

# Iteratively regularized Newton-type methods for general data misfit functionals and applications to Poisson data

Thorsten Hohage · Frank Werner

Received: 5 May 2011 / Revised: 25 July 2012 / Published online: 12 October 2012  
© The Author(s) 2012. This article is published with open access at Springerlink.com

**Abstract** We study Newton type methods for inverse problems described by non-linear operator equations  $F(u) = g$  in Banach spaces where the Newton equations  $F'(u_n; u_{n+1} - u_n) = g - F(u_n)$  are regularized variationally using a general data misfit functional and a convex regularization term. This generalizes the well-known iteratively regularized Gauss–Newton method (IRGNM). We prove convergence and convergence rates as the noise level tends to 0 both for an a priori stopping rule and for a Lepskiĭ-type a posteriori stopping rule. Our analysis includes previous order optimal convergence rate results for the IRGNM as special cases. The main focus of this paper is on inverse problems with Poisson data where the natural data misfit functional is given by the Kullback–Leibler divergence. Two examples of such problems are discussed in detail: an inverse obstacle scattering problem with amplitude data of the far-field pattern and a phase retrieval problem. The performance of the proposed method for these problems is illustrated in numerical examples.

**Mathematics Subject Classification (2000)** 65J15 · 65J20 · 78A46 · 65K10

## 1 Introduction

This study has been motivated by applications in photonic imaging, e.g. positron emission tomography [45], deconvolution problems in astronomy and microscopy [8],

---

T. Hohage · F. Werner (✉)  
Institute for Numerical and Applied Mathematics, University of Göttingen,  
Lotzestraße 16-18, 37083 Göttingen, Germany  
e-mail: f.werner@math.uni-goettingen.de

T. Hohage  
e-mail: hohage@math.uni-goettingen.de

phase retrieval problems [28] or semi-blind deconvolution problems, i.e. deconvolution with partially unknown convolution kernel [43]. In these problems, data consist of counts of photons which have interacted with the object of interest. The inverse problem of recovering the information on the object of interest from such photon counts can be formulated as an operator equation

$$F(u) = g \tag{1}$$

if one introduces an operator  $F : \mathfrak{B} \subset \mathcal{X} \rightarrow \mathcal{Y}$  mapping a mathematical description  $u \in \mathfrak{B}$  of the object of interest to the photon density  $g \in \mathcal{Y} \subset L^1(\mathbb{M})$  on the manifold  $\mathbb{M}$  at which measurements are taken. In this paper we focus on problems where the operator  $F$  is nonlinear.

For fundamental physical reasons, photon count data are described by a Poisson process with the exact data  $g^\dagger$  as mean if read-out noise and finite averaging volume of detectors is neglected. Ignoring this a priori information often leads to non-competitive reconstruction methods.

To avoid technicalities in this introduction, let us consider a discrete version where the exact data vector  $g^\dagger$  belongs to  $[0, \infty)^J$ , and  $g_j^\dagger$  is the expected number of counts of the  $j$ th detector. Then the observed count data are described by a vector  $g^{\text{obs}} \in \mathbb{N}_0^J$  of  $J$  independent Poisson distributed random variables with mean  $g^\dagger$ . A continuous version will be discussed in Sect. 6. Since  $\mathbf{P}[g^{\text{obs}} | g] = \prod_{j=1}^J e^{-g_j} g_j^{g_j^{\text{obs}}} / g_j^{\text{obs}}!$  the negative log-likelihood data misfit functional  $\mathcal{S}(g^{\text{obs}}; g) = -\ln \mathbf{P}[g^{\text{obs}} | g] + c$  (with a constant  $c$  independent of  $g$ ) is given by

$$\mathcal{S}(g^{\text{obs}}; g) := \begin{cases} \sum_{j=1}^J [g_j - g_j^{\text{obs}} \ln g_j], & g \geq 0 \text{ and } \{j : g_j^{\text{obs}} > 0, g_j = 0\} = \emptyset, \\ \infty, & \text{else.} \end{cases} \tag{2}$$

Here and in the following we use the conventions  $0 \ln 0 := 0$  and  $\ln 0 = -\infty$ . Setting  $g^{\text{obs}} = g^\dagger$  and subtracting the minimal value  $\sum_{j=1}^J [g_j^\dagger - g_j^\dagger \ln g_j^\dagger]$  attained at  $g = g^\dagger$ , we obtain a discrete version of the *Kullback–Leibler divergence*

$$\mathbb{KL}(g^\dagger; g) := \begin{cases} \sum_{j=1}^J \left[ g_j - g_j^\dagger - g_j^\dagger \ln \left( \frac{g_j}{g_j^\dagger} \right) \right] & g \geq 0, \{j : g_j^\dagger > 0, g_j = 0\} = \emptyset, \\ \infty, & \text{else.} \end{cases} \tag{3}$$

Note that both  $\mathcal{S}$  and  $\mathbb{KL}$  are convex in their second arguments.

A standard way to solve perturbed nonlinear operator equations (1) is the Gauß-Newton method. If  $F'$  denotes the Gateaux derivative of  $F$ , it is given by given by  $u_{n+1} \in \operatorname{argmin}_{u \in \mathfrak{B}} \|F(u_n) + F'(u_n; u - u_n) - g^{\text{obs}}\|^2$ . As explained above, for data errors with a non-Gaussian distribution it is in general not appropriate to use a squared norm as data misfit functional. Therefore, we will consider general data misfit functionals  $\mathcal{S} : \mathcal{Y}^{\text{obs}} \times \mathcal{Y} \rightarrow (-\infty, \infty]$  where  $\mathcal{Y}^{\text{obs}}$  is a space of (possibly discrete) observations  $g^{\text{obs}}$ . Since inverse problems are typically ill-posed in the sense that  $F$  and its derivatives  $F'(u_n; \cdot)$  do not have continuous inverses, regularization has to

be used. Therefore, we add a proper convex penalty functional  $\mathcal{R} : \mathcal{X} \rightarrow (-\infty, \infty]$ , which should be chosen to incorporate *a priori* knowledge about the unknown solution  $u^\dagger$ . This leads to the iteratively regularized Newton-type method

$$u_{n+1} \in \operatorname{argmin}_{u \in \mathfrak{B}} \left[ \mathcal{S} \left( g^{\text{obs}}; F(u_n) + F'(u_n; u - u_n) \right) + \alpha_n \mathcal{R}(u) \right] \tag{4a}$$

which will be analyzed in this paper. The regularization parameters  $\alpha_n$  are chosen such that

$$\alpha_0 \leq 1, \quad \alpha_n \searrow 0, \quad 1 \leq \frac{\alpha_n}{\alpha_{n+1}} \leq C_{\text{dec}} \quad \text{for all } n \in \mathbb{N} \tag{4b}$$

for some constant  $C_{\text{dec}}$ , typically  $\alpha_n = \alpha_0 C_{\text{dec}}^{-n}$  with  $C_{\text{dec}} = 3/2$ .

If  $\mathcal{Y} = \mathbb{R}^J$ ,  $F(u) = (F_j(u))_{j=1, \dots, d}$ , and  $\mathcal{S}$  is given by (2), we obtain the convex minimization problems

$$u_{n+1} \in \operatorname{argmin}_{u \in \mathfrak{B}_n} \left[ \sum_{j=1}^J [F_j(u_n) + F'_j(u_n; u - u_n) - g_j^{\text{obs}} \ln(F_j(u_n) + F'_j(u_n; u - u_n))] + \alpha_n \mathcal{R}(u) \right] \tag{5}$$

in each Newton step where

$$\mathfrak{B}_n := \left\{ u \in \mathfrak{B} \mid \mathcal{S} \left( g^{\text{obs}}; F(u) + F'(u_n; u - u_n) \right) < \infty \right\}.$$

In principle, several methods for the solution of (5) are available. In particular we mention inverse scale space methods [12,37] for linear operator equations and total variation penalties  $\mathcal{R}$ . EM-type methods cannot readily be used for the solution of the convex minimization problems (5) (or subproblems of the inverse scale space method as in [12]) if  $F'(u_n; \cdot)$  is not positivity preserving as in our examples. A simple algorithm for the solution of subproblems of the type (5) is discussed in Sect. 7. We consider the design of more efficient algorithms for minimizing the functionals (5) for large scale problems as an important problem for future research.

The most common choice of the data misfit functional is  $\mathcal{S}(\hat{g}; g) = \|g - \hat{g}\|_{\mathcal{Y}}^2$ , with a Hilbert space norm  $\|\cdot\|_{\mathcal{Y}}$ . This can be motivated by the case of (multi-variate) Gaussian errors. If the penalty term is also given by a Hilbert space norm  $\mathcal{R}(u) = \|u - u_0\|_{\mathcal{X}}^2$ , (4) becomes the iteratively regularized Gauss–Newton method (IRGNM) which is one of the most popular methods for solving nonlinear ill-posed operator equations [2,3,9,31]. If the penalty term  $\|u - u_0\|_{\mathcal{X}}^2$  is replaced by  $\|u - u_n\|_{\mathcal{X}}^2$  one obtains the Levenberg–Marquardt method, which is well-known in optimization and has first been analyzed as regularization method in [20]. Recently, a generalization of the IRGNM to Banach spaces has been proposed and analyzed by Kaltenbacher and Hofmann [30].

As an alternative to (4) we mention Tikhonov-type or variational regularization methods of the form

$$\hat{u}_\alpha \in \operatorname{argmin}_{u \in \mathfrak{B}} \left[ \mathcal{S} \left( g^{\text{obs}}; F(u) \right) + \alpha \mathcal{R}(u) \right]. \quad (6)$$

Here  $\alpha > 0$  is a regularization parameter. For nonlinear operators this is in general a non-convex optimization problem even if  $\mathcal{S}(g^{\text{obs}}; \cdot)$  and  $\mathcal{R}$  are convex. Hence, (6) may have many local minima and it cannot be guaranteed that the global minimum can be found numerically. Let us summarize some recent convergence results on this method: Bardsley [4] shows stability and convergence for linear operators and  $\mathcal{S} = \mathbb{K}\mathbb{L}$ . Benning and Burger [7] prove rates of convergence for linear operators under the special source condition  $F^* \omega \in \partial \mathcal{R}(u^\dagger)$ . Generalizations to nonlinear operators and general variational source conditions were published simultaneously by Bot and Hofmann [11], Flemming [16], and Grasmair [19].

Given some rule to choose the stopping index  $n_*$  our main results (Theorems 2.3 and 4.2) establish rates of convergence of the method (4), i.e. uniform estimates of the error of the final iterate in terms of some data noise level **err**

$$\left\| u_{n_*} - u^\dagger \right\| \leq C \varphi(\mathbf{err}) \quad (7)$$

for some increasing, continuous function  $\varphi : [0, \infty) \rightarrow [0, \infty)$  satisfying  $\varphi(0) = 0$ . For the classical deterministic error model  $\|g^{\text{obs}} - g\| \leq \delta$  and  $\mathcal{S}(g^{\text{obs}}; g) = \|g - g^{\text{obs}}\|^r$  with some  $r \geq 1$  we have  $\mathbf{err} = \delta^r$ . In this case we recover most of the known convergence results on the IRGNM for weak source conditions. Our main results imply error estimates for Poisson data provided a concentration inequality holds true. In this case  $\mathbf{err} = \frac{1}{\sqrt{t}}$  where  $t$  can be interpreted as an exposure time proportional to the expected total number of photons, and an estimate of the form (7) holds true with the left hand side replaced by an expected error.

As opposed to a Hilbert or Banach space setting our data misfit functional  $\mathcal{S}$  does not necessarily fulfill a triangle inequality. Therefore, it is necessary to use more general formulations of the noise level and the tangential cone condition, which controls the degree of nonlinearity of the operator  $F$ . Both coincide with the usual assumptions if  $\mathcal{S}$  is given by a norm. Our analysis uses variational methods rather than methods based on spectral theory, which have recently been studied in the context of inverse problems by a number of authors (see, e.g., [13, 24, 30, 40, 42]).

The plan of this paper is as follows: In the following section we formulate our first main convergence theorem (Theorem 2.3) and discuss its assumptions. The proof will be given in Sect. 3. In the following Sect. 4 we discuss the case of additive variational inequalities and state a convergence rates result for a Lepskiĭ-type stopping rule (Theorem 4.2). In Sect. 5 we compare our result to previous results on the iteratively regularized Gauss–Newton method. Section 6 is devoted to the special case of Poisson data, which has been our main motivation. We conclude our paper with numerical results for an inverse obstacle scattering problem and a phase retrieval problem in optics in Sect. 7.

## 2 Assumptions and convergence theorem with a priori stopping rule

Throughout the paper we assume the following mapping and differentiability properties of the forward operator  $F$ :

**Assumption 1** (*Assumptions on  $F$  and  $\mathcal{R}$* ) Let  $\mathcal{X}$  and  $\mathcal{Y}$  be Banach spaces and let  $\mathfrak{B} \subset \mathcal{X}$  a convex subset. Assume that the forward operator  $F : \mathfrak{B} \rightarrow \mathcal{Y}$  and the penalty functional  $\mathcal{R} : \mathcal{X} \rightarrow (-\infty, \infty]$  have the following properties:

1.  $F$  is injective.
2.  $F : \mathfrak{B} \rightarrow \mathcal{Y}$  is continuous, the first variations

$$F'(u; v - u) := \lim_{t \searrow 0} \frac{1}{t} (F(u + t(v - u)) - F(u))$$

exist for all  $u, v \in \mathfrak{B}$ , and  $h \mapsto F'(u; h)$  can be extended to a bounded linear operator  $F'[u] \in L(\mathcal{X}, \mathcal{Y})$  for all  $u \in \mathfrak{B}$ .

3.  $\mathcal{R}$  is proper and convex, and  $\mathfrak{B} \cap \text{dom}(\mathcal{R}) \neq \emptyset$ .

At interior points  $u \in \mathfrak{B}$  the second assumption amounts to Gateaux differentiability of  $F$ .

To motivate our assumptions on the data misfit functional, let us consider the case that  $g^{\text{obs}} = F(u^\dagger) + \xi$ , and  $\xi$  is Gaussian white noise on the Hilbert space  $\mathcal{Y}$ , i.e.  $\langle \xi, g \rangle \sim N(0, \|g\|^2)$  and  $\mathbf{E}\langle \xi, g \rangle \langle \xi, \tilde{g} \rangle = \langle g, \tilde{g} \rangle$  for all  $g, \tilde{g} \in \mathcal{Y}$ . If  $\mathcal{Y} = \mathbb{R}^J$ , then the negative log-likelihood functional is given by  $\mathcal{S}(g^{\text{obs}}; g) = \|g - g^{\text{obs}}\|_2^2$ . However, in an infinite dimensional Hilbert space  $\mathcal{Y}$  we have  $\|g^{\text{obs}}\|_{\mathcal{Y}} = \infty$  almost surely, and  $\mathcal{S}(g^{\text{obs}}; \cdot) \equiv \infty$  is obviously not a useful data misfit term. Therefore, one formally subtracts  $\|g^{\text{obs}}\|_{\mathcal{Y}}^2$  (which is independent of  $g$ ) to obtain  $\mathcal{S}(g^{\text{obs}}; g) := \|g\|_{\mathcal{Y}}^2 - 2\langle g^{\text{obs}}, g \rangle_{\mathcal{Y}}$ . For exact data  $g^\dagger$  we can of course use the data misfit functional  $\mathcal{T}(g^\dagger; g) = \|g - g^\dagger\|_{\mathcal{Y}}^2$ . As opposed to  $\mathcal{S}$ , the functional  $\mathcal{T}$  is nonnegative and does indeed describe the size of the error in the data space  $\mathcal{Y}$ . It will play an important role in our analysis.

