SIAM J. NUMER. ANAL.
Vol. 47, No. 3, pp. 1827–1846

ITERATIVELY REGULARIZED GAUSS–NEWTON METHOD FOR NONLINEAR INVERSE PROBLEMS WITH RANDOM NOISE[*]

FRANK BAUER[†], THORSTEN HOHAGE[‡], AND AXEL MUNK[§]

Abstract. We study the convergence of regularized Newton methods applied to nonlinear operator equations in Hilbert spaces if the data are perturbed by random noise. It is shown that the expected square error is bounded by a constant times the minimax rates of the corresponding linearized problem if the stopping index is chosen using a priori knowledge of the smoothness of the solution. For unknown smoothness the stopping index can be chosen adaptively based on Lepskiĭ's balancing principle. For this stopping rule we establish an oracle inequality, which implies order optimal rates for deterministic errors, and optimal rates up to a logarithmic factor for random noise. The performance and the statistical properties of the proposed method are illustrated by Monte Carlo simulations.

Key words. iterative regularization methods, nonlinear statistical inverse problems, convergence rates, a posteriori stopping rules, oracle inequalities

AMS subject classifications. 65J22, 62G99, 65J20, 65J15

DOI. 10.1137/080721789

1. Introduction.

In this paper we study the solution of nonlinear ill-posed operator equations

$$F(a) = u, \tag{1.1}$$

assuming that the exact data u are perturbed by random noise. Here $F : D(F) \subset \mathcal{X} \to \mathcal{Y}$ is a nonlinear operator between separable Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ which is Fréchet differentiable on its domain $D(F)$.

Whereas the solution of nonlinear operator equations by iterative regularization methods with deterministic errors has been studied intensively over the last decade (see Bakushinskiĭ and Kokurin [2] and the references therein), we are not aware of any convergence and convergence rate results of iterative regularization methods for nonlinear inverse problems with random noise, although in many practical examples iterative methods are frequently applied where the noise is random rather than deterministic. Vice versa, nonlinear regularization methods are rarely used in a statistical context, and our aim is to explore the potential use of these methods in classical fields of applications such as econometrics [22] and financial statistics [10]. In particular, we will derive rates of convergence under rather general assumptions.

We will consider the case that the error in the data consists of both deterministic and stochastic parts, but we are mainly interested in the situation where the stochastic noise is dominant. More precisely, we assume that the measured data u^{obs} are

[*] Received by the editors April 24, 2008; accepted for publication (in revised form) January 26, 2009; published electronically May 6, 2009. This research was supported by the Graduiertenkolleg 1023, the DFG research group FOR 916, and the Upper Austrian Technology and Research Promotion.

http://www.siam.org/journals/sinum/47-3/72178.html

[†] Department of Mathematics, Fuzzy Logic Laboratorium Linz-Hagenberg, University of Linz, 4232 Hagenberg, Austria (frank.bauer@jku.at).

[‡] Institut für Numerische und Angewandte Mathematik, Georg-August Universität Göttingen, D-37083 Göttingen, Germany (hohage@math.uni-goettingen.de).

[§] Institut für Mathematische Stochastik, Georg-August Universität Göttingen, D-37077 Göttingen, Germany (munk@math.uni-goettingen.de).

described by a Hilbert space process

$$u^{\mathrm{obs}} = F(a^\dagger) + \delta\eta + \sigma\xi, \tag{1.2a}$$

where $\delta \geq 0$ describes the deterministic error level, $\eta \in \mathcal{Y}$, $\|\eta\| \leq 1$ denotes the normalized deterministic error, $\sigma^2 \geq 0$ determines the variance of the stochastic noise, and ξ is a Hilbert space process in $\mathcal{Y}$. We assume that $\xi_\varphi := \langle \xi, \varphi \rangle$ is a random variable with $\mathbf{E}\left[\xi_\varphi\right] = 0$ and $\mathbf{Var}\xi_\varphi < \infty$ for any test vector $\varphi \in \mathcal{Y}$, and that the covariance operator $\mathbf{Cov}_\xi : \mathcal{Y} \to \mathcal{Y}$, characterized by $\langle \mathbf{Cov}_\xi \varphi, \psi \rangle_{\mathcal{Y}} = \mathbf{E}\left[\xi_\varphi \xi_\psi\right]$, satisfies the normalization condition

$$\|\mathbf{Cov}_\xi\|_{\mathcal{Y}} \leq 1. \tag{1.2b}$$

Note that in the case of white noise ($\mathbf{Cov}_\xi = I_{\mathcal{Y}}$) the noise ξ is not in the Hilbert space with probability 1; nevertheless this prominent situation is covered in our setting in the weak formulation as above. For a further discussion of the noise model (1.2) we refer to [5], where it is shown that it incorporates several discrete noise models where the data consists of a vector of n measurements of a function u at different points. In this case σ is proportional to $n^{-1/2}$; i.e., the limit $n \to \infty$ corresponds to the limit $\sigma \to 0$. A particular discrete noise model will be discussed in section 5.

In this paper we provide a convergence analysis for the class of generalized Gauss-Newton methods given by

$$\hat{a}_{k+1} := a_0 + g_{\alpha_{k+1}}\left(F'[\hat{a}_k]^* F'[\hat{a}_k]\right) F'[\hat{a}_k]^* \Big(u^{\mathrm{obs}} - F(\hat{a}_k) + F'[\hat{a}_k](\hat{a}_k - a_0) \Big) \tag{1.3a}$$

with $g_\alpha(\lambda) := \frac{(\lambda+\alpha)^m - \alpha^m}{\lambda(\lambda+\alpha)^m}$ corresponding to m-times iterated Tikhonov regularization with $m \in \mathbb{N}$. Standard Tikhonov regularization is included as the special case $m = 1$. For deterministic errors (i.e., $\sigma = 0$ in (1.2a)) the convergence of the iteration (1.3a) has been studied in [1, 2, 3]. Typically the explicit formula (1.3a) is not used in implementations, but $\hat{a}_{k+1}$ is computed by solving m linear least squares problems (see Figure 2.1). The advantage of iterated Tikhonov regularization over ordinary Tikhonov regularization is a higher qualification number (see [12]). For simplicity, we assume that the regularization parameters are chosen of the form

$$\alpha_k = \alpha_0 q^k \tag{1.3b}$$

with $q \in (0, 1)$ and $\alpha_0 > 0$.

Let us comment on the differences when treating measurement errors as random instead of deterministic: From a practical point of view the most important difference is the choice of the stopping index. Whereas the discrepancy principle, as the most common deterministic stopping rule, works reasonably well for discrete random noise models with small data vectors, the performance of the discrepancy principle becomes arbitrarily bad as the size of the data vector increases. This is further discussed and numerically demonstrated in section 5. The same holds true for the deterministic version of Lepskiĭ's balancing principle as studied for the iteration (1.3a) in [3]. From a theoretical point of view the rates of convergence are different for deterministic and random noise, and in the latter case they depend not only on the source condition, but also on the operator F and the covariance operator $\mathbf{Cov}_\xi$ of the noise.

Actually our analysis also provides an improvement of known results for purely deterministic errors, i.e., $\sigma = 0$. This is achieved by showing an oracle inequality, which is a well-established technique in statistics (cf. [8, 9]), but is rarely used in numerical analysis thus far. To our knowledge the only deterministic oracle inequality has been shown by Mathé and Pereverzev (see [19, 21]) for Lepskiĭ's balancing principle for linear problems. Theorem 4.1 below is a generalization of this result to nonlinear problems. As shown in Remark 4.3 this provides error estimates which are better by an arbitrarily large factor than any known error deterministic estimates for any nonlinear inversion method in the limit $\delta \to 0$.

An important alternative to the iteratively regularized Gauss–Newton method is nonlinear Tikhonov regularization, for which convergence and convergence rate results for random noise have been obtained by O'Sullivan [23], Bissantz, Hohage, and Munk [4], Loubes and Ludeńa [17], and Hohage and Pricop [14]. In this paper we show order optimal rates of convergence under less restrictive assumptions on the operator than in [14, 17, 23] and for a range of smoothness classes instead of a single one as in [4].

The paper is organized as follows: In the following section we show that the total error can be decomposed into an approximation error, a propagated data error, and a nonlinearity error, and that the last error component is dominated by the sum of the first two error components (Lemma 2.2). This will be fundamental for the rest of this paper. In section 3 we prove order optimal rates of convergence if the smoothness of the solution is known and the stopping index is chosen appropriately. Adaptation to unknown smoothness by Lepskiĭ's balancing principle is discussed in section 4, and an oracle inequality for nonlinear inverse problems is shown. The paper is completed by numerical simulations for a parameter identification problem in a differential equation which illustrate how well the theoretical rates of convergence are met and compare the performances of the balancing principle and the discrepancy principle.

2. Error decomposition.

In this section we will analyze the error

$$E_k = \hat{a}_k - a^\dagger \tag{2.1}$$

of the iteration (1.3). We set $\hat{a}_0 := a_0$, i.e., $E_0 = a_0 - a^\dagger$. Since lower bounds for the expected square error are typically not available for nonlinear inverse problems, we compare our upper bounds on the error with lower bounds for the linearized inverse problem

$$Ta = u_{\mathrm{lin}}^{\mathrm{obs}}, \qquad u_{\mathrm{lin}}^{\mathrm{obs}} = Ta^\dagger + \delta\eta + \sigma\xi \tag{2.2}$$

with the operator $T := F'[a^\dagger]$. It is a fundamental observation due to Bakushinskiĭ [1], which transfers directly from deterministic errors to random noise, that the total error

$$E_{k+1} = E_{k+1}^{\mathrm{app}} + E_{k+1}^{\mathrm{noi}} + E_{k+1}^{\mathrm{nl}} \tag{2.3a}$$

in (2.1) can be decomposed into an *approximation error* E_{k+1}^{app}, a *propagated data noise error* E_{k+1}^{noi}, and a *nonlinearity error* E_{k+1}^{nl} given by

$$\begin{aligned} E_{k+1}^{\mathrm{app}} &:= r_{\alpha_{k+1}}(T^*T)E_0, \\ E_{k+1}^{\mathrm{noi}} &:= g_{\alpha_{k+1}}(T_k^*T_k)T_k^*(\delta\eta + \sigma\xi), \\ E_{k+1}^{\mathrm{nl}} &:= g_{\alpha_{k+1}}(T_k^*T_k)T_k^*\left(F(a^\dagger) - F(\hat{a}_k) + T_kE_k\right) \\ &\quad + \left(r_{\alpha_{k+1}}(T_k^*T_k) - r_{\alpha_{k+1}}(T^*T)\right)E_0. \end{aligned} \tag{2.3b}$$

Here $T_k := F'[\hat{a}_k]$ and $r_\alpha(\lambda) = 1 - \lambda g_\alpha(\lambda) = \left(\frac{\alpha}{\alpha+\lambda}\right)^m$. If F is linear, then $T = T_k = F$, and the Taylor remainder $F(a^\dagger) - F(\hat{a}_k) + T_k E_k$ vanishes. Hence $E^{\mathrm{nl}}_{k+1} = 0$; i.e., the nonlinearity error vanishes for the linearized equation (2.2). This can also be seen as follows: If F is linear, then the iteration formula (1.3a) reduces to the nonrecursive formula $\hat{a}_k = a_0 + g_{\alpha_k}(T^*T)T^*(u^\delta_{\mathrm{lin}} - Ta_0)$, which is the underlying linear regularization method with initial guess a_0 and regularization parameter α_k applied to (2.2). The approximation error E^{app}_k agrees exactly in the linear and the nonlinear case, and the data noise error E^{noi}_k differs only by the operator T and T_k. The goal of the following analysis is to show that the nonlinearity error $\|E^{\mathrm{nl}}_k\|$ can be bounded in terms of sharp estimates of $\|E^{\mathrm{app}}_k\| + \|E^{\mathrm{noi}}_k\|$ (Lemma 2.2).

