,y_{n-1}^\delta-y_n^\delta\|^2.
\end{aligned}
$$

If $k<k_\delta$, we apply the stopping criterion to obtain

$$\max(\rho_1^2,\rho_2^2)\tau^2\delta^2\leq\gamma\rho_1^2\|r_k^\delta\|^2+\rho_2^2\|s_k^\delta\|^2.$$

Therefore, we have

$$
\begin{aligned}
\gamma^{1/2}\rho_1\delta\|r_k^\delta\|&=\tau^{1/2}\delta\max(\rho_1,\rho_2)^{1/2}\frac{\gamma^{1/2}\rho_1\|r_k^\delta\|}{\tau^{1/2}\max(\rho_1,\rho_2)^{1/2}}\\
&\leq\frac{\tau\max(\rho_1,\rho_2)\delta^2+\frac{\gamma\rho_1^2\|r_k^\delta\|^2}{\tau\max(\rho_1,\rho_2)}}{2}
\end{aligned}
$$

$$\begin{aligned}&\leq \frac{2\gamma\rho_1^2\|r_k^\delta\|^2+\rho_2^2\|s_k^\delta\|^2}{2\tau\max(\rho_1,\rho_2)}\\&\leq \frac{1}{\tau}\left(\gamma\rho_1\|r_k^\delta\|^2+\rho_2\|s_k^\delta\|^2\right),\quad (1\leq k<k_\delta).\end{aligned}\tag{39}$$

By using inequalities (39) and (38), we derive

$$\begin{aligned}(E1)&\geq c_2\sum_{k=m+1}^{n}E_k^\delta,\\ c_2&=\min\left\{\frac{c_0}{\rho_2},\frac{1-|\gamma-1|}{\gamma}-\frac{1}{\tau}\max(\gamma^{-1/2},\gamma^{1/2})\right\}.\end{aligned}$$

To ensure that $c_2>0$, we choose γ satisfying,

$$0<\frac{1}{\tau^2}<\gamma<\tilde{\gamma}(\tau)<2,$$

where $\tilde{\gamma}$ is the solution of $\gamma+\frac{\gamma^{3/2}}{\tau}=2$ and $\tilde{\gamma}\in(1,2)$. By using the inequality (29), we also have

$$(E2)\leq 3\epsilon\sum_{k=m+1}^{n}E_k^\delta+\frac{\rho_2(1+3/(2\epsilon))}{4c_0}E_m^\delta.$$

Choose $\epsilon=c_2/6$, we thus obtain the inequality (30).

Finally, we prove that Algorithm 1 terminates in finite steps for any $\delta>0$. Assume that the algorithm doesn't terminate in finite steps, then for any k there holds

$$\gamma\rho_1^2\|r_k^\delta\|^2+\rho_2^2\|s_k^\delta\|^2>\max(\rho_1^2,\rho_2^2)\tau^2\delta^2.$$

Hence,

$$E_k^\delta\geq\gamma\rho_1\|r_k^\delta\|^2+\rho_2\|s_k^\delta\|^2>\max(\rho_1,\rho_2)\tau^2\delta^2.$$

By (30), we conclude

$$(n-m)\frac{c_2}{2}\max(\rho_1,\rho_2)\tau^2\delta^2\leq C(m).$$

Letting $n\to\infty$ then yields a contradiction. □

Lemma 11 *Let γ satisfy*

$$\frac{1}{\tau^2}<\gamma<\tilde{\gamma}(\tau)\ \textit{and}\ |\gamma-1|\left(\frac{1}{2\gamma}+\frac{1}{\tau\gamma^{1/2}}\right)<\frac{c_2}{2}.$$