It may seem cumbersome to work with two different types data misfit functionals  $\mathcal{S}$  and  $\mathcal{T}$ . A straightforward idea to fix the free additive constant in  $\mathcal{S}$  is to introduce  $\tilde{\mathcal{S}}(g^{\text{obs}}; g) := \mathcal{S}(g^{\text{obs}}; g) - \tilde{s}$  with  $\tilde{s} := \inf_{g \in \mathcal{Y}} \mathcal{S}(g^{\text{obs}}; g)$  such that  $\tilde{\mathcal{S}}(g^{\text{obs}}; \cdot)$  is nonnegative and  $\tilde{\mathcal{S}}(g^\dagger; g) = \mathcal{T}(g^\dagger; g)$ . However,  $\tilde{s} = -\infty$  a.s. A better choice of the additive constant is  $s = \mathbf{E}\mathcal{S}(g^{\text{obs}}; g) - \mathcal{T}(g^\dagger; g) = -\|g^\dagger\|^2$  since for this choice the error has the convenient representation  $\mathcal{S}(g^{\text{obs}}; g) + \|g^\dagger\|^2 - \mathcal{T}(g^\dagger; g) = -2\langle \xi, g \rangle_{\mathcal{Y}}$ , and the expected error  $\mathbf{E}|\mathcal{S}(g^{\text{obs}}; g) - s - \mathcal{T}(g^\dagger; g)|^2$  is minimized. Note that  $s$  depends on the unknown  $g^\dagger$ , but this does not matter since the value of  $s$  does not affect the numerical algorithms. Bounds on  $\sup_{g \in \tilde{\mathcal{Y}}} |\langle \xi, g \rangle_{\mathcal{Y}}|$  with high probabilities for certain subsets  $\tilde{\mathcal{Y}} \subset \mathcal{Y}$  (concentration inequalities) have been studied intensively in probability theory (see e.g. [34]). Such results can be used in case of Gaussian errors to show that the following deterministic error assumption holds true with high probability and uniform bounds on  $\mathbf{err}(g)$  for  $g \in \tilde{\mathcal{Y}}$ .

**Assumption 2** (*data errors, properties of  $\mathcal{S}$  and  $\mathcal{T}$* ) Let  $u^\dagger \in \mathfrak{B} \subset \mathcal{X}$  be the exact solution and denote by  $g^\dagger := F(u^\dagger) \in \mathcal{Y}$  the exact data. Let  $\mathcal{Y}^{\text{obs}}$  be a set containing all possible observations and  $g^{\text{obs}} \in \mathcal{Y}^{\text{obs}}$  the observed data. Assume that:

1. The fidelity term  $\mathcal{T} : F(\mathfrak{B}) \times \mathcal{Y} \rightarrow [0, \infty]$  with respect to exact data fulfills  $\mathcal{T}(g^\dagger; g^\dagger) = 0$ .
2.  $\mathcal{T}$  and the fidelity term  $\mathcal{S} : \mathcal{Y}^{\text{obs}} \times \mathcal{Y} \rightarrow (-\infty, \infty]$  with respect to noisy data are connected as follows: There exists a constant  $C_{\text{err}} \geq 1$  and functionals  $\mathbf{err} : \mathcal{Y} \rightarrow [0, \infty]$  and  $\mathfrak{s} : F(\mathfrak{B}) \rightarrow (-\infty, \infty)$  such that

$$\mathcal{S}(g^{\text{obs}}; g) - \mathfrak{s}(g^\dagger) \leq C_{\text{err}} \mathcal{T}(g^\dagger; g) + C_{\text{err}} \mathbf{err}(g) \tag{8a}$$

$$\mathcal{T}(g^\dagger; g) \leq C_{\text{err}} (\mathcal{S}(g^{\text{obs}}; g) - \mathfrak{s}(g^\dagger)) + C_{\text{err}} \mathbf{err}(g) \tag{8b}$$

for all  $g \in \mathcal{Y}$ .

*Example 2.1* 1. Additive deterministic errors in Banach spaces. Assume that  $\mathcal{Y}^{\text{obs}} = \mathcal{Y}$ ,

$$\|g^{\text{obs}} - g^\dagger\| \leq \delta, \quad \text{and} \quad \mathcal{S}(g_2; g_1) = \mathcal{T}(g_2; g_1) = \|g_1 - g_2\|_Y^r$$

with  $r \in [1, \infty)$ . Then it follows from the simple inequalities  $(a + b)^r \leq 2^{r-1}(a^r + b^r)$  and  $|a - b|^r + b^r \geq 2^{1-r}a^r$  that (8) holds true with  $\mathbf{err} \equiv \|\|g^{\text{obs}} - g^\dagger\|_Y\|^r$ ,  $\mathfrak{s} \equiv 0$  and  $C_{\text{err}} = 2^{r-1}$ .

2. For randomly perturbed data a general recipe for the choice of  $\mathcal{S}$ ,  $\mathcal{T}$  and  $\mathfrak{s}$  is to define  $\mathcal{S}$  as the log-likelihood functional,  $\mathfrak{s}(g^\dagger) := \mathbf{E}_{g^\dagger} \mathcal{S}(g^{\text{obs}}; g^\dagger)$  and  $\mathcal{T}(g^\dagger; g) := \mathbf{E}_{g^\dagger} \mathcal{S}(g^{\text{obs}}; g) - \mathfrak{s}(g^\dagger)$ . Then we always have  $\mathcal{T}(g^\dagger; g^\dagger) = 0$ , but part 2. of Assumption 2 has to be verified case by case.
3. *Poisson data.* For discrete Poisson data we have already seen in the introduction that the general recipe of the previous point yields  $\mathcal{S}$  given by (2),  $\mathcal{T} = \mathbb{KL}$  and  $\mathfrak{s}(g^\dagger) = \sum_{j=1}^J [g_j^\dagger - g_j^\dagger \ln(g_j^\dagger)]$ . It is easy to see that  $\mathbb{KL}(g^\dagger; g) \geq 0$  for all  $g^\dagger$  and  $g$ . Then (8) holds true with  $C_{\text{err}} = 1$  and

$$\mathbf{err}(g) = \begin{cases} \left| \sum_{j=1}^J \ln(g_j) (g_j^{\text{obs}} - g_j^\dagger) \right|, & g \geq 0, \{j : g_j = 0, g_j^\dagger + g_j^{\text{obs}} > 0\} = \emptyset \\ \infty, & \text{else.} \end{cases}$$

Obviously, it will be necessary to show that  $\mathbf{err}(g)$  is finite and even small in some sense for all  $g$  for which the inequalities (8) are applied (see Sect. 6).

To simplify our notation we will assume in the following analysis that  $\mathfrak{s} \equiv 0$  or equivalently replace  $\mathcal{S}(g^{\text{obs}}; g)$  by  $\mathcal{S}(g^{\text{obs}}; g) - \mathfrak{s}(g^\dagger)$ . As already mentioned in the motivation of Assumption 2, it is not relevant that  $\mathfrak{s}(g^\dagger)$  is unknown since the value of this additive constant does not influence the iterates  $u_n$  in (4a).

Typically  $\mathcal{S}$  and  $\mathcal{T}$  will be convex in their second arguments, but we do not need this property in our analysis. However, without convexity it is not clear if the numerical solution of (4a) is easier than the numerical solution of (6).

**Assumption 3** (*Existence*) For any  $n \in \mathbb{N}$  the problem (4a) has a solution.

*Remark 2.2* By standard arguments the following properties are sufficient to ensure existence of a solution to (4a) for convex  $\mathcal{S}(g^{\text{obs}}; \cdot)$  (see [16, 24, 39]): There are possibly weaker topologies  $\tau_{\mathcal{X}}, \tau_{\mathcal{Y}}$  on  $\mathcal{X}, \mathcal{Y}$  respectively such that

1.  $\mathfrak{B}$  is sequentially closed w.r.t.  $\tau_{\mathcal{X}}$ ,
2.  $F'(u; \cdot)$  is sequentially continuous w.r.t.  $\tau_{\mathcal{X}}$  and  $\tau_{\mathcal{Y}}$  for all  $u \in \mathfrak{B}$ ,
3. the penalty functional  $\mathcal{R} : \mathcal{X} \rightarrow (-\infty, \infty]$  is sequentially lower semi-continuous with respect to  $\tau_{\mathcal{X}}$ ,
4. the sets  $M_{\mathcal{R}}(c) := \{u \in \mathcal{X} \mid \mathcal{R}(u) \leq c\}$  are sequentially pre-compact with respect to  $\tau_{\mathcal{X}}$  for all  $c \in \mathbb{R}$  and
5. the data misfit term  $\mathcal{S}(g^{\text{obs}}; \cdot) : \mathcal{Y} \rightarrow (-\infty, \infty]$  is sequentially lower semi-continuous w.r.t.  $\tau_{\mathcal{Y}}$ , and  $\inf_{u \in \mathfrak{B}} \mathcal{S}(g^{\text{obs}}; F(u_n) + F'(u_n; u - u_n)) > -\infty$ .

Note that for our analysis we do not require that the solution to (4a) is unique or depends continuously on the data  $g^{\text{obs}}$  even though these properties are desirable for other reasons. Obviously, uniqueness in (4a) is given if  $\mathcal{S}$  is convex and  $\mathcal{R}$  is strictly convex, and there are reasonable assumptions on  $\mathcal{S}$  which guarantee continuous dependence, cf. [39].

All known convergence rate results for nonlinear ill-posed problems under weak source conditions assume some condition restricting the degree of nonlinearity of the operator  $F$ . Here we use a generalization of the tangential cone condition which was introduced in [21] and is frequently used for the analysis of regularization methods for nonlinear inverse problems. It must be said, however, that for many problems it is very difficult to show that this condition is satisfied (or not satisfied). Since  $\mathcal{S}$  does not necessarily fulfill a triangle inequality we have to use a generalized formulation of the tangential cone condition, which follows from the standard formulation if  $\mathcal{S}$  is given by the power of a norm (cf. Lemma 5.2).

**Assumption 4** (*Generalized tangential cone condition*)

- (A) There exist constants  $\eta$  (later assumed to be sufficiently small) and  $C_{\text{tc}} \geq 1$  such that for all  $g^{\text{obs}} \in \mathcal{Y}^{\text{obs}}$

$$\begin{aligned} & \frac{1}{C_{\text{tc}}} \mathcal{S}(g^{\text{obs}}; F(v)) - \eta \mathcal{S}(g^{\text{obs}}; F(u)) \\ & \leq \mathcal{S}(g^{\text{obs}}; F(u) + F'(u; v - u)) \\ & \leq C_{\text{tc}} \mathcal{S}(g^{\text{obs}}; F(v)) + \eta \mathcal{S}(g^{\text{obs}}; F(u)) \quad \text{for all } u, v \in \mathfrak{B}. \end{aligned} \tag{9a}$$

- (B) There exist constants  $\eta$  (later assumed to be sufficiently small) and  $C_{\text{tc}} \geq 1$  such that

$$\begin{aligned} & \frac{1}{C_{\text{tc}}} \mathcal{T}(g^\dagger; F(v)) - \eta \mathcal{T}(g^\dagger; F(u)) \\ & \leq \mathcal{T}(g^\dagger; F(u) + F'(u; v - u)) \\ & \leq C_{\text{tc}} \mathcal{T}(g^\dagger; F(v)) + \eta \mathcal{T}(g^\dagger; F(u)) \quad \text{for all } u, v \in \mathfrak{B}. \end{aligned} \tag{9b}$$

This condition ensures that the nonlinearity of  $F$  fits together with the data misfit functionals  $\mathcal{S}$  or  $\mathcal{T}$ . Obviously, it is fulfilled with  $\eta = 0$  and  $C_{tc} = 1$  if  $F$  is linear.

It is well-known that for ill-posed problems rates of convergence can only be obtained under an additional “smoothness condition” on the solution (see [15, Prop. 3.11]). In a Hilbert space setting such conditions are usually formulated as source conditions in the form

$$u^\dagger - u_0 = \varphi \left( F' \left[ u^\dagger \right]^* F' \left[ u^\dagger \right] \right) \omega \tag{10}$$

for some  $\omega \in \mathcal{X}$  where  $\varphi : [0, \infty) \rightarrow [0, \infty)$  is a so-called *index function*, i.e.  $\varphi$  is continuous and monotonically increasing with  $\varphi(0) = 0$ . Such general source conditions were systematically studied in [23, 36]. The most common choices of  $\varphi$  are discussed in Sect. 5.

To formulate similar source conditions in Banach spaces, we first have to introduce Bregman distances, which will also be used to measure the error of our approximate solutions (see [13]): Let  $u^* \in \partial\mathcal{R}(u^\dagger)$  be a subgradient (e.g.  $u^* = u^\dagger - u_0$  if  $\mathcal{R}(u) = \frac{1}{2}\|u - u_0\|^2$  with a Hilbert norm  $\|\cdot\|$ ). Then the Bregman distance of  $\mathcal{R}$  between  $u$  and  $u^\dagger$  is given by

$$\mathcal{D}_{\mathcal{R}}^{u^*}(u, u^\dagger) := \mathcal{R}(u) - \mathcal{R}(u^\dagger) - \langle u^*, u - u^\dagger \rangle.$$

If  $\mathcal{X}$  is a Hilbert space and  $\mathcal{R}(u) = \frac{1}{2}\|u - u_0\|_{\mathcal{X}}^2$ , we have  $\mathcal{D}_{\mathcal{R}}^{u^*}(u, u^\dagger) = \frac{1}{2}\|u - u^\dagger\|_{\mathcal{X}}^2$ . Moreover, if  $\mathcal{X}$  is a  $q$ -convex Banach space ( $1 < q < \infty$ ) and  $\mathcal{R}(u) = \|u\|_{\mathcal{X}}^q$ , then there exists a constant  $C_{bd} > 0$  such that

$$\|u - u^\dagger\|_{\mathcal{X}}^q \leq C_{bd} \mathcal{D}_{\mathcal{R}}^{u^*}(u, u^\dagger) \tag{11}$$

for all  $u \in \mathcal{X}$  (this follows from [47, Eqs. (2.16)' and (2.17)']). In those cases, convergence rates w.r.t. the Bregman distance also imply rates w.r.t. the Banach space norm.

Now we can formulate the following variational formulation of the source condition (10), which is a slight variation of the one proposed in [30]:

**Assumption 5A** (*Multiplicative variational source condition*) There exists  $u^* \in \partial\mathcal{R}(u^\dagger) \subset \mathcal{X}'$ ,  $\beta \geq 0$  and a concave index function  $\varphi : (0, \infty) \rightarrow (0, \infty)$  such that

$$\langle u^*, u^\dagger - u \rangle \leq \beta \mathcal{D}_{\mathcal{R}}^{u^*}(u, u^\dagger)^{\frac{1}{2}} \varphi \left( \frac{\mathcal{T}(g^\dagger; F(u))}{\mathcal{D}_{\mathcal{R}}^{u^*}(u, u^\dagger)} \right) \quad \text{for all } u \in \mathfrak{B}. \tag{12}$$

Moreover, we assume that

$$t \mapsto \frac{\varphi(t)}{\sqrt{t}} \quad \text{is monotonically decreasing.} \tag{13}$$



As noted in [30] using Jensen’s inequality, a Hilbert space source condition (10) for which  $(\varphi^2)^{-1}$  is convex implies the variational inequality

$$\left| \langle u^*, u - u^\dagger \rangle \right| \leq \omega \| \| u - u^\dagger \| \varphi \left( \frac{\| F' [u^\dagger] (u - u^\dagger) \|^2}{\| u - u^\dagger \|^2} \right). \tag{14}$$

The tangential cone condition now shows that an inequality of type (12) is valid and hence, in a Hilbert space setup Assumption 5 is weaker than (10) at least for linear operators. As opposed to [30] we have omitted absolute values on the left hand side of (12) since they are not needed in the proofs, and this form may allow for better index functions  $\varphi$  if  $u^\dagger$  is on the boundary of  $\mathfrak{B}$ .