Approximation error. We will assume that there exists $w \in \mathcal{Y}$ such that

$$a_0 - a^\dagger = T^*w \qquad \text{with } \|w\| \le \rho \tag{2.4}$$

for some $\rho > 0$. This is equivalent to the existence of $\tilde{w} \in \mathcal{X}$ with $\|\tilde{w}\| \le \rho$ such that $a_0 - a^\dagger = (T^*T)^{1/2}\tilde{w}$ (see [12, Prop. 2.18]). Later we will require ρ to be sufficiently small, which expresses the usual closeness condition on the initial guess required for the convergence of Newton's method. Note, however, that we require not only smallness of $a_0 - a^\dagger$ in the norm $\|\cdot\|_{\mathcal{X}}$ but smallness in the stronger norm $\|(T^*)^\dagger \cdot \|_{\mathcal{X}}$. It is well known (see [12] or (3.3) below) that under assumption (2.4) the approximation error of iterated Tikhonov regularization is bounded by

$$\|E^{\mathrm{app}}_k\| \le C_r \rho \sqrt{\alpha_k}, \qquad k \in \mathbb{N}. \tag{2.5}$$

Moreover, the approximation error satisfies

$$\|E^{\mathrm{app}}_{k+1}\| \le \|E^{\mathrm{app}}_k\| \le \gamma_{\mathrm{app}} \|E^{\mathrm{app}}_{k+1}\|, \qquad k \in \mathbb{N}, \tag{2.6a}$$

with $\gamma_{\mathrm{app}} := q^{-m}$. If α_0 is chosen sufficiently large, we also have

$$\|E_0\| \le \gamma_{\mathrm{app}} \|E^{\mathrm{app}}_1\|. \tag{2.6b}$$

All the inequalities in (2.6) can be reduced to inequalities for real-valued functions with the help of spectral theory [12]. The second inequality in (2.6a) follows from

$$r_{\alpha_k}(t) = \left(\frac{\alpha_k}{\alpha_k + t}\right)^m \le \left(\frac{\alpha_k}{\alpha_{k+1} + t}\right)^m = \frac{1}{q^m} r_{\alpha_{k+1}}(t), \qquad t \ge 0,$$

and the first inequality holds since $r_\alpha(t) = \left(1 - \frac{t}{\alpha+t}\right)^m$ is monotonically increasing in α. Finally, (2.6b) follows from $\inf_{t\in[0,\bar{t}]} \left(\frac{\alpha_1}{q(\alpha_1+t)}\right)^m = \left(\frac{\alpha_0}{q\alpha_0+\bar{t}}\right)^m \ge 1$ for $\alpha_0 \ge \frac{\bar{t}}{(1-q)}$, where $\bar{t} := \|T^*T\|$.

Remark 2.1. Note that the second inequality in (2.6a) rules out regularization methods with infinite qualification such as Landweber iteration as an alternative to iterated Tikhonov regularization. Although the regularized Newton method (1.3a) also converges for such linear regularization methods [2, 15] and convergence is even faster for smooth solutions, the estimate (2.15) in Lemma 2.2 will be violated in general as it contains the norm of the approximation error $\|E_{\mathrm{app}}\|$ itself instead of an estimate. This estimate is crucial to achieve the improvement discussed in Remark 4.3. The results of this paper also hold true for other spectral regularization methods satisfying (2.6) and (2.12) below, but since iterated Tikhonov regularization is by far the most common choice, we have decided to restrict ourselves to this case for simplicity.

Propagated data noise error. The deterministic part of E_k^{noi} can be estimated by the well-known operator norm bound

$$\|g_{\alpha_k}(T_k^* T_k)T_k^*\| \leq \frac{C_g}{\sqrt{\alpha_k}} \tag{2.7}$$

with a constant C_g depending only on m. This bound cannot be used to estimate the stochastic part of E_k^{noi} or, more precisely, the variance term

$$V(a,\alpha) := \|g_\alpha(F'[a]^* F'[a])F'[a]^* \xi\|^2$$

with $a \in D(F)$ and $\alpha > 0$, since $\|\xi\| = \infty$ almost surely for typical noise processes ξ such as white noise. We assume that there exists a known function $\varphi_{\mathrm{noi}} : (0, \alpha_0] \to (0, \infty)$ such that

$$\left(\mathbf{E}\left[V(a,\alpha)\right]\right)^{1/2} \leq \varphi_{\mathrm{noi}}(\alpha) \qquad \text{for } \alpha \in (0,\alpha_0] \text{ and } a \in D(F). \tag{2.8a}$$

Such a condition is satisfied if $F'[a]$ is Hilbert–Schmidt for all $a \in D(F)$ and the singular values of these operators have the same rate of decay for all $a \in D(F)$. Estimates of the form (2.8a) have been derived for spectral regularization methods under general assumptions in [5]. We further assume that

$$1 < \underline{\gamma}_{\mathrm{noi}} \leq \frac{\varphi_{\mathrm{noi}}(\alpha_{k+1})}{\varphi_{\mathrm{noi}}(\alpha_k)} \leq \overline{\gamma}_{\mathrm{noi}}, \qquad k \in \mathbb{N}, \tag{2.8b}$$

for some constants $\underline{\gamma}_{\mathrm{noi}}, \overline{\gamma}_{\mathrm{noi}} < \infty$. Moreover, we assume an exponential inequality of the form

$$\mathbb{P}\left\{V(a,\alpha) \geq \tau \mathbf{E}\left[V(a,\alpha)\right]\right\} \leq c_1 e^{-c_2 \tau} \qquad \text{for all } a \in D(F),\ \alpha \in (0,\alpha_0],\ \tau \geq 1 \tag{2.8c}$$

with constants $c_1, c_2 > 0$. Such an exponential inequality is derived for Gaussian noise processes ξ in the appendix. If ξ is white noise and the singular values of $F'[a]$ decay at the rate $\sigma_j(F'[a]) \sim j^{-\beta}$ with $\beta > \frac{1}{2}$ uniformly for all $a \in D(F)$, then we can choose φ_{noi} of the form

$$\varphi_{\mathrm{noi}}(\alpha) = C_{\mathrm{noi}} \alpha^{-c}, \tag{2.9}$$

with $c := \frac{1}{2} + \frac{1}{4\beta}$ and some constant $C_{\mathrm{noi}} \in (0, \infty)$ (see [5]). For colored noise c may also have values smaller than $\frac{1}{2}$.

By virtue of the exponential inequality (2.8c) the probability is very small that the stochastic propagated data noise error $V(\widehat{a}_k, \alpha_k)$ at any Newton step k is much larger than the expected value $\mathbf{E}\left[V(\widehat{a}_k, \alpha_k)\right]$. We will distinguish between a "good case" where the propagated data noise is "small" at all Newton steps, and a "bad case" where the noise is "large" in at least one Newton step. The "good case" is analyzed in Lemma 2.2 below. The proof of Theorem 3.2 will require a rather elaborate distinction between what is small and what is large in order to derive the optimal rate of convergence. This distinction will be described by functions $\tau(k,\sigma)$ satisfying

$$\tau(k+1,\sigma)\varphi_{\mathrm{noi}}(\alpha_{k+1})^2 \geq \tau(k,\sigma)\varphi_{\mathrm{noi}}(\alpha_k)^2, \qquad \sigma > 0,\ k = 1, 2, \ldots, \overline{k} - 1. \tag{2.10a}$$

For given $\tau = \tau(k, \sigma)$ and a given maximal iteration number $\overline{k}$, the "good event" $A_{\tau,\overline{k}}$ is that all iterates $\hat{a}_k$ belong to $D(F)$ for $k = 1, \ldots, \overline{k}$ (and hence are well defined) and that the propagated data noise error is bounded by

$$\|E_k^{\mathrm{noi}}\| \leq \Phi_{\mathrm{noi}}(k) \qquad \text{for } k = 1, \ldots, \overline{k} \text{ with}$$

$$\Phi_{\mathrm{noi}}(k) := \sqrt{\tau(k,\sigma)}\sigma\varphi_{\mathrm{noi}}(\alpha_k) + \delta \frac{C_g}{\sqrt{\alpha_k}}. \tag{2.10b}$$

Nonlinearity error. In the following we will assume that the Fréchet derivative F' of F satisfies a Lipschitz condition with constant $L > 0$; i.e.,

$$\|F'[a_1] - F'[a_2]\| \leq L\|a_1 - a_2\| \qquad \text{for all } a_1, a_2 \in D(F). \tag{2.11}$$

If the straight line connecting $a^\dagger$ and $\hat{a}_k$ is contained in $D(F)$, the Taylor remainder in (2.3) is bounded as $\|F(a^\dagger) - F(\hat{a}_k) + T_kE_k\| \leq (L/2)\|E_k\|^2$. Now it follows from (2.7) that the first term in the definition of E_{k+1}^{nl} is bounded by $C_gL/(2\sqrt{\alpha_{k+1}})\|E_k\|^2$.

To bound the second term in the definition of E_{k+1}^{nl}, we need the assumption (2.4) and the estimate

$$\|\left(r_{\alpha_k}(T_k^*T_k) - r_{\alpha_k}(T^*T)\right)(T^*T)^{1/2}\| \leq C_{\mathrm{nl}}\|T - T_k\|, \tag{2.12}$$

which can be shown with the help of the Riesz–Dunford formula; see Bakushinskiĭ and Kokurin [2, section 4.1]. Using (2.12), the source condition (2.4), and (2.11), the second term can be bounded by $\widetilde{C}_{\mathrm{nl}}\rho\|E_k\|$ with a constant $\widetilde{C}_{\mathrm{nl}} > 0$.

In summary we obtain the following recursive estimate of the nonlinearity error:

$$\|E_{k+1}^{\mathrm{nl}}\| \leq \frac{LC_g}{2\sqrt{\alpha_{k+1}}}\|E_k\|^2 + \widetilde{C}_{\mathrm{nl}}\rho\|E_k\|. \tag{2.13}$$

LEMMA 2.2. *Assume that the ball $B_{2R}(a_0)$ with center a_0 and radius $2R > 0$ is contained in $D(F)$ and that* (1.2), (1.3), (2.8), *and* (2.11) *hold true. Moreover, let α_0 be sufficiently large that* (2.6b) *is satisfied.*

Then there exists $\rho > 0$ specified in the proof such that the following holds true for all $a^\dagger \in B_R(a_0)$ satisfying the source condition (2.4) *and all $\delta, \sigma \geq 0$: If* (2.10) *defining the "good event" is satisfied with $\overline{k} = K_{\max}$ defined by*

$$K_{\max} := \max\left\{k \in \mathbb{N} : \Phi_{\mathrm{noi}}(k)\alpha_k^{-1/2} \leq C_{\mathrm{stop}}\right\} \text{ and} \tag{2.14}$$

$$0 < C_{\mathrm{stop}} \leq \min\left(\frac{1}{8LC_g}, \frac{R}{4\sqrt{\alpha_0}}\right),$$

then $\hat{a}_k \in B_R(a^\dagger)$ and

$$\|E_k^{\mathrm{nl}}\| \leq \gamma_{\mathrm{nl}}\left(\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)\right) \qquad \text{for all } k = 1, \ldots, K_{\max}. \tag{2.15}$$

Here $\gamma_{\mathrm{nl}} := 8LC_gC_{\mathrm{stop}}$ satisfies $\gamma_{\mathrm{nl}} \leq 1$.