If $m < k_\delta - 1$, then the following estimates hold true

$$\begin{aligned} D_{\mu^\delta_{k_\delta+1}} f(\hat{y}, y^\delta_{k_\delta}) + \tilde{c} E^\delta_{k_\delta} \le & D_{\mu^\delta_{m+1}} f(\hat{y}, y^\delta_m) + |\gamma - 1| \rho_1 \| r^\delta_m \| \left(\frac{\| r^\delta_m \|}{2} + \delta \right) \\ & + \gamma^{-1/2} \max(\rho_1, \rho_2) \tau \delta^2 + \rho_2 \langle y^\delta_{m-1} - y^\delta_m, W(\hat{x} - x^\delta_{m+1}) \rangle \\ & + \frac{3\rho_2}{4} \| W(\hat{x} - x^\delta_m) \|^2 + C_1 \| s^\delta_m \|^2 + C_2 E^\delta_m, \end{aligned} \tag{40}$$

and

$$\begin{aligned} |\langle \mu^\delta_{k_\delta+1}, y^\delta_{k_\delta} - \hat{y} \rangle| \le & |\langle \mu^\delta_{m+1}, y^\delta_{k_\delta} - \hat{y} \rangle| + \gamma^{-1/2} \max(\rho_1, \rho_2) \tau \delta^2 \\ & + C_3 \sum_{k=m}^{k_\delta} E^\delta_k + \frac{\rho_2}{2} \| W(x^\delta_{k_\delta} - \hat{x}) \|^2, \end{aligned} \tag{41}$$

where constants $\tilde{c}$, C_1, C_2, C_3 only depend on c_0, ρ_2, τ and γ.

Proof Choose $k = k_\delta - 1$, implement (28), (31), (32) and notice the strongly convexity of f, we can derive

$$\begin{aligned} & D_{\mu^\delta_{k_\delta+1}} f(\hat{y}, y^\delta_{k_\delta}) + c_0 \| y^\delta_{k_\delta} - y^\delta_{k_\delta-1} \|^2 \\ & \quad \le D_{\mu^\delta_{k_\delta}} f(\hat{y}, y^\delta_{k_\delta-1}) - \rho_2 \| s^\delta_{k_\delta} \|^2 - \rho_1 \| r^\delta_{k_\delta} \|^2 \\ & \qquad + \rho_2 \langle y^\delta_{k_\delta-2} - y^\delta_{k_\delta-1}, W(\hat{x} - x^\delta_{k_\delta}) \rangle - \rho_2 \langle y^\delta_{k_\delta-1} - y^\delta_{k_\delta}, W(\hat{x} - x^\delta_{k_\delta}) \rangle \\ & \qquad - (\gamma - 1) \rho_1 \langle r^\delta_{k_\delta}, r^\delta_{k_\delta-1} \rangle + (\gamma - 1) \rho_1 \langle r^\delta_{k_\delta-1}, b - b^\delta \rangle + \gamma^{-1/2} \max(\rho_1, \rho_2) \tau \delta^2. \end{aligned}$$

Three items appearing in the right-hand side of above inequality can be estimated below

$$\begin{aligned} & \langle y^\delta_{k_\delta-2} - y^\delta_{k_\delta-1}, W(\hat{x} - x^\delta_{k_\delta}) \rangle \\ & \quad = \langle y^\delta_{k_\delta-2} - y^\delta_{k_\delta-1}, W(\hat{x} - x^\delta_m) \rangle + \sum_{k=m}^{k_\delta-1} \langle y^\delta_{k_\delta-2} - y^\delta_{k_\delta-1}, W(x^\delta_k - x^\delta_{k+1}) \rangle \\ & \quad \le \frac{1}{4} \| W(\hat{x} - x^\delta_m) \|^2 + \| y^\delta_{k_\delta-2} - y^\delta_{k_\delta-1} \|^2 \\ & \qquad + \sum_{k=m}^{k_\delta-1} \langle y^\delta_{k_\delta-2} - y^\delta_{k_\delta-1}, s^\delta_k + y^\delta_k - y^\delta_{k+1} - s^\delta_{k+1} \rangle \\ & \quad = \frac{1}{4} \| W(\hat{x} - x^\delta_m) \|^2 + \langle y^\delta_{k_\delta-2} - y^\delta_{k_\delta-1}, s^\delta_m - s^\delta_{k_\delta} \rangle \end{aligned}$$