In many recent publications [11, 16, 25, 42] variational source conditions in additive rather than multiplicative form have been used. Such conditions will be discussed in Sect. 4.

Since we use a source condition with a general index function  $\varphi$ , we need to restrict the nonlinearity of  $F$  with the help of a tangential cone condition. Nevertheless, we want to mention that for  $\varphi(t) = t^{1/2}$  in (12) our convergence analysis also works under a generalized Lipschitz assumption, but this lies beyond the aims of this paper. The cases  $\varphi(t) = t^\nu$  with  $\nu > \frac{1}{2}$  where similar results are expected are not covered by Assumption 5, since for the motivation in the Hilbert space setup we needed to assume that  $(\varphi^2)^{-1}$  is convex, which is not the case for  $\nu > \frac{1}{2}$ .

In our convergence analysis we will use the following two functions, which are both index functions as well as their inverses:

$$\begin{aligned} \Theta(t) &:= t\varphi^2(t), \\ \vartheta(t) &:= \sqrt{\Theta(t)} = \sqrt{t}\varphi(t) \end{aligned} \tag{15}$$

We are now in a position to formulate our convergence result with a priori stopping rule:

**Theorem 2.3** *Let Assumption 1, 2, 3, 4A or 4B and 5A hold true, and suppose that  $\eta, \mathcal{D}_{\mathcal{R}}^{u^*}(u_0, u^\dagger)$  and  $\mathcal{T}(g^\dagger; F(u_0))$  are sufficiently small. Then the iterates  $u_n$  defined by (4) with exact data  $g^{\text{obs}} = g^\dagger$  fulfill*

$$\mathcal{D}_{\mathcal{R}}^{u^*}(u_n, u^\dagger) = \mathcal{O}(\varphi^2(\alpha_n)), \tag{16a}$$

$$\mathcal{T}(g^\dagger; F(u_n)) = \mathcal{O}(\Theta(\alpha_n)) \tag{16b}$$

as  $n \rightarrow \infty$ . For noisy data define

$$\mathbf{err}_n := \frac{1}{C_{\text{err}}} \mathbf{err}(F(u_{n+1})) + 2\eta C_{\text{tc}} \mathbf{err}(F(u_n)) + C_{\text{tc}} C_{\text{err}} \mathbf{err}(g^\dagger) \tag{17a}$$

in case of Assumption 4A or

$$\mathbf{err}_n := \mathbf{err} \left( F(u_n) + F'(u_n; u_{n+1} - u_n) \right) + C_{\text{err}} \mathbf{err} \left( F(u_n) + F'(u_n; u^\dagger - u_n) \right) \tag{17b}$$

under Assumption 4B, and choose the stopping index  $n_*$  by

$$n_* := \min \{ n \in \mathbb{N} \mid \Theta(\alpha_n) \leq \tau \mathbf{err}_n \} \tag{18}$$

with a sufficiently large parameter  $\tau \geq 1$ . Then (16) holds for  $n \leq n_*$  and the following convergence rates are valid:

$$\mathcal{D}_{\mathcal{R}}^{u^*} \left( u_{n_*}, u^\dagger \right) = \mathcal{O} \left( \varphi^2 \left( \Theta^{-1} \left( \mathbf{err}_{n_*} \right) \right) \right), \tag{19a}$$

$$\mathcal{T} \left( g^\dagger; F(u_{n_*}) \right) = \mathcal{O} \left( \mathbf{err}_{n_*} \right). \tag{19b}$$

### 3 Proof of Theorem 2.3

We will split the proof into two main parts. For brevity we will denote

$$d_n := \mathcal{D}_{\mathcal{R}}^{u^*} \left( u_n, u^\dagger \right)^{\frac{1}{2}}, \tag{20}$$

$$s_n := \mathcal{T} \left( g^\dagger; F(u_n) \right). \tag{21}$$

Let us now start with the following

**Lemma 3.1** *Let the assumptions of Theorem 2.3 hold true. Then we have a recursive error estimate of the form*

$$\alpha_n d_{n+1}^2 + \frac{1}{C_{\text{tc}} C_{\text{err}}} s_{n+1} \leq \eta \left( C_{\text{err}} + \frac{1}{C_{\text{err}}} \right) s_n + \alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right) + \mathbf{err}_n \tag{22a}$$

in the case of 4B and

$$\alpha_n d_{n+1}^2 + \frac{1}{C_{\text{tc}} C_{\text{err}}} s_{n+1} \leq 2\eta C_{\text{err}} s_n + \alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right) + \mathbf{err}_n \tag{22b}$$

in the case of 4A for all  $n \in \mathbb{N}$ .

*Proof* Due to (12) we have

$$\begin{aligned} \mathcal{R}(u_{n+1}) - \mathcal{R}(u^\dagger) &= \mathcal{D}_{\mathcal{R}}^{u^*} \left( u_{n+1}, u^\dagger \right) - \left\langle u^*, u^\dagger - u_{n+1} \right\rangle \\ &\geq d_{n+1}^2 - \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right). \end{aligned} \tag{23}$$

From the minimality condition (4a) with  $u = u^\dagger$  we obtain

$$\begin{aligned} &\alpha_n \left( \mathcal{R}(u_{n+1}) - \mathcal{R}(u^\dagger) \right) + \mathcal{S} \left( g^{\text{obs}}; F(u_n) + F'(u_n; u_{n+1} - u_n) \right) \\ &\leq \mathcal{S} \left( g^{\text{obs}}; F(u_n) + F'(u_n; u^\dagger - u_n) \right), \end{aligned} \tag{24}$$

and putting (23) and (24) together we find that

$$\begin{aligned} &\alpha_n d_{n+1}^2 + \mathcal{S} \left( g^{\text{obs}}; F(u_n) + F'(u_n; u_{n+1} - u_n) \right) \\ &\leq \mathcal{S} \left( g^{\text{obs}}; F(u_n) + F'(u_n; u^\dagger - u_n) \right) + \alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right). \end{aligned} \tag{25}$$

– In the case of 4B we use (8), which yields

$$\begin{aligned} &\alpha_n d_{n+1}^2 + \frac{1}{C_{\text{err}}} \mathcal{T} \left( g^\dagger; F(u_n) + F'(u_n; u_{n+1} - u_n) \right) \\ &\leq C_{\text{err}} \mathcal{T} \left( g^\dagger; F(u_n) + F'(u_n; u^\dagger - u_n) \right) + \alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right) + \mathbf{err}_n \end{aligned}$$

and (9b) with  $v = u^\dagger, u = u_n$  leads to

$$\begin{aligned} &\alpha_n d_{n+1}^2 + \frac{1}{C_{\text{err}}} \mathcal{T} \left( g^\dagger; F(u_n) + F'(u_n; u_{n+1} - u_n) \right) \\ &\leq \eta C_{\text{err}} s_n + \alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right) + \mathbf{err}_n. \end{aligned}$$

By (9b) with  $v = u_{n+1}, u = u_n$  we obtain (22a).

– In the case of 4A we are able to apply (9a) with  $v = u^\dagger, u = u_n$  and (9a) with  $v = u_{n+1}$  and  $u = u_n$  to (25) to conclude

$$\begin{aligned} &\alpha_n d_{n+1}^2 + \frac{1}{C_{\text{tc}}} \mathcal{S} \left( g^{\text{obs}}; F(u_{n+1}) \right) \\ &\leq 2\eta \mathcal{S} \left( g^{\text{obs}}; F(u_n) \right) + C_{\text{tc}} \mathcal{S} \left( g^{\text{obs}}; F(u^\dagger) \right) + \alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right). \end{aligned}$$

Due to (8) and Assumption 2.2 this yields (22b). □

Before we deduce the convergence rates from the recursive error estimates (22) respectively, we note some inequalities for the index functions defined in (15) and their inverses:

*Remark 3.2* 1. We have

$$\varphi \left( \vartheta^{-1} (Ct) \right) \leq \max \left\{ \sqrt{C}, 1 \right\} \varphi \left( \vartheta^{-1} (t) \right) \tag{26}$$

$$\varphi^2 \left( \Theta^{-1} (Ct) \right) \leq \max \left\{ \sqrt{C}, 1 \right\} \varphi^2 \left( \Theta^{-1} (t) \right) \tag{27}$$

for all  $t \geq 0$  and  $C > 0$  if defined, where each inequality follows from two applications of the monotonicity assumption (13) (see [30, Remark 2]).

2. Since  $\varphi$  is concave, we have

$$\varphi (\lambda t) \leq \lambda \varphi (t) \quad \text{for all } t \text{ sufficiently small and } \lambda \geq 1 \tag{28}$$

3. Equation (28) implies the following inequality for all  $t$  sufficiently small and  $\lambda \geq 1$ :

$$\Theta (\lambda t) \leq \lambda^3 \Theta (t) \tag{29}$$

The following induction proof follows along the lines of a similar argument in the proof of [30, Theorem 1]:

**Lemma 3.3** *Let the assumptions of Theorem 2.3 hold. Then an estimate of the kind (22a) implies*

$$d_n \leq C_1 \varphi (\alpha_n), \tag{30}$$

$$s_n \leq C_2 \Theta (\alpha_n) \tag{31}$$

for all  $n \leq n_*$  in case of noisy data and for all  $n \in \mathbb{N}$  in case of exact data where (due to  $\eta$  sufficiently small)

$$C_2 = \max \left\{ 4\beta^2 (C_{tc} C_{err} C_{dec})^3, \frac{2C_{tc} C_{err} C_{dec}^3}{\tau \left( 1 - 2C_{dec}^3 C_{tc} C_{err} \eta \left( C_{err} + \frac{1}{C_{err}} \right) \right)} \right\},$$

$$C_1 = \max \left\{ \sqrt{2\beta^4 C_2}, \sqrt{2 (\eta C_2 (C_{err} + 1/C_{err}) + 1/\tau) C_{dec}} \right\}.$$

Since (22b) is of the same form as (22a) (only the constants differ), (30) and (31) are (with slightly changed constants) also valid under (22b).

*Proof* For  $n = 0$  (30) and (31) are guaranteed by the assumption that  $d_0$  and  $s_0$  are small enough. For the induction step we observe that (22a) together with (18) and the induction hypothesis for  $n \leq n_* - 1$  implies

$$\alpha_n d_{n+1}^2 + \frac{1}{C_{tc} C_{err}} s_{n+1} \leq C_{\eta, \tau} \Theta (\alpha_n) + \alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right)$$

where  $C_{\eta, \tau} = \eta C_2 (C_{err} + 1/C_{err}) + 1/\tau$ . Now we distinguish between two cases:

Case 1  $\alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right) \leq C_{\eta, \tau} \Theta (\alpha_n)$ .

In that case we find

$$\alpha_n d_{n+1}^2 + \frac{1}{C_{tc} C_{err}} s_{n+1} \leq 2C_{\eta, \tau} \Theta (\alpha_n)$$

which by  $\Theta (t) / t = \varphi^2 (t)$ , (28) and (29) implies

$$d_{n+1} \leq \sqrt{2C_{\eta, \tau}} \varphi (\alpha_n) = \sqrt{2C_{\eta, \tau}} \varphi \left( \frac{\alpha_n}{\alpha_{n+1}} \alpha_{n+1} \right) \leq \sqrt{2C_{\eta, \tau}} C_{dec} \varphi (\alpha_{n+1}),$$

$$s_{n+1} \leq 2C_{tc} C_{err} C_{\eta, \tau} \Theta (\alpha_n) \leq 2C_{tc} C_{err} C_{\eta, \tau} C_{dec}^3 \Theta (\alpha_{n+1}).$$

The assertions now follow by  $\sqrt{2C_{\eta, \tau}} C_{dec} \leq C_1$  and  $2C_{tc} C_{err} C_{\eta, \tau} C_{dec}^3 \leq C_2$  which is ensured by the definition of  $C_2$ .

Case 2  $\alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right) > C_{\eta, \tau} \Theta (\alpha_n)$ .

In that case we find

$$\alpha_n d_{n+1}^2 + \frac{1}{C_{tc} C_{err}} s_{n+1} \leq 2\alpha_n \beta d_{n+1} \varphi \left( \frac{s_{n+1}}{d_{n+1}^2} \right).$$

If  $d_{n+1} = 0$ , then this implies  $s_{n+1} = 0$  and hence the assertion is trivial. By multiplying with  $\sqrt{s_{n+1}}$  and dividing by  $d_{n+1}^2$  we have

$$\alpha_n \sqrt{s_{n+1}} + \frac{1}{C_{tc} C_{err}} \frac{s_{n+1}}{d_{n+1}^2} \sqrt{s_{n+1}} \leq 2\beta \alpha_n \vartheta \left( \frac{s_{n+1}}{d_{n+1}^2} \right). \tag{32}$$

Considering only the first term on the left hand side of (32) this is

$$\vartheta^{-1} \left( \frac{\sqrt{s_{n+1}}}{2\beta} \right) \leq \frac{s_{n+1}}{d_{n+1}^2} \tag{33}$$

and by considering only the second term on the left hand side of (32)

$$\Phi \left( \frac{s_{n+1}}{d_{n+1}^2} \right) \sqrt{s_{n+1}} \leq 2\beta C_{tc} C_{err} \alpha_n \tag{34}$$

where  $\Phi (t) = \sqrt{t} / \varphi (t) = t / \vartheta (t)$ . Plugging (33) into (34) using the monotonicity of  $\Phi$  by (13) we find

$$\Phi \left( \vartheta^{-1} \left( \frac{\sqrt{s_{n+1}}}{2\beta} \right) \right) \sqrt{s_{n+1}} \leq 2\beta C_{tc} C_{err} \alpha_n.$$

Since  $\Phi(\vartheta^{-1}(t)) = \vartheta^{-1}(t)/t$  this shows

$$\vartheta^{-1}\left(\frac{\sqrt{s_{n+1}}}{2\beta}\right) \leq C_{\text{tc}}C_{\text{err}}\alpha_n.$$

Hence,

$$s_{n+1} \leq 4\beta^2\Theta(C_{\text{tc}}C_{\text{err}}\alpha_n)$$

which by (29) and  $4\beta^2(C_{\text{dec}}C_{\text{tc}}C_{\text{err}})^3 \leq C_2$  implies  $s_{n+1} \leq C_2\Theta(\alpha_{n+1})$ .

Now from  $\vartheta(t) = \sqrt{t}\varphi(t)$  we find  $b^2\left(\varphi\left(\vartheta^{-1}\left(\frac{\sqrt{a}}{b}\right)\right)\right)^2 = a/\vartheta^{-1}\left(\frac{\sqrt{a}}{b}\right)$  and hence by (33)

$$\begin{aligned} d_{n+1}^2 &\leq 4\beta^2\left(\varphi\left(\vartheta^{-1}\left(\frac{\sqrt{s_{n+1}}}{2\beta}\right)\right)\right)^2 \\ &\leq 4\beta^2\left(\varphi\left(\vartheta^{-1}\left(\frac{\sqrt{C_2}}{2\beta}\vartheta(\alpha_{n+1})\right)\right)\right)^2 \\ &\leq 2\beta\sqrt{C_2}\varphi(\alpha_{n+1})^2 \\ &\leq C_1^2\varphi(\alpha_{n+1})^2 \end{aligned}$$

where we used (26),  $C_2 \geq 4\beta^2$  due to  $C_{\text{dec}}C_{\text{tc}}C_{\text{err}} \geq 1$  and  $\sqrt{2\beta}\sqrt[4]{C_2} \leq C_1$ .

Therefore, we have proven that (30) and (31) hold for all  $n \leq n_*$  (or in case of exact data for all  $n \in \mathbb{N}$ ).