Proof. We prove this by induction in k, starting with the induction step. If the assertion holds true for $k-1$ ($k = 2, \ldots, K_{\max}$), then we obtain from (2.3a), (2.6a), and (2.10) that

$$\begin{aligned} \|E_{k-1}\| &\leq (1+\gamma_{\mathrm{nl}})\left(\|E_{k-1}^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k-1)\right) \\ &\leq (1+\gamma_{\mathrm{nl}})\left(\gamma_{\mathrm{app}}\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)\right). \end{aligned} \tag{2.16}$$

Hence, it follows from (2.13) and the inequality $(x+y)^2 \leq 2x^2 + 2y^2$ that

$$\begin{aligned} \|E_k^{\mathrm{nl}}\| \leq{}& \widetilde{C}_{\mathrm{nl}}\rho(1+\gamma_{\mathrm{nl}})\left(\gamma_{\mathrm{app}}\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)\right) \\ &+ \frac{LC_g}{\sqrt{\alpha_k}}(1+\gamma_{\mathrm{nl}})^2\left(\gamma_{\mathrm{app}}^2\|E_k^{\mathrm{app}}\|^2 + \Phi_{\mathrm{noi}}^2(k)\right). \end{aligned} \tag{2.17}$$

```
Input: $u^{\mathrm{obs}}, \delta, \sigma, F, a_0, R, K_{\max}$
$k := 0$; $\hat{a}_0 := a_0$
while $k < K_{\max}$ and $\|\hat{a}_k - a_0\| \leq 2R$
  $\hat{a}_{k+1}^{(0)} := a_0$
  for $j = 1, \dots, m$
    $\hat{a}_{k+1}^{(j)} := \mathrm{argmin}_{a \in \mathcal{X}} \left\{ \|F'[\hat{a}_k](a - \hat{a}_k) + F(\hat{a}_k) - u^{\mathrm{obs}}\|_{\mathcal{Y}}^2 + \alpha_{k+1} \|a - \hat{a}_{k+1}^{(j-1)}\|_{\mathcal{X}}^2 \right\}$
  $\hat{a}_{k+1} := \hat{a}_{k+1}^{(m)}$; $k := k + 1$
end
if $(\|\hat{a}_k - a_0\| > 2R)$      $K_* := 0$
else      choose $K_* \in \{0, 1, \dots, K_{\max}\}$ by (3.4), (3.7), (4.3), respectively
Output: $\hat{a}_{K_*}$
```

FIG. 2.1. *Algorithm: Iteratively regularized Gauss–Newton method with m-times iterated Tikhonov regularization for random noise.*

If $\rho \leq \gamma_{\mathrm{nl}}/(2\widetilde{C}_{\mathrm{nl}}(1+\gamma_{\mathrm{nl}})\gamma_{\mathrm{app}})$, the first term on the right-hand side of (2.17) is bounded by $(\gamma_{\mathrm{nl}}/2)\left(\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)\right)$. Now we estimate the second term in (2.17). Using (2.5) we obtain

$$\frac{\|E_k^{\mathrm{app}}\|}{\sqrt{\alpha_k}} \leq C_r \rho \leq \frac{\gamma_{\mathrm{nl}}}{2LC_g(1+\gamma_{\mathrm{nl}})^2 \gamma_{\mathrm{app}}^2}$$

for ρ sufficiently small, so $\frac{LC_g}{\sqrt{\alpha_k}}(1+\gamma_{\mathrm{nl}})^2 \gamma_{\mathrm{app}}^2 \|E_k^{\mathrm{app}}\|^2 \leq (\gamma_{\mathrm{nl}}/2)\|E_k^{\mathrm{app}}\|$. Moreover, it follows from the inequality $\gamma_{\mathrm{nl}} \leq 1$ and the definition of $K_{\max}$ that

$$\frac{LC_g}{\sqrt{\alpha_k}}(1+\gamma_{\mathrm{nl}})^2 \Phi_{\mathrm{noi}}^2(k) \leq \frac{4LC_g}{\sqrt{\alpha_k}} \Phi_{\mathrm{noi}}^2(k) \leq 4LC_g C_{\mathrm{stop}} \Phi_{\mathrm{noi}}(k) \leq \frac{\gamma_{\mathrm{nl}}}{2} \Phi_{\mathrm{noi}}(k).$$

Therefore, the second line in (2.17) is also bounded by $(\gamma_{\mathrm{nl}}/2)\left(\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)\right)$, which yields (2.15) for k. This can be used to show that $\hat{a}_k \in D(F)$: If ρ is sufficiently small, then $\|E_k^{\mathrm{app}}\| \leq C_r \rho \sqrt{\alpha_0} \leq R/(2+2\gamma_{\mathrm{nl}})$. Moreover, it follows from the monotonicity of Φ_{noi} (cf. (2.10)), (2.14), and $\gamma_{\mathrm{nl}} \leq \min(1, LC_g \alpha_0^{-1/2}/2)$ that

$$\Phi_{\mathrm{noi}}(k) \leq \Phi_{\mathrm{noi}}(K_{\max}) \leq C_{\mathrm{stop}} \alpha_{K_{\max}}^{1/2} \leq C_{\mathrm{stop}} \alpha_0^{1/2} \leq \frac{R}{4} \leq \frac{R}{2+2\gamma_{\mathrm{nl}}}.$$

Together with (2.15) this shows that $\|E_k\| \leq (1+\gamma_{\mathrm{nl}})(\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)) \leq R$; i.e., $\hat{a}_k \in B_R(a^\dagger) \subset D(F)$.

It remains to establish the assertion for $k = 1$. Due to assumption (2.6b), $\|E_0\|$ is bounded by the right-hand side of (2.16) with $k = 1$. Now the assertion for $k = 1$ follows as above. □

Since we do not assume the stochastic noise ξ to be bounded, there is a positive probability that $\hat{a}_k \notin D(F)$ at each Newton step k. Therefore, we stop the iteration if $\|\hat{a}_k - a_0\| \geq 2R$ for some k, and we choose the initial guess a_0 as estimator of $a^\dagger$ in this case. The algorithm is summarized in Figure 2.1.

3. Convergence results for known smoothness. In the following we will assume more smoothness for $a_0 - a^\dagger$ than in (2.4). This is expressed in terms of source conditions of the form

$$a_0 - a^\dagger = \Lambda(T^*T)\tilde{w},$$

where Λ is continuous and monotonely increasing with $\Lambda(0) = 0$ (see [20] for the linear case). If

$$\sup_{t\in[0,\bar{t}]} \Lambda(t)|r_\alpha(t)| \leq C_\Lambda \Lambda(\alpha), \qquad \alpha \in [0, \alpha_0], \tag{3.1}$$

then $\|E_k^{\text{app}}\| \leq C_\Lambda \Lambda(\alpha_k)\|\tilde{w}\|$. The most important case is $\Lambda(t) = t^\mu$, and the condition

$$a_0 - a^\dagger = (T^*T)^\mu \tilde{w} \qquad \text{with } \|\tilde{w}\| \leq \tilde{\rho} \tag{3.2}$$

is called a Hölder source condition. The largest number $\mu > 0$ for which (3.1) holds true with this choice of Λ is called the *qualification* μ_0 of the linear regularization method. The qualification of Tikhonov regularization is $\mu_0 = 1$, and the qualification of m-times iterated Tikhonov regularization is $\mu_0 = m$ (see [12]). We obtain

$$\|E_k^{\text{app}}\| \leq C_\mu \alpha_k^\mu \tilde{\rho} \qquad \text{for } 0 \leq \mu \leq \mu_0. \tag{3.3}$$

Let us first consider deterministic errors, i.e., $\sigma = 0$. The following result shows that the same rate of convergence can be achieved for the nonlinear inverse problem (1.2) as for the linearized problem (2.2).

THEOREM 3.1 (deterministic errors). *Let the assumptions of Lemma* 2.2 *hold true with* $\sigma = 0$, *and let*

$$K_* := \min\{K_{\max}, K\}, \qquad K := \operatorname*{argmin}_{k\in\mathbb{N}} \left(\|E_k^{\text{app}}\| + \delta \frac{C_g}{\sqrt{\alpha_k}}\right). \tag{3.4}$$

Then there exist constants $C, \delta_0 > 0$ *such that*

$$\|\hat{a}_{K_*} - a^\dagger\| \leq C \inf_{k\in\mathbb{N}} \left(\|E_k^{\text{app}}\| + \delta \frac{C_g}{\sqrt{\alpha_k}}\right) \qquad \textit{for all } \delta \in (0, \delta_0]. \tag{3.5}$$

(Note that $E_k^{\text{app}} = r_{\alpha_k}(T^*T)(a_0 - a^\dagger)$ *is well defined for all* $k \in \mathbb{N}$ *even if* $\hat{a}_k$ *is not well defined for all* k.*)*

Proof. If $K \leq K_{\max}$, then (3.5) follows with $\Phi_{\text{noi}}(k) = \delta C_g/\sqrt{\alpha_k}$ and $C = 1 + \gamma_{\text{nl}}$ from Lemma 2.2 and the error decomposition (2.3). On the other hand, using (2.5) and (2.14) we get

$$\|E_k^{\text{app}}\| + \Phi_{\text{noi}}(k) \geq \Phi_{\text{noi}}(k) > C_{\text{stop}}\sqrt{\alpha_{K_{\max}}} \geq \frac{1}{1 + C_r\rho/C_{\text{stop}}} \left(\|E_{K_{\max}}^{\text{app}}\| + \Phi_{\text{noi}}(K_{\max})\right)$$

for $k > K_{\max}$. Therefore, (3.5) holds true with $C = 1 + C_r\rho/C_{\text{stop}}$ if $K \geq K_{\max}$ and hence $K_* = K_{\max}$. □

In particular, for Hölder source conditions (3.2) with $\mu \in [\frac{1}{2}, \mu_0]$, plugging (3.3) into (3.5), we obtain the rate

$$\|\hat{a}_{K_*} - a^\dagger\| = O\left(\tilde{\rho}^{\frac{1}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}}\right), \qquad \delta \to 0, \tag{3.6}$$

first shown by Bakushinskiĭ [1] for $\mu = 1$ and by Blaschke, Neubauer, and Scherzer [6] for $\mu \in [1/2, 1]$. Explicit rates for other source conditions follow analogously.

We now turn to the general noise model (1.2).