$$
\begin{aligned}
&+\| y^\delta_{k_\delta-2}-y^\delta_{k_\delta-1}\|^2+\sum_{k=m}^{k_\delta-1}\langle y^\delta_{k_\delta-2}-y^\delta_{k_\delta-1}, y^\delta_k-y^\delta_{k+1}\rangle\\
&\le \frac{1}{4}\|W(\hat{x}-x^\delta_m)\|^2+\frac{\epsilon}{2}\|s^\delta_m\|^2+\frac{\epsilon}{2}\|s^\delta_{k_\delta}\|^2+\frac{1+\epsilon}{\epsilon}\| y^\delta_{k_\delta-2}-y^\delta_{k_\delta-1}\|^2\\
&+\frac{1}{4\epsilon}\sum_{k=m}^{k_\delta-1}\| y^\delta_k-y^\delta_{k-1}\|^2+\epsilon(k_\delta-m)\| y^\delta_{k_\delta-2}-y^\delta_{k_\delta-1}\|^2\\
&\le \frac{1}{4}\|W(\hat{x}-x^\delta_m)\|^2+\frac{\epsilon}{2}\|s^\delta_m\|^2+\frac{\epsilon}{2}\|s^\delta_{k_\delta}\|^2\\
&+\left(1+\frac{5}{4\epsilon}\right)\sum_{k=m}^{k_\delta-1}\| y^\delta_k-y^\delta_{k+1}\|^2+\epsilon(k_\delta-m)\| y^\delta_{k_\delta-2}-y^\delta_{k_\delta-1}\|^2; \qquad (42)
\end{aligned}
$$

Let $\xi=\frac{\epsilon}{2}$ in Lemma 5, we obtain

$$
\begin{aligned}
-\langle y^\delta_{k_\delta-1}-y^\delta_{k_\delta}, W(\hat{x}-x^\delta_{k_\delta})\rangle &\le \frac{1}{4}\|W(\hat{x}-x^\delta_m)\|^2+\frac{\epsilon}{2}(\|s^\delta_m\|^2+\|s^\delta_{k_\delta}\|^2)\\
&+\left(1+\frac{5}{4\epsilon}\right)\sum_{k=m}^{k_\delta-1}\| y^\delta_k-y^\delta_{k+1}\|^2+\epsilon(k_\delta-m)\| y^\delta_{k_\delta-1}-y^\delta_{k_\delta}\|^2; \qquad (43)
\end{aligned}
$$

$$
\begin{aligned}
-(\gamma-1)\rho_1\langle r^\delta_{k_\delta}, r^\delta_{k_\delta-1}\rangle &\le |\gamma-1|\rho_1\frac{\|r^\delta_{k_\delta}\|^2+\|r^\delta_{k_\delta-1}\|^2}{2}\\
&\le \frac{|\gamma-1|}{2\gamma}\left(\gamma\rho_1\|r^\delta_{k_\delta}\|^2+E^\delta_{k_\delta-1}\right). \qquad (44)
\end{aligned}
$$

The stopping criterion (28) and (39) yield

$$
(\gamma-1)\rho_1\langle r^\delta_{k_\delta-1}, b-b^\delta\rangle \le |\gamma-1|\rho_1\|r^\delta_{k_\delta-1}\|\delta \le \frac{|\gamma-1|}{\tau\gamma^{1/2}}E^\delta_{k_\delta-1}. \qquad (45)
$$

By further considering (42)–(45) and applying Lemma 9, we obtain

$$
\begin{aligned}
D_{\mu^\delta_{k_\delta+1}}f(\hat{y}, y^\delta_{k_\delta})+\tilde{c}E^\delta_{k_\delta} &\le D_{\mu^\delta_{k_\delta}}f(\hat{y}, y^\delta_{k_\delta-1})+c_3\sum_{k=m+1}^{k_\delta-1}E^\delta_k+\frac{\rho_2}{2}\|W(\hat{x}-x^\delta_m)\|^2\\
&+\tilde{C}_1\|s^\delta_m\|^2+\tilde{C}_2E^\delta_m+\gamma^{-1/2}\max(\rho_1,\rho_2)\tau\delta^2
\end{aligned}
$$

with

$$c_3 = 2\epsilon + \frac{|\gamma - 1|}{2\gamma} + \frac{|\gamma - 1|}{\tau \gamma^{1/2}}$$

where $\tilde{C}_1, \tilde{C}_2$ only depend on c_0, ρ_2, τ and γ. Let ϵ be small enough, and γ in the neighborhood of 1, we allow $c_3 < c_2/2$. By using (30), we thus obtain (40).