With these two lemmas at hand we are able to complete the Proof of Theorem 2.3: Inserting (18) into (30) and (31) we find using (27)

$$\mathcal{D}_{\mathcal{R}}^{u^*}(u_{n_*}, u^\dagger) \leq C_1\varphi^2(\alpha_{n_*}) = \mathcal{O}\left(\varphi^2\left(\Theta^{-1}(\mathbf{err}_{n_*})\right)\right)$$

and

$$\mathcal{T}\left(g^\dagger; F(u_{n_*})\right) \leq C_2\Theta(\alpha_{n_*}) = \mathcal{O}(\mathbf{err}_{n_*}).$$

### 4 A Lepskii-type stopping rule and additive source conditions

In this section we will present a convergence rates result under the following variational source condition in additive form:

**Assumption 5B** There exists  $u^* \in \partial\mathcal{R}(u^\dagger) \subset \mathcal{X}'$ , parameters  $\beta_1 \in [0, 1/2)$ ,  $\beta_2 > 0$  (later assumed to be sufficiently small), and a strictly concave, differentiable index function  $\varphi$  satisfying  $\varphi'(t) \nearrow \infty$  as  $t \searrow 0$  such that

$$\left\langle u^*, u^\dagger - u \right\rangle \leq \beta_1\mathcal{D}_{\mathcal{R}}^{u^*}(u, u^\dagger) + \beta_2\varphi\left(\mathcal{T}\left(g^\dagger; F(u)\right)\right) \quad \text{for all } u \in \mathfrak{B}. \quad (35)$$

A special case of condition (35), motivated by the *benchmark condition*  $u^* = F[u^\dagger]^* \omega$  was first introduced in [24] to prove convergence rates of Tikhonov-type regularization in Banach spaces (see also [42]). Flemming [16] uses them to prove convergence rates for nonlinear Tikhonov regularization (6) with general  $\mathcal{S}$  and  $\mathcal{R}$ . Bot & Hofmann [11] prove convergence rates for general  $\varphi$  and introduce the use of Young’s inequality which we will apply in the following. Finally, Hofmann & Yamamoto [25] prove equivalence in the Hilbert space case for  $\varphi(t) = \sqrt{t}$  in (10) and (35) (with different  $\varphi$ , cf. [25, Prop. 4.4]) and almost equivalence for  $\varphi(t) = t^\nu$  with  $\nu < \frac{1}{2}$  in (10) (again with different  $\varphi$  in (35), cf. [25, Prop. 6.6 and Prop. 6.8]) under a suitable nonlinearity condition. Latest research results show that a classic Hilbert space source conditions (10), which have natural interpretations in a number of important examples, relates to (35) in a way that one obtains order optimal rates (see [17]). Nevertheless, this can be seen much easier for multiplicative variational source conditions [see (14)].

The additive structure of the variational inequality will facilitate our proof and the result will give us the possibility to apply a Lepskiĭ-type stopping rule. We remark that for  $\varepsilon \neq 0$  in Assumption 2 it is not clear how to formulate an implementable discrepancy principle.

Given  $\varphi$  in (35), we construct the following further index functions as in [11], which will be used in our convergence theorem:

$$\psi(t) = \begin{cases} \frac{1}{\varphi'(\varphi^{-1}(t))} & \text{if } t > 0, \\ 0 & \text{if } t = 0, \end{cases} = \begin{cases} (\varphi^{-1})'(t) & \text{if } t > 0, \\ 0 & \text{if } t = 0, \end{cases} \tag{36a}$$

$$\Psi(t) = \int_0^t \psi^{-1}(s) \, ds, \quad t \geq 0, \tag{36b}$$

$$\Lambda = \inf \left\{ g \mid \sqrt{g} \text{ concave index function, } g(t) \geq \frac{\Psi(t)}{t} \text{ for } t \geq 0 \right\}. \tag{36c}$$

The definition (36c) ensures that  $\sqrt{\Lambda}$  is concave, which by (4b) implies

$$(\Lambda(\alpha_n))^{\frac{1}{q}} \leq C_{\text{dec}}^{\frac{2}{q}} (\Lambda(\alpha_{n-1}))^{\frac{1}{q}} \tag{37}$$

for all  $q \geq 1$  and  $n \in \mathbb{N}$ . Since for linear problems  $\sqrt{\Psi(\alpha_n)/\alpha_n}$  is a bound on the approximation error (see [11]) and since for Tikhonov regularization the approximation error decays at most of the order  $O(\alpha_n)$ , we expect that  $t \mapsto \sqrt{\Psi(t)}/t$  is “asymptotically concave” in the sense that  $\lim_{t \searrow 0} \Lambda(t)t/\Psi(t) = 1$ , so we don’t loose anything by replacing  $\Psi(t)/t$  by  $\Lambda(t)$ . Indeed, it is easy to see that this is the case for logarithmic and Hölder type source conditions with  $\nu \leq 1$ , and in the latter case  $t \mapsto \sqrt{\Psi(t)}/t$  itself is concave everywhere.

**Lemma 4.1** *Let Assumption 1, 2, 3, 4A or 4B and 5B hold true and assume that there exists a uniform upper bound  $\mathbf{err}_n \leq \mathbf{err}$  for the error terms  $\mathbf{err}_n$  in Theorem 2.3. Then, with the notation (20), the error of the iterates  $u_n$  defined by (4) for  $n \geq 1$  can*

be bounded by the sum of an approximation error bound  $\Phi_{\text{app}}(n)$ , a propagated data noise error bound  $\Phi_{\text{noi}}(n)$  and a nonlinearity error bound  $\Phi_{\text{nl}}(n)$ ,

$$d_n^2 \leq \Phi_{\text{nl}}(n) + \Phi_{\text{app}}(n) + \Phi_{\text{noi}}(n) \tag{38}$$

where

$$\begin{aligned} \Phi_{\text{nl}}(n) &:= 2\eta C_{\text{NL}} \frac{s_{n-1}}{\alpha_{n-1}}, \\ \Phi_{\text{app}}(n) &:= 2\beta_2 \Lambda(\alpha_{n-1}), \\ \Phi_{\text{noi}}(n) &:= 2 \frac{\mathbf{err}}{\alpha_{n-1}}. \end{aligned}$$

and  $C_{\text{NL}} := \max\{2C_{\text{err}}, C_{\text{err}} + 1/C_{\text{err}}\}$ . Moreover, if  $\eta$  and  $\beta_2$  are sufficiently small, the estimate

$$\Phi_{\text{nl}}(n) \leq \gamma_{\text{nl}} (\Phi_{\text{noi}}(n) + \Phi_{\text{app}}(n)) \tag{39}$$

holds true with

$$\gamma_{\text{nl}} := \max \left\{ \frac{C_{\text{dec}}^2 \bar{\gamma}}{1 - C_{\text{dec}}^2 \bar{\gamma}}, \frac{\Phi_{\text{nl}}(1)}{\Phi_{\text{app}}(1) + \Phi_{\text{noi}}(1)} \right\}, \quad \bar{\gamma} := \frac{\eta C_{\text{dec}} C_{\text{NL}}}{\frac{1}{C_{\text{tc}} C_{\text{err}}} - \beta_2}.$$

*Proof* Similar to the proof of Lemma 3.1 the assumptions imply the iterative estimate

$$\alpha_n (1 - \beta_1) d_{n+1}^2 + \frac{1}{C_{\text{tc}} C_{\text{err}}} s_{n+1} \leq \eta \left( C_{\text{err}} + \frac{1}{C_{\text{err}}} \right) s_n + \alpha_n \beta_2 \varphi(s_{n+1}) + \mathbf{err}$$

for all  $n \in \mathbb{N}$  in case of of 4B and

$$\alpha_n (1 - \beta_1) d_{n+1}^2 + \frac{1}{C_{\text{tc}} C_{\text{err}}} s_{n+1} \leq 2\eta C_{\text{err}} s_n + \alpha_n \beta_2 \varphi(s_{n+1}) + \mathbf{err}$$

for all  $n \in \mathbb{N}$  in case of 4A. Now Young’s inequality  $ab \leq \int_0^a \psi(t) dt + \int_0^b \psi^{-1}(s) ds$  (cf. [22, Thm. 156]) with the index function  $\psi$  defined in (36a) applied to the second-last term yields

$$\alpha_n \beta_2 \varphi(s_{n+1}) \leq \beta_2 s_{n+1} + \beta_2 \Psi(\alpha_n).$$

This shows that

$$\alpha_n (1 - \beta_1) d_{n+1}^2 + \left( \frac{1}{C_{\text{tc}} C_{\text{err}}} - \beta_2 \right) s_{n+1} \leq \eta C_{\text{NL}} s_n + \beta_2 \Psi(\alpha_n) + \mathbf{err} \tag{40}$$



for all  $n \in \mathbb{N}$  both in case 4A and in case 4B. Together with  $1/(1 - \beta_1) \leq 2$  and  $\frac{\Psi(t)}{t} \leq \Lambda(t)$  this yields

$$d_{n+1}^2 \leq 2\eta C_{NL} \frac{s_n}{\alpha_n} + 2\beta_2 \Lambda(\alpha_n) + 2 \frac{\mathbf{err}}{\alpha_n}.$$

for all  $n \geq 0$  which is by definition (38).

From (40) we conclude that

$$s_{n+1} \leq \frac{\eta C_{NL}}{\frac{1}{C_{lc} C_{err}} - \beta_2} s_n + \frac{\beta_2}{\frac{1}{C_{lc} C_{err}} - \beta_2} \Psi(\alpha_n) + \frac{\mathbf{err}}{\frac{1}{C_{lc} C_{err}} - \beta_2}.$$

Now multiplying by  $2\eta C_{NL}/\alpha_{n+1}$  we find

$$\Phi_{nl}(n + 2) \leq \bar{\gamma} \Phi_{nl}(n + 1) + \bar{\gamma} \Phi_{app}(n + 1) + \bar{\gamma} \Phi_{noi}(n + 1)$$

for all  $n \in \mathbb{N}$ . Now we prove (39) by induction: For  $n = 1$  the assertion is true by the definition of  $\gamma_{nl}$ . Now let (39) hold for some  $n$ . Then by the inequality above, the induction hypothesis, (37), and the monotonicity of  $\Phi_{noi}$  we find that

$$\begin{aligned} \Phi_{nl}(n + 1) &\leq \bar{\gamma} \Phi_{nl}(n) + \bar{\gamma} \Phi_{app}(n) + \bar{\gamma} \Phi_{noi}(n) \\ &\leq \bar{\gamma} (1 + \gamma_{nl}) (\Phi_{app}(n) + \Phi_{noi}(n)) \\ &\leq C_{dec}^2 \bar{\gamma} (1 + \gamma_{nl}) (\Phi_{app}(n + 1) + \Phi_{noi}(n + 1)). \end{aligned}$$

The definition of  $\gamma_{nl}$  implies  $C_{dec}^2 \bar{\gamma} (1 + \gamma_{nl}) \leq \gamma_{nl}$  and hence the assertion is shown. □

Lemma 4.1 allows us to apply the Lepskii balancing principle as developed in [5,6,35,36] as a posteriori stopping rule. Since the balancing principle requires a metric on  $\mathcal{X}$  we assume that there exists a constant  $C_{bd} > 0$  and an exponent  $q > 1$  such that

$$\|u - u^\dagger\|_{\mathcal{X}}^q \leq C_{bd} \mathcal{D}_{\mathcal{R}}^{u^*}(u, u^\dagger) \quad \text{for all } u \in \mathcal{X}. \tag{41}$$

As already mentioned in (11) this is the case for  $\mathcal{R}(u) = \|u\|_{\mathcal{X}}^q$  with a  $q$ -convex Banach space  $\mathcal{X}$  and  $1 < q < \infty$ , but (41) can be valid also for other choices of  $\mathcal{R}$ .

Now it follows from Lemma 4.1 by inserting (41) and taking the  $q$ -th root that

$$\|u_n - u^\dagger\|_{\mathcal{X}} \leq C_{bd}^{\frac{1}{q}} \left( \Phi_{nl}(n)^{\frac{1}{q}} + \Phi_{app}(n)^{\frac{1}{q}} + \Phi_{noi}(n)^{\frac{1}{q}} \right).$$

Whereas  $\Phi_{app}$  and  $\Phi_{nl}$  are typically unknown, it is important to note that the error component  $\Phi_{noi}$  is known if an error bound **err** is available. Therefore, the following

Lepskiĭ balancing principle can be implemented:

$$N_{\max} := \min \left\{ n \in \mathbb{N} \mid C_{\text{bd}}^{\frac{1}{q}} \Phi_{\text{noi}}(n)^{\frac{1}{q}} \geq 1 \right\} \tag{42a}$$

$$n_{\text{bal}} := \min \left\{ n \in \{1, \dots, N_{\max}\} \mid \forall m \geq n \ \| u_n - u_m \|_{\mathcal{X}} \leq c \Phi_{\text{noi}}^{\frac{1}{q}}(m) \right\} \tag{42b}$$

Moreover, it is important to note that  $\Phi_{\text{noi}}$  is increasing and  $\Phi_{\text{app}}$  is decreasing. Therefore, the general theory developed in the references above can be applied, and we obtain the following convergence result:

**Theorem 4.2** (Convergence rates under Assumption 5B) *Let the assumptions of Lemma 4.1 hold true and assume that  $\mathcal{D}_{\mathcal{R}}^{u^*}(u_0, u^\dagger)$  and  $\mathcal{S}(g^\dagger; F(u_0))$  are sufficiently small.*

1. Exact data:

*Then the iterates  $(u_n)$  defined by (4) with exact data  $g^{\text{obs}} = g^\dagger$  fulfill*

$$\mathcal{D}_{\mathcal{R}}^{u^*}(u_n, u^\dagger) = \mathcal{O}(\Lambda(\alpha_n)), \quad n \rightarrow \infty. \tag{43}$$

2. A priori stopping rule:

*For noisy data and the stopping rule*

$$n_* := \min \{ n \in \mathbb{N} \mid \Psi(\alpha_n) \leq \mathbf{err} \}$$

*with  $\Psi$  defined in (36b) we obtain the convergence rate*

$$\mathcal{D}_{\mathcal{R}}^{u^*}(u_{n_*}, u^\dagger) = \mathcal{O}\left(\Lambda\left(\Psi^{-1}(\mathbf{err})\right)\right), \quad \mathbf{err} \rightarrow 0. \tag{44}$$

3. Lepskiĭ-type stopping rule:

*Assume that (41) holds true for some  $q > 1$ . Then the Lepskiĭ balancing principle (42b) with  $c = C_{\text{bd}}^{\frac{1}{q}} 4(1 + \gamma_{\text{nl}})$  leads to the convergence rate*

$$\|u_{n_{\text{bal}}} - u^\dagger\|_{\mathcal{X}}^q = \mathcal{O}\left(\Lambda\left(\Psi^{-1}(\mathbf{err})\right)\right), \quad \mathbf{err} \rightarrow 0.$$

*Proof* By (38) and (39) we find  $d_n^2 \leq (1 + \gamma_{\text{nl}})(\Phi_{\text{app}}(n) + \Phi_{\text{noi}}(n))$  which implies part 1 and

$$d_{n_*}^2 \leq (1 + \gamma_{\text{nl}}) \left( 2\beta_2 \Lambda(\alpha_{n_*-1}) + 2 \frac{\mathbf{err}}{\alpha_{n_*-1}} \right).$$

Using the definition of  $n_*$  and (37) we have

$$\frac{\mathbf{err}}{\alpha_{n_*-1}} \leq \frac{\Psi(\alpha_{n_*-1})}{\alpha_{n_*-1}} \leq \Lambda(\alpha_{n_*-1}) \leq C_{\text{dec}}^2 \Lambda(\alpha_{n_*}).$$

Using the definition of  $n_*$  again we obtain  $\alpha_{n_*} \leq \Psi^{-1}(\mathbf{err})$ . Putting these estimates together yields (44).