THEOREM 3.2 (general noise model). *Consider an iterative regularization method described by* (1.3) *and a noise model* (1.2), (2.8). *Assume that* $B_{2R}(a_0) \subset D(F)$, *that* F *satisfies* (2.11), *and that a Hölder source condition* (3.2) *with* $\mu \in (1/2, \mu_0]$ *and* ρ *sufficiently small holds true. Define*

$$K := \operatorname*{argmin}_{k\in\mathbb{N}} \left(\|E_k^{\mathrm{app}}\| + \delta \frac{C_g}{\sqrt{\alpha_k}} + \sigma\varphi_{\mathrm{noi}}(\alpha_k) \right). \tag{3.7}$$

If $\|\hat{a}_k - a_0\| \leq 2R$ *for* $k = 1, \ldots, K$, *set* $K_* := K$; *otherwise* $K_* := 0$. *Then the expected square error of the regularized Newton iteration converges at the same rate as the expected square error of the corresponding linear regularization method; i.e., there exist constants* $C > 1$ *and* $\delta_0, \sigma_0 > 0$ *such that*

$$\left(\mathbf{E}\left[\|\hat{a}_{K_*} - a^\dagger\|^2\right]\right)^{1/2} \leq C \min_{k\in\mathbb{N}} \left(\|E_k^{\mathrm{app}}\| + \delta \frac{C_g}{\sqrt{\alpha_k}} + \sigma\varphi_{\mathrm{noi}}(\alpha_k) \right) \tag{3.8}$$

for all $\delta \in (0, \delta_0]$ *and* $\sigma \in (0, \sigma_0]$

Proof. We define the "bounded noise" events $A_1 \subset A_2 \subset \cdots \subset A_{J(\sigma)}$ with $J(\sigma) := \lfloor \ln(\sigma^{-2})/c_2 \rfloor$ and c_2 from (2.8c) by

$$A_j := \left\{ \hat{a}_k \in B_{2R}(a_0) \text{ and } \|E_k^{\mathrm{noi}}\| \leq \frac{\delta C_g}{\sqrt{\alpha_k}} + \sqrt{\tau_j(k,\sigma)}\sigma\varphi_{\mathrm{noi}}(\alpha_k), \quad k = 1, \ldots, K \right\},$$
$$\text{with } \tau_j(k,\sigma) := j + \frac{\ln\kappa}{c_2}(K - k) \text{ and } \kappa > 1 \tag{3.9}$$

(cf. (2.10b)). Due to (2.8b) we can choose $\kappa > 1$ sufficiently small such that τ_j satisfies condition (2.10a) for all $j = 1, \ldots, J$. Definition (3.9) is motivated by the fact that an unusually large propagated data noise term E_k^{noi} at the first Newton steps, where the total error is dominated by E_k^{app}, has less effect than an unusually large propagated data noise error at the last Newton steps. From Lemmas 2.2 and 3.4 it follows that the iterates $\widehat{a}_1, \ldots, \widehat{a}_K$ remain in $B_R(a^\dagger)$ if the bounds on $\|E_k^{\mathrm{noi}}\|$, $k = 1, \ldots, K$, in the definition (3.9) of A_j with $j \leq J(\sigma)$ are satisfied and σ is sufficiently small. Together with (2.8c) this yields for the probability of the complementary event

$$\begin{aligned} \mathbb{P}(\mathcal{C}A_j) &\leq c_1 \sum_{k=1}^{K} \exp(-c_2\tau_j(k,\sigma)) = c_1 \exp(-c_2 j) \sum_{k=1}^{K} \kappa^{k-K} \\ &\leq c_1 \exp(-c_2 j) \sum_{m=0}^{\infty} \kappa^{-m} = \frac{c_1 \exp(-c_2 j)}{1 - \kappa^{-1}}. \end{aligned}$$

By definition, we have $K_* = K$ for the events A_j. Applying Lemma 2.2 and the inequality $(x + y + z)^2 \leq 3x^2 + 3y^2 + 3z^2$ and using $\gamma_{\mathrm{nl}} \leq 1$ we obtain

$$\|\hat{a}_{K_*} - a^\dagger\|^2 \leq 6\|E_K^{\mathrm{app}}\|^2 + 6\delta^2 \frac{C_g^2}{\alpha_K} + 6\tau_j(K,\sigma)\sigma^2\varphi_{\mathrm{noi}}(\alpha_K)^2 =: B_j \quad \text{for the event } A_j.$$

Moreover, by the definition of the algorithm we have an error bound $\|\hat{a}_{K_*} - a^\dagger\| \leq \|\hat{a}_{K_*} - a_0\| + \|a_0 - a^\dagger\| \leq 3R$ for the event $\mathcal{C}A_{J(\sigma)}$. Note that $J(\sigma)$ is chosen such that $e^{-c_2 J(\sigma)} \leq \sigma^2$. Hence,

$$\begin{aligned}
\mathbf{E}\left[\|\hat{a}_{K_*} - a^\dagger\|^2\right] &\leq \mathbb{P}(A_1)B_1 + \sum_{j=2}^{J(\sigma)} \mathbb{P}(A_j \setminus A_{j-1})B_j + \mathbb{P}(\mathcal{C}A_{J(\sigma)})9R^2 \\
&\leq 6\|E_K^{\text{app}}\|^2 + 6\delta^2 \frac{C_g^2}{\alpha_K} + 6\left(\sigma\varphi_{\text{noi}}(\alpha_K)\right)^2 \frac{c_1}{1-\kappa^{-1}} \sum_{j=1}^{J(\sigma)-1} j e^{-c_2 j} \\
&\quad + \frac{c_1 e^{-c_2 J(\sigma)}}{1-\kappa^{-1}} 9R^2 \\
&\leq C\left(\|E_K^{\text{app}}\| + \delta\frac{C_g}{\sqrt{\alpha_k}} + \sigma\varphi_{\text{noi}}(\alpha_K)\right)^2
\end{aligned}$$

for some constant $C > 0$. In the second line we have used that $\mathbb{P}(A_1) + \sum_{j=2}^{J(\sigma)} \mathbb{P}(A_j \setminus A_{j-1}) + \mathbb{P}(\mathcal{C}A_{J(\sigma)}) = 1$ and $\mathbb{P}(A_j \setminus A_{j-1}) \leq \mathbb{P}(\mathcal{C}A_{j-1})$. Now (3.8) follows as in the proof of Theorem 3.1. □

The proof of Theorem 3.2 is completed by the following two lemmas.

LEMMA 3.3. *Setting* $\tilde{\Phi}_{\text{noi}}(k) := \delta\frac{C_g}{\sqrt{\alpha_k}} + \sigma\varphi_{\text{noi}}(\alpha_k)$, $\underline{\underline{\gamma}}_{\text{noi}} := \min\{\underline{\gamma}_{\text{noi}}, q^{-1/2}\}$, *and* $\overline{\overline{\gamma}}_{\text{noi}} := \max\{\overline{\gamma}_{\text{noi}}, q^{-1/2}\}$, *the optimal stopping index* K *defined in* (3.7) *satisfies the following bounds:*

$$\frac{\tilde{\Phi}_{\text{noi}}(K)}{\|E_K^{\text{app}}\|} \leq \frac{\gamma_{\text{app}} - 1}{1 - \underline{\underline{\gamma}}_{\text{noi}}^{-1}}, \tag{3.10}$$

$$K \geq \sup\left\{\underline{K} \in \mathbb{N} : \|E_1^{\text{app}}\|\gamma_{\text{app}}^{1-\underline{K}} > \inf_{l\in\mathbb{N}}\left(C_r\rho\sqrt{\alpha_l} + \tilde{\Phi}_{\text{noi}}(1)\overline{\overline{\gamma}}_{\text{noi}}^{l-1}\right)\right\}. \tag{3.11}$$

Proof. It follows from (2.8b) that

$$1 < \underline{\underline{\gamma}}_{\text{noi}} \leq \frac{\tilde{\Phi}_{\text{noi}}(k+1)}{\tilde{\Phi}_{\text{noi}}(k)} \leq \overline{\overline{\gamma}}_{\text{noi}}, \qquad k \in \mathbb{N}. \tag{3.12}$$

To show (3.10), assume on the contrary that

$$(\gamma_{\text{app}} - 1)\|E_K^{\text{app}}\| < (1 - \underline{\underline{\gamma}}_{\text{noi}}^{-1})\tilde{\Phi}_{\text{noi}}(K). \tag{3.13}$$

Then we obtain using (2.6a) and (3.12)

$$\|E_{K-1}^{\text{app}}\| + \tilde{\Phi}_{\text{noi}}(K-1) \leq \gamma_{\text{app}}\|E_K^{\text{app}}\| + \underline{\underline{\gamma}}_{\text{noi}}^{-1}\tilde{\Phi}_{\text{noi}}(K) < \|E_K^{\text{app}}\| + \tilde{\Phi}_{\text{noi}}(K)$$

in contradiction to the definition of K. Therefore, (3.13) is false, and (3.10) holds true.

To show (3.11), assume that

$$C_r\rho\sqrt{\alpha_l} + \tilde{\Phi}_{\text{noi}}(1)\overline{\overline{\gamma}}_{\text{noi}}^{l-1} < \|E_1^{\text{app}}\|\gamma_{\text{app}}^{1-\underline{K}} \tag{3.14}$$

for some $\underline{K}, l \geq 1$. Then for all $k \leq \underline{K}$ we have

$$\begin{aligned}\|E_l^{\rm app}\| + \tilde{\Phi}_{\rm noi}(l) &\leq C_r \rho \sqrt{\alpha_l} + \tilde{\Phi}_{\rm noi}(1)\overline{\overline{\gamma}}_{\rm noi}^{\,l-1} \\ &< \|E_1^{\rm app}\|\gamma_{\rm app}^{1-\underline{K}} \leq \|E_{\underline{K}}^{\rm app}\| \leq \|E_k^{\rm app}\| \\ &\leq \|E_k^{\rm app}\| + \tilde{\Phi}_{\rm noi}(k)\end{aligned}$$

using (2.5), (2.6a), and (3.12). Therefore, it follows from the definition of K that $K > \underline{K}$. Taking the infimum over l in (3.14) and then the supremum over $\underline{K}$ yields (3.11). □

LEMMA 3.4. *Under the assumptions of Theorem* 3.2 *there exists* $\sigma_0 > 0$ *such that*

$$K \leq K_{\max}(\sigma, j) \qquad \textit{for all } \sigma \leq \sigma_0 \textit{ and } j = 1, \ldots, J(\sigma) := \lfloor \ln(\sigma^{-2})/c_2 \rfloor \tag{3.15}$$

with $K_{\max}(\sigma, j)$ *defined as in Lemma* 2.2 *using* τ_j *in* (3.9):

$$K_{\max}(\sigma, j) := \max\left\{k \in \mathbb{N} : \left(\delta \frac{C_g}{\sqrt{\alpha_k}} + \tau_j(k, \sigma)\sigma\varphi_{\rm noi}(\alpha_k)\right)\alpha_k^{-1/2} \leq C_{\rm stop}\right\}.$$