We prove the final inequality. By Lemma 8, we have

$$\begin{aligned}
&\langle \mu^\delta_{k_\delta+1} - \mu^\delta_{m+1}, y^\delta_{k_\delta} - \hat{y}\rangle = \rho_2 \sum_{k=m+1}^{k_\delta} \langle s^\delta_k, y^\delta_{k_\delta} - \hat{y}\rangle \\
=& -\rho_2 \sum_{k=m+1}^{k_\delta} \langle s^\delta_k, s^\delta_{k_\delta}\rangle + \rho_2 \sum_{k=m+1}^{k_\delta} \langle s^\delta_k, W(x^\delta_{k_\delta} - \hat{x})\rangle \\
=& -\rho_2 \sum_{k=m+1}^{k_\delta} \langle s^\delta_k, s^\delta_{k_\delta}\rangle + \rho_2 \sum_{k=m+1}^{k_\delta} \langle (y^\delta_{k-1} - y^\delta_k) - (y^\delta_{k-2} - y^\delta_{k-1}), W(x^\delta_{k_\delta} - \hat{x})\rangle \\
&- \rho_1 \sum_{k=m+1}^{k_\delta} \left\{ \langle r^\delta_k, r^\delta_{k_\delta}\rangle + \langle r^\delta_k, b^\delta - b\rangle \right\} - (\gamma - 1)\rho_1 \sum_{k=m+1}^{k_\delta} \left\{ \langle r^\delta_{k-1}, r^\delta_{k_\delta}\rangle + \langle r^\delta_{k-1}, b^\delta - b\rangle \right\} \\
=& -\rho_2 \sum_{k=m+1}^{k_\delta} \langle s^\delta_k, s^\delta_{k_\delta}\rangle - \gamma\rho_1 \sum_{k=m+1}^{k_\delta - 1} \langle r^\delta_k, r^\delta_{k_\delta}\rangle - (\gamma - 1)\rho_1 \langle r^\delta_m, r^\delta_{k_\delta}\rangle \\
&- \rho_1 \|r^\delta_{k_\delta}\|^2 - \gamma\rho_1 \sum_{k=m+1}^{k_\delta - 1} \langle r^\delta_k, b^\delta - b\rangle - (\gamma - 1)\rho_1 \langle r^\delta_m, b^\delta - b\rangle \\
&- \rho_1 \langle r^\delta_{k_\delta}, b^\delta - b\rangle + \rho_2 \langle (y^\delta_m - y^\delta_{k_\delta}) - (y^\delta_{m-1} - y^\delta_{k_\delta - 1}), W(x^\delta_{k_\delta} - \hat{x})\rangle.
\end{aligned}$$

Take the absolute value on both sides and apply the triangle inequality, we derive

$$\begin{aligned}
&|\langle \mu^\delta_{k_\delta+1} - \mu^\delta_{m+1}, y^\delta_{k_\delta} - \hat{y}\rangle| \\
\le& \frac{1}{2} \sum_{k=m+1}^{k_\delta - 1} \left\{ \gamma\rho_1 \|r^\delta_k\|^2 + \rho_2 \|s^\delta_k\|^2 \right\} + \rho_1 \|r^\delta_{k_\delta}\|^2 + \rho_2 \|s^\delta_{k_\delta}\|^2 + (k_\delta - m - 1)\frac{\rho_2}{2}\|s^\delta_{k_\delta}\|^2 \\
&+ \frac{\gamma\rho_1}{2}(k_\delta - m - 1)\|r^\delta_{k_\delta}\|^2 + |\gamma - 1|\rho_1 |\langle r^\delta_m, r^\delta_{k_\delta}\rangle| + \gamma\rho_1\delta \sum_{k=m+1}^{k_\delta - 1} \|r^\delta_k\| \\
&+ |\gamma - 1|\rho_1\delta\|r^\delta_m\| + \rho_1 \|r^\delta_{k_\delta}\|\delta + \rho_2 |\langle y^\delta_{k_\delta - 1} - y^\delta_{k_\delta}, W(x^\delta_{k_\delta} - \hat{x})\rangle| \\
&+ \rho_2 |\langle y^\delta_{m-1} - y^\delta_m, W(x^\delta_{k_\delta} - \hat{x})\rangle|
\end{aligned}$$