To prove part 3 assume that  $\mathbf{err}$  is sufficiently small in the following. We use again  $d_n^2 \leq (1 + \gamma_{nl}) (\Phi_{\text{app}}(n) + \Phi_{\text{noi}}(n))$ , which yields by (41) the estimate

$$\|u_n - u^\dagger\|_{\mathcal{X}} \leq C_{\text{bd}}^{\frac{1}{q}} (1 + \gamma_{nl})^{\frac{1}{q}} \left( \Phi_{\text{app}}(n)^{\frac{1}{q}} + \Phi_{\text{noi}}(n)^{\frac{1}{q}} \right)$$

for all  $n \in \{1, \dots, N_{\text{max}}\}$ . Define  $\psi(j) := 2C_{\text{bd}}^{\frac{1}{q}} (1 + \gamma_{nl})^{\frac{1}{q}} \Phi_{\text{noi}}(N_{\text{max}} + 1 - j)$  and  $\phi(j) := 2C_{\text{bd}}^{\frac{1}{q}} (1 + \gamma_{nl})^{\frac{1}{q}} \Phi_{\text{app}}(N_{\text{max}} + 1 - j)$  and note that  $\phi(1) \leq \psi(1)$  if and only if  $\Phi_{\text{app}}(N_{\text{max}}) \leq 1$ . This is the case if  $N_{\text{max}}$  is sufficiently large which holds true for sufficiently small  $\mathbf{err}$  as assumed. Thus by (37) we can apply [35, Cor. 1] to gain

$$\|u_{n_{\text{bal}}} - u^\dagger\|_{\mathcal{X}} \leq 6(1 + \gamma_{nl})^{\frac{1}{q}} C_{\text{dec}}^{\frac{2}{q}} C_{\text{bd}}^{\frac{1}{q}} \min_{n \leq N_{\text{max}}} \left( \Phi_{\text{app}}(n)^{\frac{1}{q}} + \Phi_{\text{noi}}(n)^{\frac{1}{q}} \right).$$

If we can show that  $n_* \in \{1, \dots, N_{\text{max}}\}$  we obtain the assertion as in part 2. Since by definition  $\alpha_{n_*-1} > \Psi^{-1}(\mathbf{err})$ , we have

$$\Phi_{\text{noi}}(n_*) = 2 \frac{\mathbf{err}}{\alpha_{n_*-1}} < 2 \frac{\mathbf{err}}{\Psi^{-1}(\mathbf{err})} \leq 2\Lambda \left( \Psi^{-1}(\mathbf{err}) \right)$$

and hence  $n_* \leq N_{\text{max}}$  if  $\mathbf{err}$  is sufficiently small. □

### 5 Relation to previous results

The most commonly used source conditions are Hölder-type and logarithmic source conditions, which correspond to

$$\varphi_\nu(t) := t^\nu, \quad \nu \in (0, 1/2], \tag{45a}$$

$$\bar{\varphi}_p(t) := \begin{cases} (-\ln(t))^{-p} & \text{if } 0 < t \leq \exp(-p-1), \\ 0 & \text{if } t = 0, \end{cases} \quad p > 0, \tag{45b}$$

respectively. For a number of inverse problems such source conditions have been shown to be equivalent to natural smoothness assumptions on the solution in terms of Sobolev space regularity (see [15,27]). We have restricted the range of Hölder indices to  $\nu \in (0, 1/2]$  since for  $\nu > 1/2$  the monotonicity assumption (13) is violated. By computing the second derivative, one can easily see that the functions  $\bar{\varphi}_p$  are concave on the interval  $[0, \exp(-p-1)]$ , and condition (13) is trivial. If necessary, the functions  $\bar{\varphi}_p$  can be extended to concave functions on  $[0, \infty)$  by suitable affine linear function on  $(\exp(-p-1), \infty)$ .

We note the explicit form of the abstract error estimates (19) for these classes of source conditions as a corollary:

**Corollary 5.1** (Hölder and logarithmic source conditions) *Suppose the assumptions of Theorem 2.3 hold true.*

1. *If  $\varphi$  in (12) is of the form (45a) and  $n_* := \min \left\{ n \in \mathbb{N} \mid \alpha_n \leq \tau \mathbf{err}_n^{\frac{1}{1+2\nu}} \right\}$  with  $\tau \geq 1$  sufficiently large, then*

$$\mathcal{D}_{\mathcal{R}}^{u_*} \left( u_{n_*}, u^\dagger \right) = \mathcal{O} \left( \mathbf{err}_{n_*}^{\frac{2\nu}{1+2\nu}} \right). \tag{46a}$$

2. *If  $\varphi = \bar{\varphi}_p$ ,  $\bar{n}_* := \min \left\{ n \in \mathbb{N} \mid \alpha_n^2 \leq \tau \mathbf{err}_n \right\}$  and  $\tau \geq 1$  sufficiently large, then*

$$\mathcal{D}_{\mathcal{R}}^{u_*} \left( u_{\bar{n}_*}, u^\dagger \right) = \mathcal{O} \left( \bar{\varphi}_{2p} \left( \mathbf{err}_{\bar{n}_*} \right) \right). \tag{47a}$$

*Proof* In the case of Hölder source conditions we already remarked that the conditions in Assumption 5A are satisfied  $\nu \in (0, 1/2]$ , and we have  $\Theta(t) = t^{1+2\nu}$ ,  $\Theta^{-1}(\xi) = \xi^{1/(1+2\nu)}$ .

In the case of logarithmic source conditions we have  $\Theta(t) = t \cdot \bar{\varphi}_{2p}(t)$ . The function  $\Theta^{-1}$  does not have an algebraic representation, but its asymptotic behavior at 0 can be computed:  $\Theta^{-1}(t) = \frac{t}{\bar{\varphi}_{2p}(t)} (1 + o(1))$  as  $t \searrow 0$ . This implies that  $\bar{\varphi}_p(\Theta^{-1}(t)) = \bar{\varphi}_p(t) (1 + o(1))$  as  $t \searrow 0$ . Note that the proposed stopping rule  $\bar{n}_*$ , which can be implemented without knowledge of the smoothness index  $p$ , deviates from the stopping rule

$$n_* := \min \left\{ n \in \mathbb{N} \mid \alpha_n \bar{\varphi}_{2p}(\alpha_n) \leq \tau \mathbf{err}_n \right\}$$

proposed in Theorem 2.3. Asymptotically we have  $n_* > \bar{n}_*$ , and hence (16) holds for  $n = \bar{n}_*$ . Therefore, we still get the optimal rates since

$$\mathcal{D}_{\mathcal{R}}^{u_*} \left( u_{\bar{n}_*}, u^\dagger \right) = \mathcal{O} \left( \bar{\varphi}_{2p} \left( \alpha_{\bar{n}_*} \right) \right) = \mathcal{O} \left( \bar{\varphi}_{2p} \left( \sqrt{\tau \mathbf{err}_{\bar{n}_*}} \right) \right) = \mathcal{O} \left( \bar{\varphi}_{2p} \left( \mathbf{err}_{\bar{n}_*} \right) \right).$$

□

Recall from Sect. 2 that we can choose

$$\mathbf{err} \equiv \delta^r \quad \text{if} \quad \|g^{\text{obs}} - g^\dagger\|_{\mathcal{Y}} \leq \delta \quad \text{and} \quad \mathcal{S}(g_2; g_1) = \|g_1 - g_2\|_{\mathcal{Y}}^r, \quad \mathcal{T} = \mathcal{S}$$

with  $r \in [1, \infty)$ . In particular, if  $\mathcal{X}$  and  $\mathcal{Y}$  are Hilbert spaces,  $r = 2$  and  $\mathcal{R} = \|u - u_0\|^2$  for some  $u_0 \in \mathcal{X}$ , then (46a) and (47a) translate into the rates

$$\begin{aligned} \|u_{n_*} - u\| &= \mathcal{O} \left( \delta^{\frac{2\nu}{1+2\nu}} \right), \\ \|u_{n_*} - u\| &= \mathcal{O} \left( (-\ln \delta)^{-p} \right), \end{aligned}$$

respectively, for  $\delta \rightarrow 0$  (see, e.g., [31]), which are known to be optimal for linear inverse problems.

It remains to discuss the relation of Assumption 4 to the standard tangential cone condition:

**Lemma 5.2** (tangential cone condition) *Let  $\mathcal{S}(g_2; g_1) = \mathcal{T}(g_2; g_1) = \|g_1 - g_2\|_{\mathcal{Y}}^r$ . If  $F$  fulfills the tangential cone condition*

$$\|F(u) + F'(u; v - u) - F(v)\|_{\mathcal{Y}} \leq \bar{\eta} \|F(u) - F(v)\|_{\mathcal{Y}} \quad \text{for all } u, v \in \mathfrak{B} \quad (48)$$

with  $\bar{\eta} \geq 0$  sufficiently small, then Assumptions 4A and 4B are satisfied.

*Proof* Using the inequality  $(a + b)^r \leq 2^{r-1}(a^r + b^r)$ ,  $a, b \geq 0$  we find that

$$\begin{aligned} & \|F(u) + F'(u; v - u) - g\|_{\mathcal{Y}}^r \\ & \leq \left( \|F(u) + F'(u; v - u) - F(v)\|_{\mathcal{Y}} + \|F(v) - g\|_{\mathcal{Y}} \right)^r \\ & \leq 2^{r-1} \bar{\eta}^r \|F(u) - F(v)\|_{\mathcal{Y}}^r + 2^{r-1} \|F(v) - g\|_{\mathcal{Y}}^r \\ & \leq 2^{2r-2} \bar{\eta}^r \|F(u) - g\|_{\mathcal{Y}}^r + \left( 2^{r-1} + \bar{\eta}^r 2^{2r-2} \right) \|F(v) - g\|_{\mathcal{Y}}^r. \end{aligned}$$

Moreover, with  $|a - b|^r \geq 2^{1-r}a^r - b^r$ ,  $a, b \geq 0$  we get

$$\begin{aligned} & \|F(u) + F'(u; v - u) - g\|_{\mathcal{Y}}^r \\ & \geq \left| \|F(v) - g\|_{\mathcal{Y}} - \|F(u) + F'(u; v - u) - F(v)\|_{\mathcal{Y}} \right|^r \\ & \geq 2^{1-r} \|F(v) - g\|_{\mathcal{Y}}^r - \bar{\eta}^r \|F(u) - F(v)\|_{\mathcal{Y}}^r \\ & \geq 2^{1-r} \|F(v) - g\|_{\mathcal{Y}}^r - 2^{r-1} \bar{\eta}^r \|F(u) - g\|_{\mathcal{Y}}^r - 2^{r-1} \bar{\eta}^r \|F(v) - g\|_{\mathcal{Y}}^r \\ & = \left( 2^{1-r} - 2^{r-1} \bar{\eta}^r \right) \|F(v) - g\|_{\mathcal{Y}}^r - 2^{r-1} \bar{\eta}^r \|F(u) - g\|_{\mathcal{Y}}^r \end{aligned}$$

for all  $g \in \mathcal{Y}$ . Hence, (9) holds true with  $\eta = 2^{2r-2} \bar{\eta}^r$  and

$$C_{\text{tc}} = \max \left\{ \frac{1}{2^{1-r} - 2^{r-1} \bar{\eta}^r}, 2^{r-1} + \bar{\eta}^r 2^{2r-2} \right\} \geq 1$$

if  $\bar{\eta}$  is sufficiently small. □

### 6 Convergence analysis for Poisson data

In this section we discuss the application of our results to inverse problems with Poisson data. We first describe a natural continuous setting involving Poisson processes (see e.g. [1]). The relation to the finite dimensional setting discussed in the introduction is described at the end of this section.

Recall that a Poisson process with intensity  $g^\dagger \in L^1(\mathbb{M})$  on some submanifold  $\mathbb{M} \subset \mathbb{R}^d$  can be described as a random finite set of points  $\{x_1, \dots, x_N\} \subset \mathbb{M}$  written as random measure  $G = \sum_{n=1}^N \delta_{x_n}$  such that the following conditions are satisfied:

1. For all measurable subsets  $M' \subset M$  the number  $G(M') = \#\{n : x_n \in M'\}$  is Poisson distributed with mean  $\int_{M'} g^\dagger dx$ .
2. For disjoint measurable subsets  $M'_1, \dots, M'_m \subset M$  the random variables  $G(M'_1), \dots, G(M'_m)$  are stochastically independent.

Actually, the first condition can be replaced by the weaker assumption that  $\mathbf{E}G(M') = \int_{M'} g^\dagger dx$ . In photonic imaging  $g^\dagger$  will describe the photon density on the measurement manifold  $M$ , and  $x_1, \dots, x_N$  will denote the positions of the detected photons. For a Poisson process  $G$  with intensity  $g^\dagger$  and a measurable function  $\psi : M \rightarrow \mathbb{R}$  the following equalities hold true whenever the integrals on the right hand sides exist (see [32]):

$$\mathbf{E} \int_M \psi dG = \int_M \psi g^\dagger dx, \quad \mathbf{Var} \int_M \psi dG = \int_M \psi^2 g^\dagger dx \tag{49}$$

We also introduce an exposure time  $t > 0$ . Our convergence results will describe reconstruction errors in the limit  $t \rightarrow \infty$ . Assume the data  $\tilde{G}_t$  are drawn from a Poisson process with intensity  $t g^\dagger$  and define  $G_t := \frac{1}{t} \tilde{G}_t$ . The negative log-likelihood functional is given by

$$\mathcal{S}(G_t; g) = \begin{cases} \int_M g dx - \int_M \ln g dG_t = \int_M g dx - \frac{1}{t} \sum_{n=1}^N \ln g(x_n), & g \geq 0 \\ \infty, & \text{else.} \end{cases} \tag{50}$$

Using (49) we obtain the following formulas for the mean and variance of  $\mathcal{S}(G_t; g)$  if the integrals on the right hand side exist:

$$\mathbf{E}\mathcal{S}(G_t; g) = \int_M [g - g^\dagger \ln g] dx, \quad \mathbf{Var} \mathcal{S}(G_t; g) = \frac{1}{t} \int_M (\ln g)^2 g^\dagger dx. \tag{51}$$

The term  $\mathfrak{s}(g^\dagger) = \mathbf{E}\mathcal{S}(G_t; g^\dagger) = \int_M [g^\dagger - g^\dagger \ln g^\dagger] dx$  is finite if  $g^\dagger \in L^1(M) \cap L^\infty(M)$ ,  $M$  is bounded, and  $g^\dagger \geq 0$  as assumed below. Setting

$$\mathcal{T}(g^\dagger; g) := \mathbb{KL}(g^\dagger; g) := \begin{cases} \int_M [g - g^\dagger - g^\dagger \ln g + g^\dagger \ln g^\dagger] dx, & g \geq 0 \\ \infty, & \text{else,} \end{cases} \tag{52}$$

we find that Assumption 2 holds true with  $C_{\text{err}} = 1$  and

$$\mathbf{err}(g) := \begin{cases} \left| \int_M \ln(g) (dG_t - g^\dagger dx) \right|, & g \geq 0 \\ 0, & \text{else.} \end{cases} \tag{53}$$

This motivates the following assumption:

**Assumption  $\mathcal{P}$**  With the notation of Assumption 1 assume that

1.  $\mathbb{M}$  is a compact submanifold of  $\mathbb{R}^d$ ,  $\mathcal{Y} := L^1(\mathbb{M}) \cap C(\mathbb{M})$  with norm  $\|g\|_{\mathcal{Y}} := \|g\|_{L^1} + \|g\|_{\infty}$  and

$$F(u) \geq 0 \quad \text{for all } u \in \mathfrak{B}.$$

2. For a subset  $\tilde{\mathcal{Y}} \subset \mathcal{Y}$  specified later there exist constants  $\rho_0, t_0 > 0$  and a strictly monotonically decreasing function  $\zeta : (\rho_0, \infty) \rightarrow [0, 1]$  fulfilling  $\lim_{\rho \rightarrow \infty} \zeta(\rho) = 0$  such that the concentration inequality

$$\mathbf{P} \left( \sup_{g \in \tilde{\mathcal{Y}}} \left| \int_{\mathbb{M}} \ln(g) (dG_t - g^\dagger dx) \right| \geq \frac{\rho}{\sqrt{t}} \right) \leq \zeta(\rho) \tag{54}$$

holds for all  $\rho > \rho_0$  and all  $t > t_0$ .