Proof. It follows from (2.10a), the inequalities $\tau_j(K, \sigma) \leq \tau_{J(\sigma)}(K, \sigma) \leq \ln(\sigma^{-2})/c_2$, and (3.10) that for all $j \leq J(\sigma)$ and $k \leq K$

$$\begin{aligned}\left(\delta \frac{C_g}{\sqrt{\alpha_k}} + \tau_j(k, \sigma)\sigma\varphi_{\rm noi}(\alpha_k)\right)\alpha_k^{-1/2} &\leq \left(\delta \frac{C_g}{\sqrt{\alpha_K}} + \tau_{J(\sigma)}(K, \sigma)\sigma\varphi_{\rm noi}(\alpha_K)\right)\alpha_K^{-1/2} \\ &\leq \frac{\ln(\sigma^{-2})}{c_2}\tilde{\Phi}_{\rm noi}(K)\alpha_K^{-1/2} \\ &\leq \frac{1}{c_2}\frac{\gamma_{\rm app} - 1}{1 - \underline{\gamma}_{\rm noi}^{-1}}\|E_K^{\rm app}\| \ln(\sigma^{-2})\alpha_K^{-1/2} \qquad (3.16) \\ &\leq C \ln(\sigma^{-2})\alpha_K^{\mu - 1/2}\end{aligned}$$

with $C := \frac{\rho C_\mu}{c_2}\frac{\gamma_{\rm app}-1}{1-\underline{\gamma}_{\rm noi}^{-1}}$. Furthermore, a straightforward calculation shows that the infimum in (3.11) decays at a polynomial rate in σ, so $K \geq -\kappa \ln(\sigma)$ for some $\kappa > 0$. As $\lim_{x\to\infty} xq^{cx} = 0$ for all $c > 0$, there exists $\sigma_0 > 0$ such that

$$C \ln(\sigma^{-2})\alpha_K^{\mu-1/2} \leq C\left(\frac{\alpha_0}{q}\right)^{\mu-1/2} \ln(\sigma^{-2})q^{\ln(\sigma^{-2})\cdot(\kappa/2)\cdot(\mu-1/2)} \leq C_{\rm stop}$$

for all $\sigma \in (0, \sigma_0]$. This together with (3.16) gives the assertion. □

The right-hand side of the estimate (3.8) is known to be an order optimal error bound for the linearized problem (2.2) under mild assumptions (cf. [5]). Again, more explicit bounds can easily be derived under more specific assumptions. For example, if $\delta = 0$ and $\varphi_{\rm noi}$ is given by (2.9), we obtain

$$\left(\mathbf{E}\left[\|\hat{a}_{K_*} - a^\dagger\|^2\right]\right)^{1/2} = O\left(\tilde{\rho}^{\frac{c}{\mu+c}}\sigma^{\frac{\mu}{\mu+c}}\right), \qquad \sigma \to 0. \tag{3.17}$$

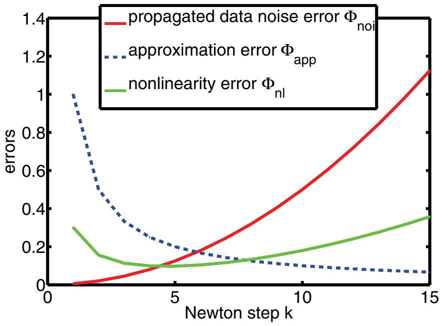

FIG. 4.1. *Illustration of setting in section* 4.

4. Adaptation by Lepskiĭ's balancing principle.

The stopping index in Theorem 3.2 cannot be computed since it depends on $\|E_k^{\mathrm{app}}\|$ and hence the smoothness of the difference $a_0 - a^\dagger$ of the initial guess and the unknown solution. In this section we address the problem of how to choose the stopping index of the Newton iteration adaptively using a Lepskiĭ-type balancing principle.

We will present the balancing principle in a general framework adapted from [19]. From our choice of notation it will be obvious how the regularized Newton methods (1.3) fit into this framework and how the stopping index can be selected for these methods. As the same method will hopefully also apply to regularization methods other than (1.3), we have decided to use this more general setting.

General abstract setting. Let $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_{K_{\max}}$ be estimators of $a^\dagger$ in a metric space (X, d), and let $\Phi_{\mathrm{noi}}, \Phi_{\mathrm{app}}, \Phi_{\mathrm{nl}} : \mathbb{N}_0 \to [0, \infty)$ be functions such that

$$d(\hat{a}_k, a^\dagger) \le \Phi_{\mathrm{noi}}(k) + \Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{nl}}(k), \qquad k \le K_{\max}. \tag{4.1}$$

We assume that Φ_{noi} is known and nondecreasing, Φ_{app} is unknown and nonincreasing, and Φ_{nl} is unknown and satisfies

$$\Phi_{\mathrm{nl}}(k) \le \gamma_{\mathrm{nl}} \left(\Phi_{\mathrm{noi}}(k) + \Phi_{\mathrm{app}}(k) \right), \qquad k = 0, \dots, K_{\max}, \tag{4.2}$$

for some $\gamma_{\mathrm{nl}} > 0$. This is illustrated in Figure 4.1. Note that under the assumptions of Lemma 2.2 these inequalities are satisfied for the iterates of the generalized Gauss–Newton method (1.3) with $\Phi_{\mathrm{app}}(k) := \|E_k^{\mathrm{app}}\|$ and $\Phi_{\mathrm{nl}}(k) := \|E_k^{\mathrm{nl}}\|$.

As in [3] we consider the following Lepskiĭ-type parameter selection rule:

$$k_{\mathrm{bal}} := \min \left\{ k \le K_{\max} : d(\hat{a}_k, \hat{a}_m) \le 4(1 + \gamma_{\mathrm{nl}}) \Phi_{\mathrm{noi}}(m), m = k+1, \dots, K_{\max} \right\}. \tag{4.3}$$

Deterministic errors. We first recall some results from [19, 21] for linear problems, i.e., $\Phi_{\mathrm{nl}} \equiv 0$. Assume that Φ_{noi} satisfies

$$\Phi_{\mathrm{noi}}(k+1) \le \overline{\gamma}_{\mathrm{noi}} \Phi_{\mathrm{noi}}(k), \qquad k = 1, 2, \dots, K_{\max} - 1, \tag{4.4}$$

for some constant $\overline{\gamma}_{\mathrm{noi}} < \infty$, and let

$$\overline{k} := \min\{k \le K_{\max} : \Phi_{\mathrm{app}}(k) \le \Phi_{\mathrm{noi}}(k)\}. \tag{4.5}$$

Then, as shown by Mathé and Pereverzev [21], the following *deterministic oracle inequality* holds true:

$$d(\hat{a}_{k_{\mathrm{bal}}}, a^\dagger) \le 6\Phi_{\mathrm{noi}}(\overline{k}) \le 6\overline{\gamma}_{\mathrm{noi}} \min_{k=1,\dots,K_{\max}} \left(\Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{noi}}(k) \right). \tag{4.6}$$

This shows that k_{bal} yields an optimal error bound up to a factor $6\overline{\gamma}_{\mathrm{noi}}$.

If (4.1) and (4.2) hold true, then Assumption 2.1 in [3] is satisfied with $\mathcal{E}(k)\delta = 2(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(k)$ (in the notation of [3]) and

$$K := \min\{k \leq K_{\max} : \Phi_{\mathrm{noi}}(k) + \Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{nl}}(k) \leq 2(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(k)\}. \tag{4.7}$$

Therefore, Theorem 2.3 in [3] implies the error estimate

$$d(\hat{a}_{k_{\mathrm{bal}}}, a^\dagger) \leq 6(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(\overline{k}). \tag{4.8}$$

We obtain the following order optimality result inspired by Mathé [19].

THEOREM 4.1 (deterministic oracle inequality). *Assume* (4.1)–(4.4). *Then*

$$d(\hat{a}_{k_{\mathrm{bal}}}, a^\dagger) \leq 6(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(\overline{k}) \leq 6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}} \min_{k\in\{1,\dots,K_{\max}\}} \left(\Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{noi}}(k)\right). \tag{4.9}$$

Proof. As $\Phi_{\mathrm{app}}(k) \leq \Phi_{\mathrm{noi}}(k)$ implies

$$\Phi_{\mathrm{noi}}(k) + \Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{nl}}(k) \leq (1+\gamma_{\mathrm{nl}})\left(\Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{noi}}(k)\right) \leq 2(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(k),$$

we have $K \leq \overline{k}$ with $\overline{k}$ defined in (4.5). Therefore, we get

$$\begin{aligned} d(\hat{a}_{k_{\mathrm{bal}}}, a^\dagger) &\overset{(4.8)}{\leq} 6(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(K) \overset{K\leq\overline{k}}{\leq} 6(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(\overline{k}) \\ &\overset{(4.6)}{\leq} 6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}} \min\left\{\Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{noi}}(k) : k = 1,\dots,K_{\max}\right\}. \qquad \square \end{aligned}$$

Obviously, we could add the term $\Phi_{\mathrm{nl}}(k)$ to $\Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{noi}}(k)$ on the right-hand side of (4.9). Hence, Theorem 4.1 implies that the Lepskiĭ rule (4.3) leads to an optimal error bound up to a factor $6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}}$ among all $k = 1,\dots,K_{\max}$. However, Theorem 4.1 even implies the stronger result that we obtain the same rates of convergence as in the linear case, which are often known to be minimax.

Setting $\Phi_{\mathrm{app}}(k) := \|E_k^{\mathrm{app}}\|$, $\Phi_{\mathrm{noi}}(k) = \delta C_g/\sqrt{\alpha_k}$, and $\Phi_{\mathrm{nl}}(k) := \|E_k^{\mathrm{nl}}\|$ yields the following oracle inequality for the Gauss–Newton iteration, where we have replaced $\{1,\dots,K_{\max}\}$ by $\mathbb{N}$ (see Theorem 3.1).