$$\leq \left(1+\max\left(\frac{1}{\gamma},1\right)\right)\sum_{k=m+1}^{k_\delta} E_k^\delta + \max(\gamma,1)\rho_1\delta\sum_{k=m}^{k_\delta-1}\|r_k^\delta\|$$
$$+\rho_1\delta\|r_{k_\delta}^\delta\| + |\gamma-1|\rho_1|\langle r_m^\delta, r_{k_\delta}^\delta\rangle|$$
$$+\rho_2|\langle y_{k_\delta-1}^\delta - y_{k_\delta}^\delta, W(x_{k_\delta}^\delta-\hat{x})\rangle| + \rho_2|\langle y_{m-1}^\delta - y_m^\delta, W(x_{k_\delta}^\delta-\hat{x})\rangle|. \tag{46}$$

Implement the stopping criterion (28) and (39), we can derive

$$\max(\gamma,1)\rho_1\delta\sum_{k=m}^{k_\delta-1}\|r_k^\delta\| \leq \max(\gamma,1)\frac{1}{\tau\gamma^{1/2}}\sum_{k=m}^{k_\delta-1}E_k^\delta, \tag{47}$$

$$\rho_1\delta\|r_{k_\delta}^\delta\| \leq \gamma^{-1/2}\max(\rho_1,\rho_2)\tau\delta^2. \tag{48}$$

Recall the definition of E_k^δ, we have

$$\rho_1|\gamma-1|\left|\langle r_m^\delta, r_{k_\delta}^\delta\rangle\right| \leq \frac{|\gamma-1|}{2\gamma}(E_m^\delta + E_{k_\delta}^\delta), \tag{49}$$

and

$$\rho_2|\langle y_{k_\delta-1}^\delta - y_{k_\delta}^\delta, W(x_{k_\delta}^\delta-\hat{x})\rangle| + \rho_2|\langle y_{m-1}^\delta - y_m^\delta, W(x_{k_\delta}^\delta-\hat{x})\rangle|$$
$$\leq \frac{\rho_2}{2}\|W(x_{k_\delta}^\delta-\hat{x})\|^2 + E_{k_\delta}^\delta + E_m^\delta. \tag{50}$$

Combine (46)–(50), we prove (41). □

Finally we present the main convergence theorem of the Relaxing ADMM for noisy data.

Theorem 4 *Let Assumption 1 (I)–(IV) hold true and the observation data b be consistent. Denote x^* be the unique solution of (2) and $y^* = Wx^*$. If k_δ is the first k satisfying the stopping criterion (28) and γ belongs to the set Φ such that*

$$\Phi = \left\{\gamma \,|\, \max\left\{\frac{2}{3},\frac{1}{\tau^2}\right\} < \gamma < \tilde{\gamma}(\tau), |\gamma-1|\left(\frac{1}{2\gamma}+\frac{1}{\tau\gamma^{1/2}}\right) < \frac{c_2}{2}\right\}$$

then the Relaxing ADMM with noisy data yields

$$x_{k_\delta}^\delta \to x^*, \quad y_{k_\delta}^\delta \to y^*, \quad Wx_{k_\delta}^\delta \to y^*,$$
$$f(y_{k_\delta}^\delta) \to f(y^*), \quad D_{\mu_{k_\delta+1}^\delta} f(y^*, y_{k_\delta}^\delta) \to 0$$

when δ tends to 0.