Concerning the second part of this assumption we cite the following result:

**Theorem 6.1** ([46, Thm. 2.1]) *Let  $\mathbb{M} \subset \mathbb{R}^d$  be a bounded domain with Lipschitz boundary and suppose  $s > \frac{d}{2}$ . For  $R \geq 1$  consider the ball  $B_s(R) := \{g \in H^s(\mathbb{M}) \mid \|g\|_{H^s(\mathbb{M})} \leq R\}$ . Then there exists a constant  $C_{\text{conc}} > 0$  depending only on  $\mathbb{M}, s$  and  $\|g^\dagger\|_{L^1(\mathbb{M})}$  such that (54) holds true with  $\tilde{\mathcal{Y}} = B_s(R)$ ,  $\zeta(\rho) = \exp\left(-\frac{\rho}{RC_{\text{conc}}}\right)$ ,  $\rho_0 = RC_{\text{conc}}$  and  $t_0 = 1$ .*

The proof of this result is based on a concentration inequality by Reynaud-Bouret [41]. Unfortunately, her result requires  $\|\ln(g)\|_{\infty} \leq b$  for all  $g \in \tilde{\mathcal{Y}}$ , and this restriction is carried over to Theorem 6.1 as  $s > \frac{d}{2}$  implies by Sobolev’s embedding theorem  $\|\ln(g)\|_{\infty} \leq C$  for all  $g \in \tilde{\mathcal{Y}}$ . To apply this result, we thus need to assume that  $\|\ln(F(u))\|_{\infty} \leq C$  for all  $u \in \mathfrak{B}$ , which does not allow for zeros of  $F(u)$  even on sets of measure 0 if  $F(u)$  is continuous and is hence very restrictive. Therefore, we introduce the following shifted version of the Kullback–Leibler divergence (3) involving an offset parameter  $\sigma \geq 0$  and a side-constraint  $g \geq -\frac{\sigma}{2}$ :

$$\mathcal{T} \left( g^\dagger; g \right) := \begin{cases} \mathbb{KL}(g^\dagger + \sigma; g + \sigma) & \text{if } g \geq -\frac{\sigma}{2} \\ \infty & \text{otherwise.} \end{cases} \tag{55}$$

Note that (52) and (55) coincide for  $\sigma = 0$ . Correspondingly, we choose

$$\mathcal{S}(G_t; g) := \begin{cases} \int_{\mathbb{M}} [g - \sigma \ln(g + \sigma)] dx - \int_{\mathbb{M}} \ln(g + \sigma) dG_t & \text{if } g \geq -\frac{\sigma}{2}, \\ \infty & \text{else} \end{cases} \tag{56}$$

as data misfit functional in (4a). Setting  $\mathfrak{s}(g^\dagger) := \int_{\mathbb{M}} [g^\dagger - (g^\dagger + \sigma) \ln(g^\dagger + \sigma)] dx$ , Assumption 2 is satisfied with

$$\mathbf{err}(g) := \begin{cases} \left| \int_{\mathbb{M}} \ln(g + \sigma) (dG_t - g^\dagger dx) \right|, & g \geq -\frac{\sigma}{2}, \\ 0 & \text{else.} \end{cases} \tag{57}$$

*Remark 6.2* (Assumptions 5A and 5B (source conditions)) Using the inequality

$$\|g_1 - g_2\|_{L^2}^2 \leq \left( \frac{4}{3} \|g_1\|_{L^\infty} + \frac{2}{3} \|g_2\|_{L^\infty} \right) \mathbb{K}\mathbb{L}(g_2; g_1)$$

(see [10, Lemma 2.2 (a)]), Assumption 5A/B with  $\mathcal{T}(g_1; g_2) = \|g_1 - g_2\|_{L^2}^2$  imply Assumption 5A/B with  $\mathcal{T}(g_1; g_2) = \mathbb{K}\mathbb{L}(g_1; g_2)$  if  $F(\mathfrak{B})$  is bounded in  $L^\infty(\mathbb{M})$ . However, Assumptions 5A/B with  $\mathcal{T}(g_1; g_2) = \mathbb{K}\mathbb{L}(g_1; g_2)$  may be fulfilled with a better index function  $\varphi$  if  $F(u^\dagger)$  is close to 0 in parts of the domain.

Before we state our convergence result, we introduce the smallest concave function larger than the rate function in Theorem 4.2:

$$\hat{\varphi} := \inf \left\{ \tilde{\varphi} \mid \tilde{\varphi} \text{ concave index function, } \tilde{\varphi}(s) \geq \Lambda \left( \Psi^{-1}(s) \right) \text{ for } s \geq 0 \right\}. \tag{58}$$

From the case of Hölder-type source conditions we expect that  $\hat{\varphi}$  will typically coincide with  $\Lambda \circ \Psi^{-1}$  at least in a neighborhood of 0 (see e.g. [25, Prop. 4.3]).

**Corollary 6.3** *Let the Assumptions 1, 3 and 5B hold true. Moreover, assume that one of the following conditions is satisfied:*

- Assumptions 4A and  $\mathcal{P}$  hold true with  $\mathcal{S}$  and  $\mathcal{T}$  given by (50) and (52) and  $\tilde{\mathcal{Y}} = F(\mathfrak{B})$ .
- Assumptions 4B and  $\mathcal{P}$  hold true with  $\mathcal{T}$  and  $\mathcal{S}$  given by (55) and (56) and

$$\begin{aligned} \tilde{\mathcal{Y}} := & \{F(u) + \sigma : u \in \mathfrak{B}\} \\ & \cup \left\{ F(u) + F'(u; v - u) + \sigma : u, v \in \mathfrak{B}, F(u) + F'(u; v - u) \geq -\frac{\sigma}{2} \right\}. \end{aligned}$$

Suppose that  $\beta_2$  is sufficiently small,  $\mathfrak{B}$  is bounded and  $\mathcal{R}$  is chosen such that (41) holds true for some  $q > 1$ , and Lepskiĭ’s balancing principle (42) is applied with  $c = C_{\text{bd}}^{\frac{1}{q}} 4(1 + \gamma_{\text{nl}})$  and  $\mathbf{err} = \frac{\tau \zeta^{-1}(1/\sqrt{t})}{\sqrt{t}}$  with a sufficiently large parameter  $\tau$  (a lower will be given in the proof). Then we obtain the following convergence rate in expectation:

$$\mathbf{E} \left\| u_{n_{\text{bal}}} - u^\dagger \right\|_{\mathcal{X}}^q \leq \mathcal{O} \left( \hat{\varphi} \left( \frac{\zeta^{-1}(1/\sqrt{t})}{\sqrt{t}} \right) \right), \quad t \rightarrow \infty. \tag{59}$$

*Proof* In the case of Assumption 4A and  $\sigma = 0$ , we find that Assumption 2 holds true with  $\mathbf{err}$  defined by (53). Assumption  $\mathcal{P}$  implies that the terms  $\mathbf{err}_n$  defined by (17a) in Theorem 2.3 satisfy

$$\mathbf{P} \left[ \sup_{n \in \mathbb{N}_0} \mathbf{err}_n \leq \frac{\tau \rho}{\sqrt{t}} \right] \geq 1 - \zeta(\rho) \tag{60}$$



for all  $\rho > \rho_0$  and  $t > t_0$  with  $\tau := 1 + 2\eta C_{\text{tc}} + C_{\text{tc}}$  due to  $C_{\text{err}} = 1$ . To show the analogous estimate in the case of Assumption 4B, recall that Assumption 2 holds true with  $\mathbf{err}$  defined by (57). From the variational characterization of  $u_{n+1}$  it follows that

$$F(u_n) + F'(u_n; u_{n+1} - u_n) \geq -\frac{\sigma}{2} \tag{61}$$

Moreover, from Assumption 4B we conclude that

$$F(u_n) + F'(u_n; u^\dagger - u_n) \geq -\frac{\sigma}{2} \tag{62}$$

This yields the inequality (60) with  $\tau := 2$  also for  $\mathbf{err}_n$  defined by (17b) using Assumption  $\mathcal{P}$ .

By virtue of (60) the sets  $E_\rho := \left\{ \sup_{n \in \mathbb{N}_0} \mathbf{err}_n \leq \frac{\tau \rho}{\sqrt{t}} \right\}$  have probability  $\geq 1 - \zeta(\rho)$  if  $\rho > \rho_0$ . Recall that  $\zeta$  is monotonically decreasing and define  $\rho(t) := \zeta^{-1}(1/\sqrt{t})$  where we assume  $t$  to be sufficiently large. We have

$$\begin{aligned} \mathbf{E} \left\| u_{\text{nbal}} - u^\dagger \right\|_{\mathcal{X}}^q &\leq 2^q \left( \max_{E_{\rho(t)}} \left\| u_{\text{nbal}} - u^\dagger \right\|_{\mathcal{X}}^q \cdot \mathbf{P}(E_{\rho(t)}) \right. \\ &\quad \left. + \sup_{u, v \in \mathfrak{B}} \|u - v\|_{\mathcal{X}}^q \mathbf{P}(E_{\rho(t)}^c) \right). \end{aligned} \tag{63}$$

Now we can apply Theorem 4.2 to obtain the error bound

$$\max_{E_{\rho(t)}} \left\| u_{\text{nbal}} - u^\dagger \right\|_{\mathcal{X}}^q \leq C_1 \hat{\varphi}(\mathbf{err}) \leq C_1 \tau \hat{\varphi} \left( \frac{\zeta^{-1}(1/\sqrt{t})}{\sqrt{t}} \right)$$

with some constant  $C_1 > 0$  for all sufficiently large  $t$ . In the last inequality we have used the concavity of  $\hat{\varphi}$ . Plugging this into (63) yields

$$\mathbf{E} \left\| u_{\text{nbal}} - u^\dagger \right\|_{\mathcal{X}}^q \leq 2^q \left( C_1 \tau \hat{\varphi} \left( \frac{\zeta^{-1}(1/\sqrt{t})}{\sqrt{t}} \right) + \frac{1}{\sqrt{t}} \sup_{u, v \in \mathfrak{B}} \|u - v\|_{\mathcal{X}}^q \right).$$

Since  $\hat{\varphi}$  is concave, there exists  $C_2 > 0$  such that  $s \leq C_2 \hat{\varphi}(s)$  for all sufficiently small  $s > 0$ . Moreover,  $\frac{1}{\sqrt{t}}$  in the second term is bounded by  $\frac{1}{\rho_0} \frac{\zeta^{-1}(1/\sqrt{t})}{\sqrt{t}}$ , and thus we obtain the assertion (59). □

If  $\zeta(\rho) = \exp(-c\rho)$  for some  $c > 0$  as discussed above, then our convergence rates result (59) means that we have to pay a logarithmic factor for adaptation to unknown smoothness by the Lepskiĭ principle. It is known (see [44]) that in some cases such a logarithmic factor is inevitable.

The most important issue is the verification of Assumption  $\mathcal{P}$ . In case of Assumption 4A this follows from the results discussed above only under the restrictive assumption that  $F(u)$  is uniformly bounded away from 0 for all  $u \in \mathfrak{B}$ . On the other hand for

the case of Assumption 4B we find from Theorem 6.1 that Assumption  $\mathcal{P}$  is satisfied under the mild condition

$$\sup_{u, v \in \mathfrak{B}} \|F(u) + F'(u, v - u)\|_{H^s} < \infty.$$

**Binning** Let us discuss the relation between the discrete data model discussed in the introduction and the continuous model above. Consider a decomposition of the measurement manifold  $\mathbb{M}$  into  $J$  measurable disjoint subdomains (bins) of positive measure  $|\mathbb{M}_j| > 0$ :

$$\mathbb{M} = \bigcup_{j=1}^J \mathbb{M}_j$$

In practice each  $\mathbb{M}_j$  may correspond to a detector counting the number of photons in  $\mathbb{M}_j$ , so the measured data are

$$\underline{g}_j^{\text{obs}} = tG_t(\mathbb{M}_j) = \#\{n \mid x_n \in \mathbb{M}_j\}, \quad j = 1, \dots, J.$$

Consider the linear operator  $S_J : L^1(\mathbb{M}) \rightarrow \mathbb{R}^J$ ,  $(S_J g)_j := \int_{\mathbb{M}_j} g \, dx$  and the mapping  $S_J^* \underline{g} := \sum_{j=1}^J |\mathbb{M}_j|^{-1} \underline{g}_j \mathbf{1}_{\mathbb{M}_j}$ , which is adjoint to  $S_J$  with respect to the  $L^2(\mathbb{M})$  inner product and the inner product  $\langle \underline{g}, \underline{h} \rangle := \sum_{j=1}^J |\mathbb{M}_j|^{-1} \underline{g}_j \underline{h}_j$ .  $P_J := S_J^* S_J$  is the  $L^2$ -orthogonal projection onto the subspace of functions, which are constant on each  $\mathbb{M}_j$ .  $S_J$  can naturally be extended to measures such that  $(S_J(G_t))_j = G_t(\mathbb{M}_j) = \frac{1}{t} \#\{n : x_n \in \mathbb{M}_j\}$ . For distinction we denote the right hand sides of Eqs. (2) and (3) by  $\underline{\mathcal{S}}_J$  and  $\underline{\mathbb{K}\mathbb{L}}_J$ , and define  $\mathcal{S}_\infty$  and  $\mathbb{K}\mathbb{L}_\infty$  by (50) and (52). Then

$$\underline{\mathcal{S}}_J(\underline{g}^{\text{obs}}; \underline{g}) = \mathcal{S}_\infty(S_J^* \underline{g}^{\text{obs}}; S_J^* \underline{g}) \quad \text{and} \quad \underline{\mathbb{K}\mathbb{L}}_J(\underline{g}^\dagger; \underline{g}) = \mathbb{K}\mathbb{L}_\infty(S_J^* \underline{g}^\dagger; S_J^* \underline{g}).$$

The discrete data model above can be treated in the framework of our analysis by choosing

$$\mathcal{S}(g^{\text{obs}}; g) := \underline{\mathcal{S}}_J\left(\frac{1}{t} g^{\text{obs}}; S_J g\right),$$

$\mathfrak{s}(g^\dagger) := \underline{\mathcal{S}}_J(S_J g^\dagger; S_J g^\dagger)$ , and  $\mathcal{T} := \mathbb{K}\mathbb{L}_\infty$ . Then Assumption 2 holds true with

$$\begin{aligned} \mathbf{err}(g) := & \left| \sum_{j=1}^J \ln((S_J g)_j) \left( \frac{1}{t} g_j^{\text{obs}} - (S_J g^\dagger)_j \right) \right| \\ & + \left| \mathbb{K}\mathbb{L}_\infty(g^\dagger; g) - \mathbb{K}\mathbb{L}_\infty(P_J g^\dagger; P_J g) \right| \end{aligned} \tag{64}$$

if  $S_J g \geq 0$ ,  $\{j : (S_J g)_j = 0, (S_J g^\dagger)_j + g_j^{\text{obs}} > 0\} = \emptyset$  and  $\mathbf{err}(g) := \infty$  else. To achieve convergence, the binning has to be refined as  $t \rightarrow \infty$ . The binning should be

chosen such that the second term on the right hand side of (64) (the discretization error) is dominated by the first term (the stochastic error) such that the reconstruction error is determined by the number of observed photons rather than discretization effects.