COROLLARY 4.2. *Let the assumptions of Lemma* 2.2 *hold true for* $\sigma = 0$, *i.e.,* $\Phi_{\mathrm{noi}}(k) = \frac{\delta C_g}{\sqrt{\alpha_k}}$. *Furthermore let* k_{bal} *be chosen as in* (4.3). *Then*

$$\|\hat{a}_{k_{\mathrm{bal}}} - a^\dagger\| \leq 6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}} \min_{k\in\mathbb{N}} \left(\|E_k^{\mathrm{app}}\| + \frac{C_g\delta}{\sqrt{\alpha_k}}\right). \tag{4.10}$$

Remark 4.3. If $a^\dagger$ belongs to the smoothness class $\mathcal{M}_{\mu,\rho} := \{a_0 + (T^*T)^\mu \tilde{w} : \|\tilde{w}\| \leq \tilde{\rho}\}$ defined by the source condition (3.2) with $\mu \in [1/2, \mu_0]$, then it follows from (4.10) and (3.3) that

$$\|\hat{a}_{k_{\mathrm{bal}}} - a^\dagger\| \leq 6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}} \min_{k\in\mathbb{N}} \left(C_\mu \alpha_k^\mu \tilde{\rho} + \frac{C_g\delta}{\sqrt{\alpha_k}}\right) = O\left(\tilde{\rho}^{\frac{1}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}}\right), \qquad \delta \to 0, \tag{4.11}$$

and for linear problems it is well known that among all possible methods this is the best possible uniform estimate over $\mathcal{M}_{\mu,\rho}$ up to a constant (see [12]). Assume now that $\mu \in [1/2, \mu_0)$. Then $\lim_{\alpha\to 0}(\lambda/\alpha)^\mu r_\alpha(\lambda) = 0$ for all $\lambda \geq 0$ and $\sup_{\lambda,\alpha} |(\lambda/\alpha)^\mu r_\alpha(\lambda)| \leq C_\mu$, and it follows from spectral theory and Lebesgue's dominated convergence theorem that $\alpha^{-\mu}\|r_\alpha(T^*T)(T^*T)^\mu\tilde{w}\| \to 0$ as $\alpha \to 0$ for all $\tilde{w} \in \mathcal{X}$; i.e.,

$$\frac{\|E_k^{\mathrm{app}}\|}{C_\mu\tilde{\rho}\alpha_k^\mu} \to 0, \qquad k \to \infty. \tag{4.12}$$

As we will show in a moment, this implies that for all $a^\dagger \in \mathcal{M}_{\mu,\rho}$

$$\frac{\min_{k\in\mathbb{N}}\left(\|E_k^{\mathrm{app}}\| + \frac{C_g\delta}{\sqrt{\alpha_k}}\right)}{\min_{k\in\mathbb{N}}\left(C_\mu\alpha_k^\mu\tilde{\rho} + \frac{C_g\delta}{\sqrt{\alpha_k}}\right)} \to 0, \qquad \delta \to 0. \tag{4.13}$$

Equation (4.13) is the deterministic analogue of what is known as superefficiency in statistics (see [7]).

To show (4.13), let $\epsilon > 0$. Using Lemma 3.3, (3.11) with $\sigma = 0$, and $\varphi_{\mathrm{noi}}(\alpha) = C_g/\sqrt{\alpha}$, we obtain that $K(\delta) := \operatorname{argmin}_{k\in\mathbb{N}}\|E_k^{\mathrm{app}}\| + \frac{C_g\delta}{\sqrt{\alpha_k}}$ tends to ∞ as $\delta \to 0$. Therefore, it follows from (4.12) that there exists δ_0 such that $\|E_{K(\delta)}^{\mathrm{app}}\| \le \epsilon C_\mu\rho\alpha_{K(\delta)}^\mu$ for all $\delta < \delta_0$. Using (3.10) in Lemma 3.3 we obtain $C_g\delta/\sqrt{\alpha_{K(\delta)}} \le C\|E_{K(\delta)}^{\mathrm{app}}\| \le C\epsilon C_\mu\tilde{\rho}\alpha_{K(\delta)}^\mu$ with $C := \frac{\gamma_{\mathrm{app}}-1}{1-\sqrt{q}}$. Now a straightforward computation shows that

$$\begin{aligned}\inf_{\alpha>0}\left(C_\mu\alpha^\mu\tilde{\rho} + \frac{C_g\delta}{\sqrt{\alpha}}\right) &\ge \tilde{C}(C\epsilon)^{2\mu/(2\mu+1)}C_\mu\tilde{\rho}\alpha_{K(\delta)}^\mu\\ &\ge \tilde{C}(C\epsilon)^{-1/(2\mu+1)}\frac{\|E_{K(\delta)}^{\mathrm{app}}\| + C\,C_g\delta/\sqrt{\alpha_{K(\delta)}}}{1+C}\end{aligned}$$

with $\tilde{C} > 0$ independent of ϵ and δ. This shows (4.13) since we can make $(C\epsilon)^{1/(2\mu+1)}$ arbitrarily small.

Equation (4.13) implies that although estimates of the form (4.11), known in deterministic regularization theory for the discrepancy principle and improved parameter choice rules [12, 16], are order optimal as *uniform* estimates over a smoothness class, they are suboptimal by an arbitrarily large factor *for each individual element* of the smoothness class in the limit $\delta \to 0$. To our knowledge it is an open question whether or not deterministic parameter choice rules other than Lepskiĭ's balancing principle are optimal in the more restrictive sense of oracle inequalities.

General noise model. We return to the general noise model (1.2).

THEOREM 4.4. *Let the assumptions of Lemma 2.2 hold true with $K_{\max}$ determined by (2.14) with $\Phi_{\mathrm{noi}}(k) := \delta\frac{C_g}{\sqrt{\alpha_k}} + \frac{\ln\sigma^{-2}}{c_2}\sigma\varphi_{\mathrm{noi}}(\alpha_k)$, and assume (3.2) with $\mu > \frac{1}{2}$. Let $K_* := k_{\mathrm{bal}}$ as in (4.3) if $\hat{a}_k \in B_{2R}(a_0)$ for $k = 1, \ldots, K_{\max}$ and $K_* := 0$ else. Then there exist constants $C, \delta_0, \sigma_0 > 0$ such that*

$$\left(\mathbf{E}\left[\|\hat{a}_{K_*} - a^\dagger\|^2\right]\right)^{1/2} \le C\min_{k\in\mathbb{N}}\left(\|E_k^{\mathrm{app}}\| + \delta\frac{C_g}{\sqrt{\alpha_k}} + (\ln\sigma^{-1})\sigma\varphi_{\mathrm{noi}}(\alpha_k)\right) \tag{4.14}$$

for $\delta \in [0, \delta_0]$ and $\sigma \in (0, \sigma_0]$.

Proof. First consider the event $A_{\tau,K_{\max}}$ of "bounded noise" defined by (2.10b) with $\tau(k,\sigma) = (\ln\sigma^{-2})/c_2$. Then the assumptions of Theorem 4.1 are satisfied with $\Phi_{\mathrm{app}}(k) := \|E_k^{\mathrm{app}}\|$ and $\Phi_{\mathrm{nl}}(k) := \|E_k^{\mathrm{nl}}\|$ due to (2.3a), (2.8b), and Lemma 2.2, and we get

$$\|\hat{a}_{k_{\mathrm{bal}}} - a^\dagger\| \le 6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}}\min_{k=1,\ldots,K_{\max}}\left(\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)\right).$$

As $\mu > \frac{1}{2}$, we can use the same arguments as in Lemma 3.4 to show that $K := \operatorname{argmin}_{k\in\mathbb{N}} (\|E_k^{\text{app}}\| + \Phi_{\text{noi}}(k)) \leq K_{\max}$ for δ, σ small enough, and hence the minimum may be taken over all $k \in \mathbb{N}_0$. Moreover, we have

$$\mathbb{P}(\mathcal{C}A_{\tau,K_{\max}}) \leq c_1 \sum_{k=1}^{K_{\max}} \exp(-\ln\sigma^{-2}) = c_1 K_{\max}\sigma^2 \leq C_{\text{tail}}\sigma^2 \ln\sigma^{-1}$$

for some constant $C_{\text{tail}} > 0$. For the last inequality we have used that $K_{\max} = O\left(\ln\sigma^{-1}\right)$ due to the definition of $K_{\max}$ and the fact that φ_{noi} is decreasing. Using again that φ_{noi} is decreasing and $K \to \infty$ as $\sigma \to 0$, we get

(4.15)
$$\mathbb{P}(\mathcal{C}A_{\tau,K_{\max}}) \leq C_{\text{tail}}\sigma^2 \ln\sigma^{-1} \leq C \min_{k\in\mathbb{N}_0} \left(\|E_k^{\text{app}}\| + \delta\frac{C_g}{\sqrt{\alpha_k}} + (\ln\sigma^{-1})\sigma\varphi_{\text{noi}}(\alpha_k)\right)^2$$

for σ small enough with a generic constant C. Hence,

$$\begin{aligned}\mathbf{E}\left[\|\hat{a}_{K_*} - a^\dagger\|^2\right] &\leq \mathbb{P}(A_{\tau,K_{\max}}) \min_{k=1,\dots,K_{\max}} \left(\|E_k^{\text{app}}\| + \Phi_{\text{noi}}(k)\right)^2 + \mathbb{P}(\mathcal{C}A_{\tau,K_{\max}})(3R)^2 \\ &\leq C^2 \min_{k=1,\dots,K_{\max}} \left(\|E_k^{\text{app}}\| + \delta\frac{C_g}{\sqrt{\alpha_k}} + (\ln\sigma^{-1})\sigma\varphi_{\text{noi}}(\alpha_k)\right)^2. \qquad \square\end{aligned}$$

In particular, for $\delta = 0$ and φ_{noi} given by (2.9), we obtain

$$\left(\mathbf{E}\left[\|\hat{a}_{K_*} - a^\dagger\|^2\right]\right)^{1/2} = O\left(\tilde{\rho}^{\frac{c}{\mu+c}}(\sigma\ln\sigma^{-1})^{\frac{\mu}{\mu+c}}\right), \qquad \sigma \to 0. \tag{4.16}$$

Comparing (4.16) to (3.17) or (4.14) to (3.8) shows that we have to pay a logarithmic factor for adaptivity. As shown in [26] it is impossible to achieve optimal rates adaptively in the general situation considered in this paper. However, for special classes of linear inverse problems which are not too ill-posed order optimal adaptive parameter choice rules have been devised (see, e.g., [9]). It remains an interesting open problem to construct order optimal adaptive stopping rules for mildly ill-posed nonlinear statistical inverse problems.

5. Numerical experiments.

To test the predicted rates of convergence with random noise and the performance of the stopping rule (4.3), we consider a simple parameter identification problem for an ordinary differential equation where the forward operator F is easy to evaluate and reliable conclusions can be obtained by Monte Carlo simulations within reasonable time. The efficiency of the iteratively regularized Gauss–Newton method to solve large-scale inverse problems has been sufficiently demonstrated in a number of previous publications (see, e.g., [13]).

Forward operator. For a given positive function $a \in L^2([0,1])$ and a given right-hand side $f \in L^2([0,1])$, let $u \in H^2_{\text{per}}([0,1])$ denote the solution to the ordinary differential equation

$$-u'' + au = f \qquad \text{in } [0,1]. \tag{5.1}$$

Here $H^s_{\text{per}}([0,1])$, $s \geq 0$, denotes the periodic Sobolev space of order 2 with norm

$$\|u\|^2_{H^s_{\text{per}}} = \sum_{j\in\mathbb{N}} (1+j^2)^s \left|\int_0^1 u(x)\exp(-2\pi ijx)\,dx\right|^2.$$

We introduce the parameter-to-solution operator

$$F : D(F) \subset L^2([0,1]) \to L^2([0,1]),$$
$$a \mapsto u.$$

If G denotes the inverse of the differential operator $-\frac{\partial^2}{\partial x^2} + 1$ with periodic boundary conditions, the differential equation (5.1) can equivalently be reformulated as an integral equation

$$u + GM_{a-1}u = Gf$$

with the multiplication operator $M_{a-1}u := (a-1)u$. To prove the Fréchet differentiability of F, it is convenient to consider M_{a-1} as an operator from $C_{\mathrm{per}}([0,1])$ to $L^2([0,1])$ and G as an operator from $L^2([0,1])$ to $C_{\mathrm{per}}([0,1])$. Then M_a is bounded and G is compact, and it is easy to show that F is analytic on the domain $D(F) := \{a \in L^2([0,1]) : I + GM_{a-1}$ boundedly invertible in $C_{\mathrm{per}}([0,1])\}$. In particular, F' satisfies the Lipschitz condition (2.11). By a Neumann series argument, $D(F)$ is open, and it contains all positive functions. Differentiation of (5.1) with respect to a shows that for a perturbation h of a the Fréchet derivative $u_h = F'[a]h$ satisfies the differential equation

$$-u_h'' + au_h = -hu \qquad \text{in } [0,1],$$

where u is the solution to (5.1). This implies that $F'[a]h = -(I + GM_{a-1})^{-1}GM_uh$.