Proof We prove the theorem in two cases.

Firstly we assume that b^{δ_i} satisfies $\|b^{\delta_i} - b\| \leq \delta_i \to 0$ such that $k_{\delta_i} = k_0 < \infty$ holds for all i. By the stopping criterion, we have

$$\gamma\rho_1^2\|r_{k_0}^{\delta_i}\|^2 + \rho_2^2\|s_{k_0}^{\delta_i}\|^2 \leq \max(\rho_1^2, \rho_2^2)\tau^2\delta_i^2.$$

Let $\delta_i \to 0$ and apply Theorem 3, we have

$$Ax_{k_0} = b, \quad Wx_{k_0} = y_{k_0}.$$

By Algorithm 1, we have $\lambda_{k_0+1} = \lambda_{k_0}$, $\mu_{k_0+1} = \mu_{k_0}$ and $\mu_{k+1} \in \partial f(y_k)$, such that

$$0 = \langle \mu_{k_0+1} - \mu_{k_0}, y_{k_0} - y_{k_0-1}\rangle \geq 2c_0\| y_{k_0} - y_{k_0-1}\|^2.$$

Hence $y_{k_0} = y_{k_0-1}$. We can easily verify that $x_{k_0+1} = x_{k_0}$ and $y_{k_0+1} = y_{k_0}$. Notice that the Relaxing ADMM with exact observation data stops at k_0. From the convergence analysis for exact data, we have

$$x_k = x^*, \quad y_k = y^*,$$

for all $k \geq k_0$. By applying Theorem 3, we have the convergence results.

In the second case, we assume that b^{δ_i} satisfies $\|b^{\delta_i} - b\| \leq \delta_i \to 0$ such that $k_i = k_{\delta_i} \to \infty$ $(i \to \infty)$. We first prove that $D_{\mu_{k_i+1}^{\delta_i}} f(y^*, y_{k_{\delta_i}}^{\delta_i}) \to 0$. By applying upper limit to (40), we have

$$\limsup_{i\to\infty} D_{\mu_{k_i+1}^{\delta_i}} f(y^*, y_{k_i}^{\delta_i}) \leq D_{\mu_{m+1}} f(y^*, y_m) + C\|W(x^* - x_m)\|^2 + C_1\|s_m\|^2$$
$$+ C_2E_m + \rho_2|\langle y_{m-1} - y_m, W(x^* - x_{m+1})\rangle| + |\gamma - 1|\frac{\rho_1}{2}\|r_m\|^2.$$

From the convergence result for the exact data, let $m \to \infty$, we obtain

$$D_{\mu_{k_i+1}^{\delta_i}} f(y^*, y_{k_{\delta_i}}^{\delta_i}) \to 0. \tag{51}$$

Because f is strongly convex, then there holds $y_{k_i}^{\delta_i} \to y^*$. Implement the stopping criterion, we derive

$$\gamma\rho_1^2\|Ax_{k_i}^{\delta_i} - b^{\delta_i}\|^2 + \rho_2^2\|Wx_{k_i}^{\delta_i} - y_{k_i}^{\delta_i}\|^2 \leq \max(\rho_1^2, \rho_2^2)\tau^2\delta_i^2.$$

Let $i \to \infty$, we obtain

$$Ax_{k_i}^{\delta_i} \to b, \quad Wx_{k_i}^{\delta_i} \to y^*.$$

From Assumption 1 (IV), we obtain $x_{k_i}^{\delta_i} \to x^*(i \to \infty)$. Hence,

$$E_{k_i}^{\delta_i} \to 0 \quad (i \to \infty).$$

Finally we prove that $f(y_{k_i}^{\delta_i}) \to f(y^*)$. From (51), we only need to show

$$\langle \mu_{k_i+1}^{\delta_i}, y^* - y_{k_i}^{\delta_i} \rangle \to 0. \tag{52}$$