### 7 Applications and computed examples

**Solution of the convex subproblems** We first describe a simple strategy to minimize the convex functional (4a) with  $\mathcal{S}$  as defined in (56) in each Newton step. For the moment we neglect the side condition  $g \geq -\sigma/2$  in (56). For simplicity we further assume that  $\mathcal{R}$  is quadratic, e.g.  $\mathcal{R}(u) = \|u - u_0\|^2$ . We approximate  $\mathcal{S}(g^{\text{obs}}; g + h)$  by the second order Taylor expansion

$$\mathcal{S}^{(2)}[g^{\text{obs}}; g](h) := \mathcal{S}(g^{\text{obs}}; g) + \int_{\mathbb{M}} \left[ \left(1 - \frac{g^{\text{obs}} + \sigma}{g + \sigma}\right) h + \frac{1}{2} \frac{g^{\text{obs}} + \sigma}{(g + \sigma)^2} h^2 \right] dx$$

and define an inner iteration

$$h_{n,l} := \operatorname{argmin}_h \left[ \mathcal{S}^{(2)} \left[ g^{\text{obs}}; F(u_n) + F'[u_n](u_{n,l} - u_n); \right] (h) + \alpha_n \mathcal{R}(u_{n,l} + h) \right] \tag{65}$$

for  $l = 0, 1, \dots$  with  $u_{n,0} := u_n$  and  $u_{n,l+1} := u_{n,l} + s_{n,l} h_{n,l}$ . Here the step-length parameter  $s_{n,l}$  is chosen as the largest  $s \in [0, 1]$  for which  $s F'[u_n] \geq -\eta\sigma - F(u_n)$  with a tuning parameter  $\eta \in [0, 1]$  (typically  $\eta = 0.9$ ). This choice of  $s_{n,l}$  ensures that  $F(u_n) + F'[u_n](u_{n,l+1} - u_n) \geq -\eta\sigma$ , i.e. (65) is a reasonable approximation to (4a), and  $\eta = 1/2$  ensures that  $u_{n,l+1}$  satisfies the side condition in (56). It follows from the first order optimality conditions, which are necessary and sufficient due to strict convexity here, that  $u_{n,l} = u_{n,l+1}$  is the exact solution  $u_{n+1}$  of (4a) if  $h_{n,l} = 0$ . Therefore, we stop the inner iteration if  $\|h_{n,l}\|/\|h_{n,0}\|$  is sufficiently small. We also stop the inner iteration if  $s_{n,l}$  is 0 or too small.

Simplifying and omitting terms independent of  $h$  we can write (65) as a least squares problem

$$h_{n,l} = \operatorname{argmin}_h \left[ \int_{\mathbb{M}} \frac{1}{2} \left( \frac{\sqrt{g^{\text{obs}} + \sigma}}{g_{n,l} + \sigma} F'[u_n] h + \frac{g_{n,l} - g^{\text{obs}}}{\sqrt{g^{\text{obs}} + \sigma}} \right)^2 dx + \alpha_n \mathcal{R}(u_{n,l} + h) \right] \tag{66}$$

with  $g_{n,l} := F(u_n) + F'[u_n](u_{n,l} - u_n)$ . (66) is solved by the CG method applied to the normal equation.

In the examples below we observed fast convergence of the inner iteration (65). In the phase retrieval problem we had problems with the convergence of the CG iteration when  $\alpha_n$  becomes too small. If the offset parameter  $\sigma$  becomes too small or if  $\sigma = 0$  convergence deteriorates in general. This is not surprising since the iteration (65) cannot be expected to converge to the exact solution  $u_{n+1}$  of (4a) if the

side condition  $F(u_n) + F'(u_n; u_{n+1} - u_n) \geq -\sigma/2$  is active at  $u_{n+1}$ . The design of efficient algorithms for this case will be addressed in future research.

**An inverse obstacle scattering problem without phase information** The scattering of polarized, transverse magnetic (TM) time harmonic electromagnetic waves by a perfect cylindrical conductor with smooth cross section  $D \subset \mathbb{R}^2$  is described by the equations

$$\Delta u + k^2 u = 0, \quad \text{in } \mathbb{R}^2 \setminus D, \tag{67a}$$

$$\frac{\partial u}{\partial n} = 0, \quad \text{on } \partial D, \tag{67b}$$

$$\lim_{r \rightarrow \infty} \sqrt{r} \left( \frac{u_s}{r} - ik u_s \right) = 0, \quad \text{where } r := |x|, u_s := u - u_i. \tag{67c}$$

Here  $D$  is compact,  $\mathbb{R}^2 \setminus D$  is connected,  $n$  is the outer normal vector on  $\partial D$ , and  $u_i = \exp(ikx \cdot d)$  is a plane incident wave with direction  $d \in \{x \in \mathbb{R}^2 : |x| = 1\}$ . This is a classical obstacle scattering problems, and we refer to the monograph [14] for further details and references. The Sommerfeld radiation condition (67c) implies the asymptotic behavior

$$u_s(x) = \frac{\exp(ik|x|)}{\sqrt{|x|}} \left( u_\infty \left( \frac{x}{|x|} \right) + \mathcal{O} \left( \frac{1}{|x|} \right) \right)$$

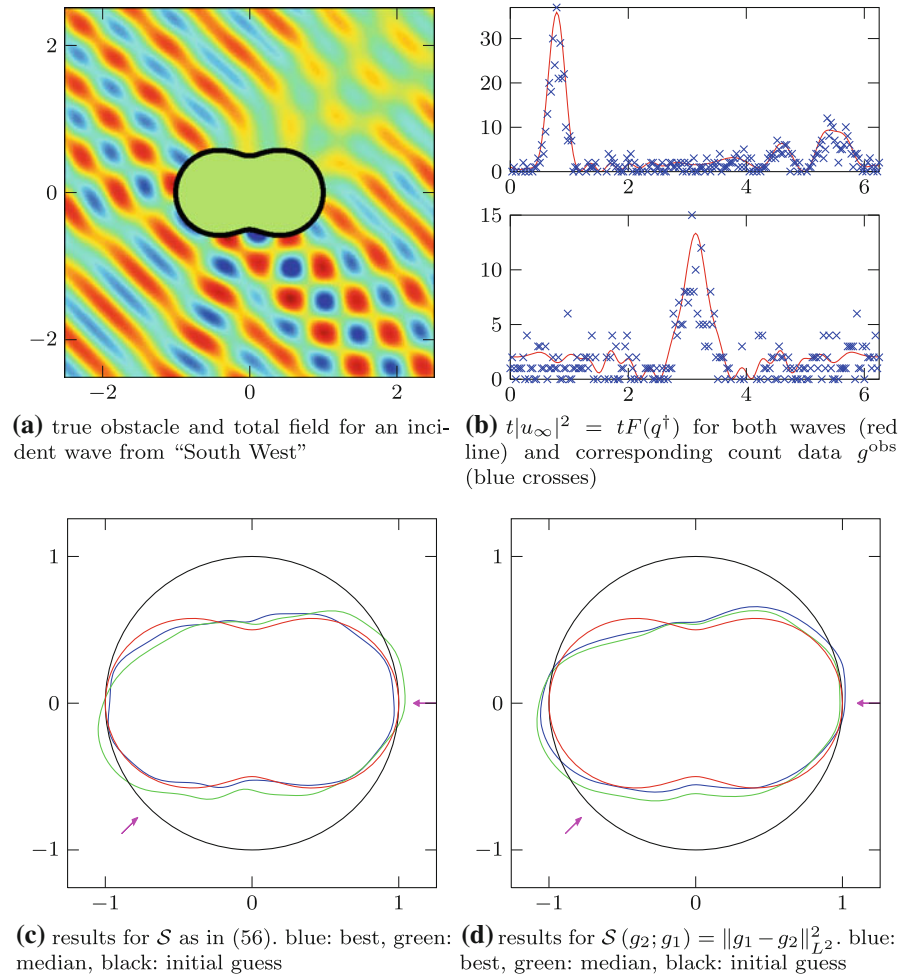
as  $|x| \rightarrow \infty$ , and  $u_\infty$  is called the *far field pattern* or *scattering amplitude* of  $u_s$ .

We consider the inverse problem to recover the shape of the obstacle  $D$  from photon counts of the scattered electromagnetic field far away from the obstacle. Since the photon density is proportional to the squared absolute value of the electric field, we have no immediate access to the phase of the electromagnetic field. Since at large distances the photon density is approximately proportional to  $|u_\infty|^2$ , our inverse problem is described by the operator equation

$$F(\partial D) = |u_\infty|^2. \tag{68}$$

A similar problem is studied with different methods and noise models by Ivanyshyn and Kress [29]. Recall that  $|u_\infty|$  is invariant under translations of  $\partial D$ . Therefore, it is only possible to recover the shape, but not the location of  $D$ . For plottings we always shift the center of gravity of  $\partial D$  to the origin. We assume that  $D$  is star-shaped and represent  $\partial D$  by a periodic function  $q$  such that  $\partial D = \{q(t)(\cos t, \sin t)^\top : t \in [0, 2\pi]\}$ . For details on the implementation of  $F$ , its derivative and adjoint we refer to [26] where the mapping  $q \mapsto u_\infty$  is considered as forward operator. Even in this situation where the phase of  $u_\infty$  is given in addition to its modulus, it has been shown in [26] that for Sobolev-type smoothness assumptions at most logarithmic rates of convergence can be expected.

As a test example we choose the obstacle shown in Fig. 1 described by  $q^\dagger(t) = \frac{1}{2} \sqrt{3 \cos^2 t + 1}$  with two incident waves from ‘‘South West’’ and from ‘‘East’’ with wave number  $k = 10$  as shown in Fig. 1. We used  $J = 200$  equidistant bins. The initial guess for the Newton iteration is the unit circle described by  $q_0 \equiv 1$ , and we

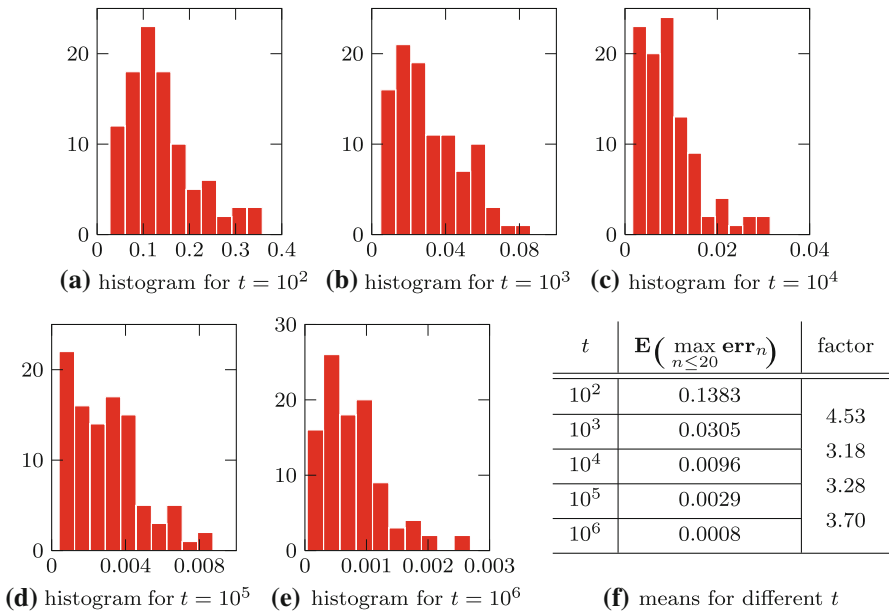


**Fig. 1** Numerical results for the inverse obstacle scattering problem (68). **c, d** Best and median reconstruction from 100 experiments with  $t = 1,000$  expected counts. See also Table 1

choose the Sobolev norm  $\mathcal{R}(q) = \|q - q_0\|_{H^s}^2$  with  $s = 1.6$  as penalty functional. The regularization parameters are chosen as  $\alpha_n = 0.5 \cdot (2/3)^n$ . Moreover, we choose an initial offset parameter  $\sigma = 0.002$ , which is reduced by  $\frac{4}{5}$  in each iteration step. The inner iteration (65) is stopped when  $\|h_{n,l}\|/\|h_{n,0}\| \leq 0.1$ , which was usually the case after about three iterations (or about five iterations for  $\|h_{n,l}\|/\|h_{n,0}\| \leq 0.01$ ).

For comparison we take the usual IRGNM, i.e. (4) with  $\mathcal{S}(\hat{g}; g) = \|g - \hat{g}\|_{L^2}^2$  and  $\mathcal{R}$  as above as well as a weighted IRGNM where  $\mathcal{S}$  is chosen to be Pearson’s  $\phi^2$ -distance:

$$\phi^2(g^{\text{obs}}; g) = \int_{\mathbb{M}} \frac{|g - g^{\text{obs}}|^2}{g^{\text{obs}}} dx. \tag{69}$$



**Fig. 2** Overview for the error terms (17b) for the inverse scattering problem. For different values of the expected total number of counts the value  $\max_{n \leq 20} \mathbf{err}_n$  has been calculated in 100 experiments. The figure shows the corresponding histograms and means. The decay of order  $\frac{1}{\sqrt{t}}$ , i.e. reduction by a factor of  $\sqrt{10} \approx 3.16$  in the table is clearly visible. All parameters are as in Fig. 1

Since in all our examples we have many zero counts, we actually used

$$S(g^{\text{obs}}; g) = \phi_c^2(g^{\text{obs}}; g) = \int_{\mathbb{M}} \frac{|g - g^{\text{obs}}|^2}{\max\{g^{\text{obs}}, c\}} dx.$$

with a cutoff-parameter  $c > 0$ .

Figure 2 lists histograms and empiric means of the error terms (17b) and shows the decay of order  $1/\sqrt{t}$  in accordance with the theoretic result from Theorem 6.1.

Error statistics of shape reconstructions from 100 experiments are shown in Table 1. The stopping index  $N$  is chosen a priori such that (the empirical version of) the expectation  $\mathbf{E}\|q_n - q^\dagger\|_{L^2}^2$  is minimal for  $n = N$ , i.e. we compare both methods with an oracle stopping rule. Note that the mean square error is significantly smaller for the Kullback–Leibler divergence than for the  $L^2$ -distance and also clearly smaller than for Pearson’s distance. Moreover the distribution of the error is more concentrated for the Kullback–Leibler divergence. For Pearson’s  $\phi^2$  distance it must be said that the results depend strongly on the cutoff parameter for the data. In our experiments  $c = 0.2$  seemed to be a good choice in general.

**A phase retrieval problem** A well-known class of inverse problems with numerous applications in optics consists in reconstructing a function  $f : \mathbb{R}^d \rightarrow \mathbb{C}$  from the mod-

**Table 1**  $L^2$ -error statistics for the inverse obstacle scattering problem (68)

$t$	$\mathcal{S}(g^{\text{obs}}; g)$	N	$\sqrt{\mathbf{E}\ q_N - q^\dagger\ _{L^2}^2}$	$\sqrt{\mathbf{Var}\ q_N - q^\dagger\ _{L^2}^2}$
100	$\ g - g^{\text{obs}}\ _{L^2}^2$	7	0.124	0.033
	$\phi_c^2(g^{\text{obs}}; g)$	2	0.122	0.018
	$\mathcal{S}$ in Eq. (56)	3	0.091	0.025
1,000	$\ g - g^{\text{obs}}\ _{L^2}^2$	9	0.106	0.014
	$\phi_c^2(g^{\text{obs}}; g)$	7	0.091	0.012
	$\mathcal{S}$ in Eq. (56)	5	0.070	0.017
10,000	$\ g - g^{\text{obs}}\ _{L^2}^2$	9	0.105	0.004
	$\phi_c^2(g^{\text{obs}}; g)$	23	0.076	0.048
	$\mathcal{S}$ in Eq. (56)	5	0.050	0.005

The log-likelihood functional (56) is compared to the standard  $L^2$  and Pearson’s  $\phi^2$  distance [cf. (69)] for different values of the expected total number of counts  $t$  with 100 experiments for each set of parameters. The error of the initial guess is  $\|q_0 - q^\dagger\|_{L^2} = 0.288$ . All parameters as in Fig. 1

ulus of its Fourier transform  $|\mathcal{F}f|$  and additional a priori information, or equivalently to reconstruct the phase  $\mathcal{F}f/|\mathcal{F}f|$  of  $\mathcal{F}f$  (see Hurt [28]).