We briefly sketch how Hölder source conditions (3.2) can be interpreted as smoothness conditions in Sobolev spaces under certain conditions. Assume that $a \in H^s \cap D(F)$ with $s > 1/2$, and $f \in C^\infty$. Here and in the following we shortly write H^s for $H^s_{\mathrm{per}}([0,1])$. Since $G : H^t \to H^{t+2}$ is an isomorphism for all $t \geq 0$ and $uv \in H^t$ for $u, v \in H^t$ with $\|uv\|_{H^t} \leq C\|u\|_{H^t}\|v\|_{H^t}$ for $t > 1/2$, it can be shown that $I + GM_a : H^t \to H^t$ is an isomorphism for $t \in [0,s]$ and that $u \in H^s$. Moreover, the operators $F'[a], F'[a]^* : H^t \to H^{\min(s,t+2)}$ are bounded, and if u has no zeros in $[0,1]$ and $t \leq s-2$, the inverses are bounded as well. Under this additional assumption, it can be shown using Heinz's inequality (see [12]) that $(F'[a]^*F'[a])^\mu : L^2 \to H^{4\mu}$ is an isomorphism if $2\lceil 2\mu \rceil \leq s$. Thus, $a, a_0 \in H^s$ implies a Hölder source condition with $\mu = s/4$. Moreover, the singular values of $F'[a]$ behave like $\sigma_j(F'[a]) \sim j^{-2}$, and this asymptotic behavior is uniform for all a with $\|a\|_{L^\infty} \leq C$.

Noise model. Let us assume that our data are n noisy measurements of $u^\dagger = F(a^\dagger)$ at equidistant points $x_j^{(n)} := \frac{j}{n}$,

$$Y_j = u^\dagger\left(x_j^{(n)}\right) + \epsilon_j, \qquad j = 1, \ldots, n. \tag{5.2}$$

The measurement errors are modeled by independent and identically distributed (i.i.d.) random variables ϵ_j satisfying

$$\mathbf{E}\left[\epsilon\right]_j = 0, \qquad \mathbf{Var}\epsilon_j = \sigma_\epsilon^2 < \infty.$$

For n even we introduce the space $\Pi_n := \mathrm{span}\left\{e^{2\pi ijx} : j = -n/2, \ldots, n/2-1\right\}$ and the linear mapping $S_n : \mathbb{R}^n \to \Pi_n$, which maps a vector $\underline{u} = (u_1, \ldots, u_n)^\top$ of nodal values to the unique trigonometric interpolation polynomial $S_n\underline{u} \in \Pi_n$ satisfying $(S_n\underline{u})(x_j^{(n)}) = u_j$, $j = 1, \ldots, n$. We will show that $u^{\mathrm{obs}} := S_n\underline{Y}$ with $\underline{Y} := (Y_1, \ldots, Y_n)^\top$ satisfies assumption (1.2a). Hence, we interpret u^{obs} as our

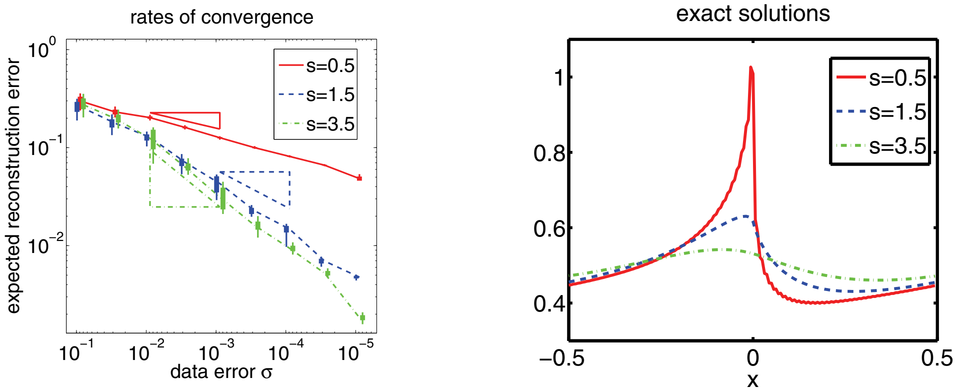

FIG. 5.1. *Left panel: rates of convergence of* $(\mathbf{E}\left[\|a^\dagger - \hat{a}_{k_{\mathrm{bal}}}\|^2\right])^{1/2}$ *as* $\sigma \to 0$ *for different smoothnesses of exact solution. The triangles indicate the rates predicted by theory, and the bars indicate the empirical standard deviations of the reconstruction error. Right panel: exact solutions.*

observed data. Since $\sqrt{n}S_n$ is unitary, u^{obs} and $\underline{Y}$ have unitarily equivalent covariance operators.

We have $\delta\eta = \mathbf{E}\left[u\right]^{\mathrm{obs}} - F(a^\dagger)$ and $\sigma\xi = u^{\mathrm{obs}} - \mathbf{E}\left[u\right]^{\mathrm{obs}}$ in (1.2a), and $\mathbf{E}\left[u\right]^{\mathrm{obs}}$ is the trigonometric interpolation polynomial of $F(a^\dagger)$ at the points $x_j^{(n)}$. Hence, by standard estimates of the trigonometric interpolation error (see, e.g., [24, Cor. 2.47]) the deterministic error of the data function u^{obs} is bounded by

$$\delta = \|\mathbf{E}\left[u\right]^{\mathrm{obs}} - u^\dagger\|_{L^2} \leq \frac{C}{n^2}\|u^\dagger\|_{H^2}.$$

Moreover, the covariance operator of the stochastic noise is given by

$$\mathbf{Cov}_{u^{\mathrm{obs}}} = \mathbf{Cov}_{S_n\underline{\epsilon}} = S_n\mathbf{Cov}_{\underline{\epsilon}}S_n^* = \frac{\sigma_\epsilon^2}{n}P_n,$$

where $\underline{\epsilon} := (\epsilon_1, \dots, \epsilon_n)^\top$ and $P_n \in \mathcal{L}(L^2([0,1]))$ is the orthogonal projection onto Π_n. Note that the stochastic noise level

$$\sigma = \frac{\sigma_\epsilon}{\sqrt{n}} \tag{5.3}$$

dominates the deterministic noise level $\delta = O(n^{-2})$ for large n.

Numerical results. As exact solutions $a^\dagger$ we used three functions of different smoothness shown in the right panel of Figure 5.1. These functions were defined in terms of Fourier coefficients such that they belong to $\bigcap_{t>s} H^t_{\mathrm{per}}([0,1])$ with $s \in \{0.5, 1.5, 3.5\}$. The initial guess was always chosen as the constant function 1. We never had to stop the iteration early because an iterate was not in the domain of definition of F.

We first tested the performance of the balancing principle for $\sigma = 0.1$ and $\sigma = 0.01$ and three different values on n (cf. (5.3)) for the curve with $s = 1.5$. For each value of σ and n, 100 independent data vectors Y were drawn from a Gaussian distribution according to the additive noise model (5.2). We chose $\Phi_{\mathrm{noi}}(k) := (\frac{1}{25}\sum_{l=1}^{25}\|g_{\alpha_k}(T_k^*T_k)T_k^*\underline{\epsilon}^{(l)}\|^2)^{1/2}$, where $\underline{\epsilon}^{(l)}$ are independent copies of the noise vector. Moreover, we set $\gamma_{\mathrm{nl}} = 0.1$ in (4.3). Note that this choice of Φ_{noi} is not fully covered by our theory since we use only an estimator of the expected value, as opposed to (2.8a) we do not have a uniform bound over $a \in D(F)$, and we dropped the logarithmic factor in Theorem 4.4. Furthermore, recall that the Lepskiĭ rule

Table 5.1

Comparison of balancing principle and discrepancy principle as stopping rules. For each value of n and σ_ϵ the algorithm was run 100 *times. The values after* ± *denote standard deviations.*

$\sigma = 0.1$ $K_{\max} = 14$		$n = 200$ $\sigma_\epsilon = 0.01$	$n = 800$ $\sigma_\epsilon = 0.02$	$n = 3200$ $\sigma_\epsilon = 0.04$
Optimal	$\min_k \Vert\hat{a}_k - a^\dagger\Vert$	0.0270 ± 0.0039	0.0263 ± 0.0036	0.0261 ± 0.0039
	$\mathrm{argmin}_k \Vert\hat{a}_k - a^\dagger\Vert$	8.92 ± 0.27	8.92 ± 0.27	8.95 ± 0.22
Balancing	$\Vert\hat{a}_{k_{\mathrm{bal}}} - a^\dagger\Vert$	0.0386 ± 0.0024	0.0383 ± 0.0022	0.0383 ± 0.0025
	k_{bal}	7.00 ± 0	7.00 ± 0	7.00 ± 0
	I_{bal}	1.44	1.47	1.48
Discrepancy	$\Vert\hat{a}_{k_{\mathrm{discr}}} - a^\dagger\Vert$	0.0532 ± 0.0048	0.0763 ± 0.0012	0.1047 ± 0.0009
	k_{discr}	6.09 ± 0.27	5.00 ± 0	4.00 ± 0
	I_{discr}	1.98	2.93	4.06

$\sigma = 0.01$ $K_{\max} = 22$		$n = 200$ $\sigma_\epsilon = 0.001$	$n = 800$ $\sigma_\epsilon = 0.002$	$n = 3200$ $\sigma_\epsilon = 0.004$
Optimal	$\min_k \Vert\hat{a}_k - a^\dagger\Vert$	0.0089 ± 0.0008	0.0088 ± 0.0009	0.0088 ± 0.0008
	$\mathrm{argmin}_k \Vert\hat{a}_k - a^\dagger\Vert$	11.32 ± 0.47	11.43 ± 0.50	11.32 ± 0.47
Balancing	$\Vert\hat{a}_{k_{\mathrm{bal}}} - a^\dagger\Vert$	0.0116 ± 0.0006	0.0116 ± 0.0007	0.0116 ± 0.0006
	k_{bal}	10.00 ± 0	10.00 ± 0	10.00 ± 0
	I_{bal}	1.31	1.33	1.32
Discrepancy	$\Vert\hat{a}_{k_{\mathrm{discr}}} - a^\dagger\Vert$	0.0116 ± 0.0006	0.0171 ± 0.0005	0.0254 ± 0.0003
	k_{discr}	10.00 ± 0	9.00 ± 0	8.00 ± 0
	I_{discr}	1.31	1.96	2.90

requires the computation of iterates up to a fixed $K_{\max}$ specified in (2.14). Usually the Lipschitz constant L involved in the definition (2.14) of $K_{\max}$ is not known exactly. Fortunately, we need only an upper bound L on the Lipschitz constant, and the experimental results are insensitive to the choice of L. As expected from the assumptions of our convergence results, the reconstructions are also insensitive to the choice of $\alpha_0 > 0$ and $q \in (0, 1)$ as long as they are sufficiently large. Further increasing α_0 and q results in more Newton steps and hence more computational effort to reach the optimal value of $\alpha_k = \alpha_0 q^k$, but no noticable difference in the reconstructions.