By applying upper limit to (41), we derive

$$\limsup_{i\to\infty} |\langle \mu_{k_i+1}^{\delta_i}, y^* - y_{k_i}^{\delta_i} \rangle| \leq C_3 \limsup_{i\to\infty} \left(\sum_{k=m}^{k_i-1} E_k^{\delta_i} + E_{k_i}^{\delta_i} \right).$$

From (30),

$$\limsup_{i\to\infty} \sum_{k=m}^{k_i-1} E_k^{\delta_i} \leq C\Big(D_{\mu_m} f(y^*, y_{m-1}) + \tilde{C}\|r_{m-1}\|^2 + \langle y_{m-2} - y_{m-1}, W(x^* - x_m)\rangle + \|W(x^* - x_{m-1})\|^2 + \|s_{m-1}\|^2 + E_{m-1}\Big).$$

Let $m \to \infty$, we thus prove (52). □

5 Conclusion

We investigate the ADMM for solving linear inverse problems. In particular, a relaxing factor is introduced to the standard algorithm allowing more flexible updating of the Lagrange multiplier. We shall emphasize that in principle another relaxing factor can be introduced to the Relaxing ADMM (3)–(6) updating both Lagrange multipliers. But the convergence analysis is much more difficult than what we have proposed in current work.

We skip the numerical simulation since most of the examples are quite robust. By choosing the relaxing factor appropriately, one can obtain early convergence of the Relaxing ADMM with respect to the standard one. The resolution of both algorithms are comparably the same.

Acknowledgements This work is supported by NSFC No.11522108, Shanghai Municipal Education Commission No.16SG01, Special Funds for Major State Basic Research Projects of China (2015CB856003) and the Programme of Introducing Talents of Discipline to Universities (number B08018). Both authors thank Xiliang Lu (Wuhan University) for helpful discussions.

References

1. S. Boyd, N. Parikh, E. Chu, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
2. L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
3. M. Burger, S. Osher, Convergence rates of convex variational regularization. Inverse Prob. **20**, 1411–1421 (2004)
4. W. Deng, W. Yin, On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**, 889–916 (2012)
5. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*. Mathematics and Its Applications, vol. 375 (Kluwer Academic, Dordrecht, 1996)
6. K. Frick, M. Grasmair, Regularization of linear ill-posed problems by the augmented Lagrangian method and variational inequalities. Inverse Prob. **28**(10), 2027–2036 (2012)
7. K. Frick, D.A. Lorenz, E. Resmerita, Morozov's principle for the augmented Lagrangian method applied to linear inverse problems. Multiscale Model. Simul. **4**, 1528–1548 (2010)
8. B. Hofmann, *Regularization for Applied Inverse and Ill-Posed Problems. A Numerical Approach*. Teubner-Texte zur Mathematik, vol. 85 (BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1986)
9. B. Hofmann, B. Kaltenbacher, C. Pöschl, O. Scherzer, A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Prob. **23**(3), 987–1010 (2007)
10. Y. Jiao, Q. Jin, X. Lu, W. Wang, Alternating direction method of multipliers for linear inverse problems. SIAM J. Numer. Anal. **54**, 2114–2137 (2016)
11. Q. Jin, W. Wang, Landweber iteration of Kaczmarz type with general non-smooth convex penalty functionals. Inverse Prob. **29**(8), 085011 (2013)
12. S. Lu, S.V. Pereverzev, *Regularization Theory for Ill-Posed Problems: Selected Topics*, vol. 58 (Walter de Gruyter, Berlin, 2013)
13. S. Osher, M. Burger, D. Goldfarb, et al., An iterative regularization method for total variation-based image restoration. Multiscale Model. Simul. **4**(2), 460–489 (2005)
14. E. Resmerita, O. Scherzer, Error estimates for non-quadratic regularization and the relation to enhancement. Inverse Prob. **22**, 801–814 (2006)
15. T. Schuster, B. Kaltenbacher, B. Hofmann, K.S. Kazimierski, *Regularization Methods in Banach Spaces*. Radon Series on Computational and Applied Mathematics, vol. 10 (Walter de Gruyter, Berlin, 2012), xii+283 pp.
16. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. **67**(2), 301–320 (2005)