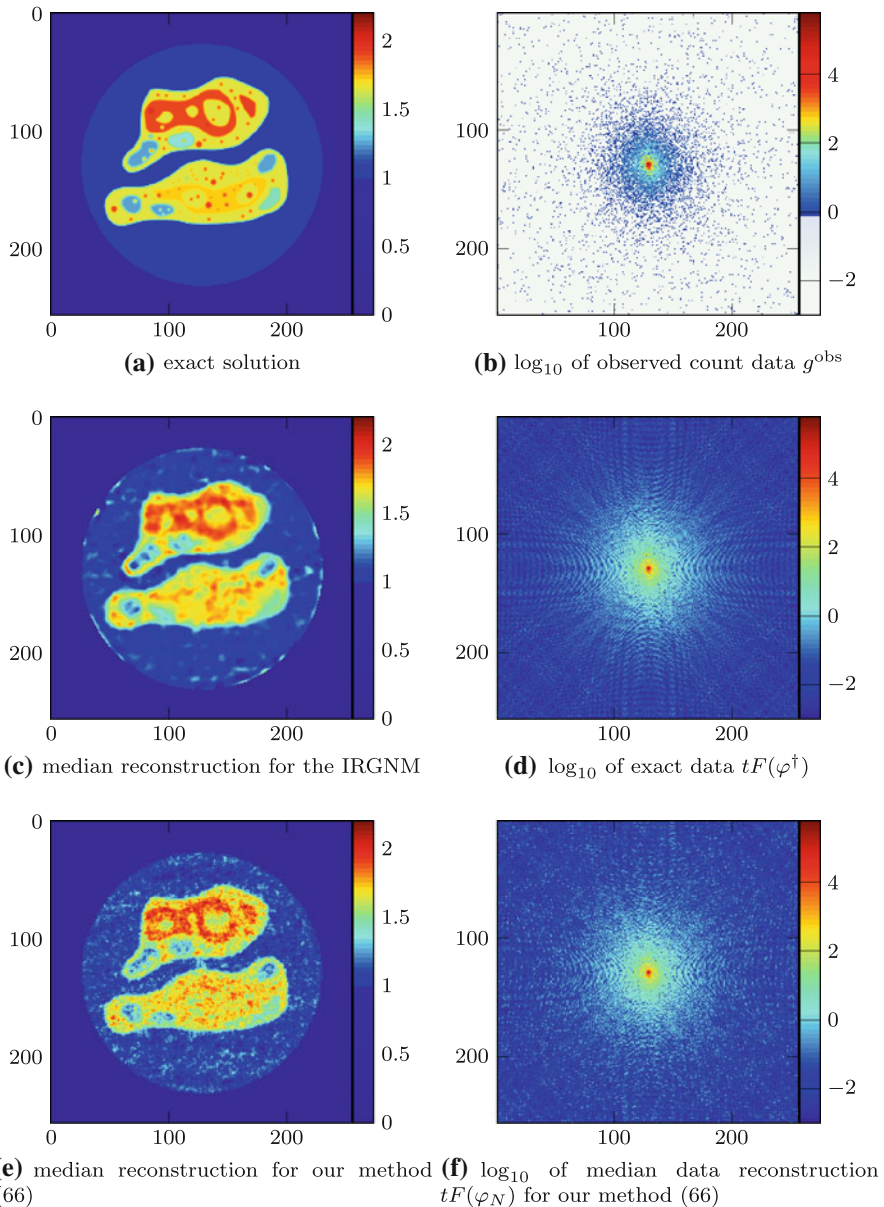
In the following we assume more specifically that  $f : \mathbb{R}^2 \rightarrow \mathbb{C}$  is of the form  $f(x) = \exp(i\varphi(x))$  with an unknown real-valued function  $\varphi$  with known compact support  $\text{supp}(\varphi)$ . For a uniqueness result we refer to Klibanov [33], although not all assumptions of this theorem are satisfied in the example below. It turns out to be particularly helpful if  $\varphi$  has a jump of known magnitude at the boundary of its support. We will assume that  $\text{supp} \varphi = B_\rho = \{x \in \mathbb{R}^2 : |x| \leq \rho\}$  and that  $\varphi \approx \chi_{B_\rho}$  close to the boundary  $\partial B_\rho$  (here  $\chi_{B_\rho}$  denotes the characteristic function of  $B_\rho$ ). This leads to an inverse problem where the forward operator is given by

$$F : H^s(B_\rho) \longrightarrow L^\infty(\mathbb{M}),$$

$$(F\varphi)(\xi) := \left| \int_{B_\rho} e^{-i\xi \cdot x} e^{i\varphi(x)} dx \right|^2. \tag{70}$$

Here  $H^s(B_\rho)$  denotes a Sobolev space with index  $s \geq 0$  and  $\mathbb{M} \subset \mathbb{R}^2$  is typically of the form  $\mathbb{M} = [-\kappa, \kappa]^2$ . The a priori information on  $\varphi$  can be incorporated in the form of an initial guess  $\varphi_0 \equiv 1$ . Note that the range of  $F$  consists of analytic functions.

The problem above occurs in optical imaging: If  $f(x') = \exp(i\varphi(x')) = u(x', 0)$  ( $x' = (x_1, x_2)$ ) denotes the values of a cartesian component  $u$  of an electric field in the plane  $\{x \in \mathbb{R}^3 : x_3 = 0\}$  and  $u$  solves the Helmholtz equation  $\Delta u + k^2 u = 0$  and a radiation condition in the half-space  $\{x \in \mathbb{R}^3 : x_3 > 0\}$ , then the intensity  $g(x') = |u(x', \Delta)|^2$  of the electric field at a measurement plane  $\{x \in \mathbb{R}^3 : x_3 = \Delta\}$  in the limit  $\Delta \rightarrow \infty$  in the *Fraunhofer approximation* is given by  $|\mathcal{F}_2 f|^2$  up to rescaling (see e.g. Paganin [38, Sec. 1.5]). If  $f$  is generated by a plane incident wave in  $x_3$  direction passing through a non-absorbing, weakly scattering object of interest in the half-space  $\{x_3 < 0\}$  close to the plane  $\{x_3 = 0\}$  and if the wave length is



**Fig. 3** Median reconstructions for the phase retrieval problem with  $t = 10^6$  expected counts

small compared to the length scale of the object, then the *projection approximation*  $\varphi(x') \approx \frac{k}{2} \int_{-\infty}^0 (n^2(x', x_3) - 1) dx_3$  is valid where  $n$  describes the refractive index of the object of interest (see e.g. [38, Sec. 2.1]). A priori information on  $\varphi$  concerning a jump at the boundary of its support can be obtained by placing a known transparent object before or behind the object of interest.



The simulated test object in Fig. 3 which represents two cells is taken from Giewekemeyer et al. [18]. We choose the initial guess  $\varphi_0 \equiv 1$ , the Sobolev index  $s = \frac{1}{2}$ , and the regularization parameters  $\alpha_n = \frac{5}{10^6} \cdot (2/3)^n$ . The photon density is approximated by  $J = 256^2$  bins. The offset parameter  $\sigma$  is initially set to  $2 \cdot 10^{-6}$  and reduced by a factor  $\frac{4}{5}$  in each iteration step. As for the scattering problem, we use an oracle stopping rule  $N := \operatorname{argmin}_n \mathbf{E} \|\varphi_n - \varphi^\dagger\|_{L^2}^2$ . As already mentioned, we had difficulties to solve the quadratic minimization problems (66) by the CG method for small  $\alpha_n$  and had to stop the iterations before residuals were sufficiently small to guarantee a reliable solution.

Nevertheless, comparing subplots (c) and (e) in Fig. 3, the median KL-reconstruction (e) seems preferable (although more noisy) since the contours are sharper and details in the interior of the cells are more clearly separated.

**Acknowledgments** We would like to thank Tim Salditt and Klaus Giewekemeyer for helpful discussions and data concerning the phase retrieval problem, Patricia Reynaud-Bouret for fruitful discussions on concentration inequalities, and two anonymous referees for their suggestions, which helped to improve the paper considerably. Financial support by the German Research Foundation DFG through SFB 755, the Research Training Group 1023 and the Federal Ministry of Education and Research (BMBF) through the project INVERS is gratefully acknowledged.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

## References

1. Antoniadis, A., Bigot, J.: Poisson inverse problems. *Ann. Stat.* **34**(5), 2132–2158 (2006)
2. Bakushinskiĭ, A.B.: The problem of the convergence of the iteratively regularized Gauss–Newton method. *Comput. Math. Math. Phys.* **32**(9), 1353–1359 (1992)
3. Bakushinskiĭ, A.B., Kokurin, M.Y.: *Iterative Methods for Approximate Solution of Inverse Problems*. Springer, Berlin (2004)
4. Bardsley, J.M.: A theoretical framework for the regularization of Poisson likelihood estimation problems. *Inverse Probl. Imaging* **4**, 11–17 (2010)
5. Bauer, F., Hohage, T.: A Lepskij-type stopping rule for regularized Newton methods. *Inverse Probl.* **21**(6), 1975 (2005)
6. Bauer, F., Hohage, T., Munk, A.: Iteratively regularized Gauss–Newton method for nonlinear inverse problems with random noise. *SIAM J. Numer. Anal.* **47**(3), 1827–1846 (2009)
7. Benning, M., Burger, M.: Error estimates for general fidelities. *Electron. Trans. Numer. Anal.* **38**, 44–68 (2011)
8. Bertero, M., Boccacci, P., Desiderà, G., Vicidomini, G.: Image deblurring with Poisson data: from cells to galaxies. *Inverse Probl.* **25**(12), 123006 (2009)
9. Blaschke, B., Neubauer, A., Scherzer, O.: On convergence rates for the Iteratively regularized Gauss–Newton method. *IMA J. Numer. Anal.* **17**(3), 421–436 (1997)
10. Borwein, J.M., Lewis, A.S.: Convergence of best entropy estimates. *SIAM J. Optim.* **1**, 191–205 (1991)
11. Bot, R.I., Hofmann, B.: An extension of the variational inequality approach for nonlinear ill-posed problems. *J. Integr. Equ. Appl.* **22**(3), 369–392 (2010)
12. Brune, C., Sawatzky, A., Burger, M.: Primal and dual Bregman methods with application to optical nanoscopy. *Int. J. Comput. Vis.* **92**(2), 211–229 (2011)
13. Burger, M., Osher, S.: Convergence rates of convex variational regularization. *Inverse Probl.* **20**(5), 1411–1422 (2004)
14. Colton, D., Kress, R.: *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd edn. Springer, Berlin (1997)
15. Engl, H., Hanke, M.: *A. Springer, Neubauer. Regularization of Inverse Problems* (1996)
16. Flemming, J.: Theory and examples of variational regularisation with non-metric fitting functionals. *J. Inverse Ill Posed Probl.* **18**(6), 677–699 (2010)

17. Flemming, J.: Generalized Tikhonov regularization—basic theory and comprehensive results on convergence rates. PhD thesis, Chemnitz University of Technology (2011)
18. Giewekemeyer, K., Krüger, S.P., Kalbfleisch, S., Bartels, M., Beta, C., Salditt, T.: X-ray propagation microscopy of biological cells using waveguides as a quasipoint source. *Phys. Rev. A* **83**, 023804 (2011)
19. Grasmair, M.: Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Probl.* **26**, 115014 (2010)
20. Hanke, M.: A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Probl.* **13**, 79–95 (1997)
21. Hanke, M., Neubauer, A., Scherzer, O.: A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.* **72**, 21–37 (1995)
22. Hardy, G.H., Littlewood, J.E., Polya, G.: *Inequalities*. Cambridge University Press, Cambridge (1967)
23. Hegland, M.: Variable Hilbert scales and their interpolation inequalities with applications to Tikhonov regularization. *Appl. Anal.* **59**(1–4), 207–223 (1995)
24. Hofmann, B., Kaltenbacher, B., Pöschl, C., Scherzer, O.: A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Probl.* **23**(3), 987–1010 (2007)
25. Hofmann, B., Yamamoto, M.: On the interplay of source conditions and variational inequalities for nonlinear ill-posed problems. *Appl. Anal.* **89**(11), 1705–1727 (2010)
26. Hohage, T.: Convergence rates of a regularized Newton method in sound-hard inverse scattering. *SIAM J. Numer. Anal.* **36**, 125–142 (1998)
27. Hohage, T.: Regularization of exponentially ill-posed problems. *Numer. Funct. Anal. Optim.* **21**, 439–464 (2000)
28. Hurt, N.E.: Phase retrieval and zero crossings, volume 52 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht (1989)
29. Ivanyshyn, O., Kress, R.: Identification of sound-soft 3D obstacles from phaseless data. *Inverse Probl. Imaging* **4**(1), 131–149 (2010)
30. Kaltenbacher, B., Hofmann, B.: Convergence rates for the iteratively regularized Gauss–Newton method in Banach spaces. *Inverse Probl.* **26**(3), 035007 (2010)
31. Kaltenbacher, B., Neubauer, A., Scherzer, O.: *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, volume 6 of *Radon Series on Computational and Applied Mathematics*. de Gruyter (2008)
32. Kingman, J.F.C.: *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press/Oxford University Press, New York (1993)
33. Klibanov, M.V.: On the recovery of a 2-D function from the modulus of its Fourier transform. *J. Math. Anal. Appl.* **323**(2), 818–843 (2006)
34. Massart, P.: *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin (2007)
35. Mathé, P.: The Lepskiï principle revisited. *Inverse Probl.* **22**(3), L11–L15 (2006)
36. Mathé, P., Pereverzev, S.: Geometry of ill-posed problems in variable Hilbert scales. *Inverse Probl.* **19**, 789–803 (2003)
37. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation-based image restoration. *Multiscale Model. Simul.* **4**(2), 460–489 (electronic) (2005)
38. Paganin, D.: *Coherent X-Ray Optics*. Oxford University Press, Oxford (2006)
39. Pöschl, C.: *Tikhonov Regularization with General Residual Term*. PhD thesis, Universität Innsbruck (2008)
40. Resmerita, E., Scherzer, O.: Error estimates for non-quadratic regularization and the relation to enhancement. *Inverse Probl.* **22**(3), 801 (2006)
41. Reynaud-Bouret, P.: Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Relat. Fields* **126**(1), 103–153 (2003)
42. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: *Variational Methods in Imaging*. Applied Mathematical Sciences. Springer, Berlin (2008)
43. Stück, R., Burger, M., Hohage, T.: The iteratively regularized Gauß–Newton method with convex constraints and applications in 4Pi microscopy. *Inverse Probl.* **28**, 015012 (2012)
44. Tsybakov, A.: On the best rate of adaptive estimation in some inverse problems. *C. R. Acad. Sci. Paris* **330**, 835–840 (2000)
45. Vardi, Y., Shepp, L. A., Kaufman, L.: A statistical model for positron emission tomography. *J. Am. Stat. Assoc.*, **80**(389), 8–37 (1985) (with discussion)

46. Werner, F., Hohage, T.: Convergence rates in expectation for Tikhonov-type regularization of Inverse Problems with Poisson data. *Inverse Probl.* **28**, 104004 (2012)
47. Xu, Z.B., Roach, G.F.: Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces. *J. Math. Anal. Appl.* **157**(1), 189–210 (1991)