The results of our first series of simulations are summarized in Table 5.1. As "inefficiency index" of a stopping rule K_* we used the number

$$I := \frac{\left(\mathbf{E}\left[\|\hat{a}_{K_*} - a^\dagger\|^2\right]\right)^{1/2}}{\left(\mathbf{E}\left[\min_{k=0,\dots,K_{\max}} \|\hat{a}_k - a^\dagger\|^2\right]\right)^{1/2}}.$$

The results displayed in Table 5.1 demonstrate that both the expected optimal error (the denominator in the previous expression) and the expected error for the balancing principle (4.3) (the numerator) depend only on $\sigma = \sigma_\epsilon/\sqrt{n}$ but not on n. This is in contrast to the discrepancy principle, which is defined in a discrete setting by

$$k_{\mathrm{discr}} := \min\{k : n^{-1/2}\|\underline{F}(\hat{a}_k) - \underline{Y}\|_{\mathbb{R}^n} \leq \tau\sigma_\epsilon\}.$$

Here $\tau = 2.1$, $\underline{Y} = (Y_j)_{j=1,\dots,n}$ is the vector defined in (5.2), $(\underline{F}(a))_j := (F(a))(x_j^{(n)})$, $j = 1, \dots, n$, and the factor $n^{-1/2}$ before the Euclidean norm in $\mathbb{R}^n$ is chosen such that $n^{-1/2}\|\underline{F}(a)\|_{\mathbb{R}^n} \approx \|F(a)\|_{L^2([0,1])}$. The discrepancy principle for the noise model (5.2) works fairly well for small n but badly for large n, as previously observed, e.g., in [18], for linear problems. The reason is that the standard deviation of the measurement error $n^{-1/2}\big(\mathrm{E}\big[\sum_{j=1}^n \epsilon_j^2\big]\big)^{1/2} = \sigma_\epsilon = \sigma\sqrt{n}$ tends to infinity as $n \to \infty$ with constant σ, and hence the discrepancy principle stops too early. In fact this happens almost always at the first step for n sufficiently large, whereas the optimal stopping index, which asymptotically depends only on σ, but not on n, may be arbitrarily large for small σ.

Finally we tested the rates of convergence with the balancing principle for the three curves $a^\dagger$ shown in the right panel of Figure 5.1. We always chose $n = 128$ and varied σ_ϵ. For each value of σ and each of the three curves $a^\dagger$ we performed 50 runs of the iteratively regularized Gauss–Newton method. The triangles in the left panel show the rates $\sigma^{1/6}$, $\sigma^{3/8}$, and $\sigma^{7/12}$, which are obtained by neglecting the logarithmic factor in (4.16) and setting $c = \frac{5}{8}$ and $\mu = \frac{1}{8}, \frac{3}{8}, \frac{7}{8}$ following the discussion above. Note that the experimental rates agree quite well with these predicted rates even for the first two functions, which are not smooth enough to be covered by our theory.

In summary we have demonstrated that the performance of the balancing principle is independent of the sample size n, whereas the discrepancy principle works well for small n, but becomes more and more inefficient as $n \to \infty$. Moreover, for the balancing principle the empirical rates of convergence match the theoretical rates very well.

Appendix. Exponential inequality for Gaussian noise. In this appendix we will prove the exponential inequality (2.8c) for the case that the noise process ξ in (1.2a) is Gaussian.

Let us consider the random variable $V = \|\Lambda\xi\|^2$ for an arbitrary linear operator $\Lambda : \mathcal{Y} \to \mathcal{X}$ such that $\mathbf{E}\left[\|\Lambda\xi\|^2\right] < \infty$. Since $\mathbf{Cov}_{\Lambda\xi} = \Lambda M \Lambda^*$ with $M := \mathbf{Cov}_\xi$, the operator $\Lambda M \Lambda^*$ is trace class; i.e., the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ of $\Lambda M \Lambda^*$ satisfy $\sum_{i=1}^\infty \lambda_i = \mathbf{E}[V] < \infty$. Note that $M^{-1/2}\xi$ is a Gaussian white noise process; i.e., the random variables $\xi_i := \left\langle M^{-1/2}\xi, \varphi_i\right\rangle$ are i.i.d. standard normal for any orthonormal system $\{\varphi_i\}$ in $\mathcal{Y}$. In particular, if $\{\varphi_i\}$ is a system of left singular vectors of $\Lambda M^{1/2}$, then $V = \sum_{i=1}^\infty \lambda_i \xi_i^2$. The random variables ξ_i^2 are i.i.d. χ_1^2 with Laplace transform

$$\mathbf{E}\left[\exp\left(t\xi_i^2\right)\right] = (1-2t)^{-1/2}, \qquad 0 < t < \frac{1}{2}.$$

Then it holds that

$$\text{(A.1)} \qquad \mathbb{P}\left(\sum_{i=1}^\infty \lambda_i(\xi_i^2 - 1) \geq \eta\sqrt{2\sum_{i=1}^\infty \lambda_i^2}\right) \leq \exp\left(\frac{1}{8} - \frac{\eta}{4}\right)$$

for any $\eta > 0$ (see [11, 25]).

Theorem A.1. *Under the above conditions we have for any $\tau \geq 1$ that*

$$\text{(A.2)} \qquad \mathbb{P}\left(V \geq \tau\mathbf{E}[V]\right) \leq \exp\left(\frac{1}{8} - \frac{C}{4}(\tau - 1)\right),$$

where $C := 2^{-1/2}\sum_{i=1}^\infty \lambda_i / (\sum_{i=1}^\infty \lambda_i^2)^{1/2}$.

Proof. Rewrite

$$\{V \geq \tau\mathbf{E}[V]\} = \left\{\sum_{i=1}^\infty \lambda_i\xi_i^2 \geq \tau\sum_{i=1}^\infty \lambda_i\right\} = \left\{\sum_{i=1}^\infty \lambda_i(\xi_i^2 - 1) \geq \eta\sqrt{2\sum_{i=1}^\infty \lambda_i^2}\right\},$$

where $\eta = (\tau - 1)C$, and apply (A.1). □

Since $\lambda_i \geq 0$, we have $\sum_{i=1}^\infty \lambda_i^2 \leq \left(\sum_{i=1}^\infty \lambda_i\right)^2$, and hence $C \geq 2^{-1/2}$ for any eigenvalue sequence (λ_i). Therefore, choosing $\Lambda := g_\alpha(F'[a]^*F'[a])F'[a]^*$, we obtain (2.8c) with $c_1 = \exp(\frac{1}{8} + \frac{1}{4\sqrt{2}})$ and $c_2 = \frac{1}{4\sqrt{2}}$.

A variation of the proof in [25] allows in principle an extension for more general (non χ^2) random variables under proper growth conditions on the Laplace transform of the ξ_i^2. This would again give an exponential bound of the form (2.8c).

REFERENCES

[1] A. B. Bakushinskiĭ, *The problem of the convergence of the iteratively regularized Gauss-Newton method*, Comput. Math. Math. Phys., 32 (1992), pp. 1353–1359.

[2] A. B. Bakushinskiĭ and M. Y. Kokurin, *Iterative Methods for Approximate Solution of Inverse Problems*, Springer, Dordrecht, The Netherlands, 2004.

[3] F. Bauer and T. Hohage, *A Lepskij-type stopping rule for regularized Newton methods*, Inverse Problems, 21 (2005), pp. 1975–1991.

[4] N. Bissantz, T. Hohage, and A. Munk, *Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise*, Inverse Problems, 20 (2004), pp. 1773–1791.

[5] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart, *Convergence rates of general regularization methods for statistical inverse problems and applications*, SIAM J. Numer. Anal., 45 (2007), pp. 2610–2636.

[6] B. Blaschke, A. Neubauer, and O. Scherzer, *On convergence rates for the iteratively regularized Gauss-Newton method*, IMA J. Numer. Anal., 17 (1997), pp. 421–436.

[7] T. T. Cai and M. G. Low, *Nonparametric estimation over shrinking neighborhoods: Superefficiency and adaptation*, Ann. Statist., 33 (2005), pp. 184–213.

[8] E. Candès, *Modern statistical estimation via oracle inequalities*, Acta Numer., 15 (2006), pp. 257–325.

[9] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov, *Oracle inequalities for inverse problems*, Ann. Statist., 30 (2002), pp. 843–874.

[10] R. Cont and P. Tankov, *Retrieving Lévy processes from option prices: Regularization of an ill-posed inverse problem*, SIAM J. Control Optim., 45 (2006), pp. 1–25.

[11] R. Dahlhaus and W. Polonik, *Nonparametric quasi-maximum likelihood estimation for Gaussian locally stationary processes*, Ann. Statist., 34 (2006), pp. 2790–2824.

[12] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1996.

[13] T. Hohage, *Fast numerical solution of the electromagnetic medium scattering problem and applications to the inverse problem*, J. Comput. Phys., 214 (2006), pp. 224–238.

[14] T. Hohage and M. Pricop, *Nonlinear Tikhonov regularization in Hilbert scales for inverse boundary value problems with random noise*, Inverse Probl. Imaging, 2 (2008), pp. 271–290.

[15] B. Kaltenbacher, *Some Newton-type methods for the regularization of nonlinear ill-posed problems*, Inverse Problems, 13 (1997), pp. 729–753.

[16] B. Kaltenbacher, *A posteriori parameter choice strategies for some Newton type methods for the regularization of nonlinear ill-posed problems*, Numer. Math., 79 (1998), pp. 501–528.

[17] J.-M. Loubes and C. Ludeńa. *Penalized estimators for nonlinear inverse problems*, ESAIM Probab. Stat., to appear.

[18] M. A. Lukas, *Comparison of parameter choice methods for regularization with discrete noisy data*, Inverse Problems, 14 (1998), pp. 161–184.

[19] P. Mathé, *The Lepskiĭ principle revisited*, Inverse Problems, 22 (2006), pp. L11–L15.

[20] P. Mathé and S. Pereverzev, *Geometry of ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.

[21] P. Mathé and S. Pereverzev, *Regularization of some linear ill-posed problems with discretized random noisy data*, Math. Comp., 75 (2006), pp. 1913–1929.

[22] W. K. Newey and J. L. Powell, *Instrumental variable estimation of nonparametric models*, Econometrica, 71 (2003), pp. 1565–1578.

[23] F. O'Sullivan, *Convergence characteristics of methods of regularization estimators for nonlinear operator equations*, SIAM J. Numer. Anal., 27 (1990), pp. 1635–1649.

[24] S. Prössdorf and B. Silbermann, *Numerical Analysis for Integral and Related Operator Equations*, Birkhäuser, Basel, 1991.

[25] A. Rohde and L. Dümbgen, *Confidence Sets for the Optimal Approximating Model—Bridging a Gap between Adaptive Point Estimation and Confidence Regions*, 2008; available online at http://arxiv.org/abs/0802.3276.

[26] A. Tsybakov, *On the best rate of adaptive estimation in some inverse problems*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 835–840.